SELF-SUPERVISED DEEP LEARNING FOR ADVANCING MACROMOLECULAR ANALYSIS IN CRYO-ELECTRON TOMOGRAPHY

FROSINA STOJANOVSKA

Ruprecht-Karls-University Heidelberg

December 2024

Self-supervised deep learning for advancing macromolecular analysis in cryo-electron tomography

Frosina Stojanovska, M.Sc.

DEFENSE COMMITTEE:
Prof. Dr. Robert Russell
Dr. Anna Kreshuk
Prof. Dr. Irmgard Sinning
Dr. Simone Mattei

SUPERVISOR:
Prof. Dr. Judith Zaugg

Heidelberg, Germany

Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Frosina Stojanovska

born in: Skopje, North Macedonia

Oral examination: 17.03.2025

SELF-SUPERVISED DEEP LEARNING FOR ADVANCING MACROMOLECULAR ANALYSIS IN CRYO-ELECTRON TOMOGRAPHY

Referees:
Prof. Dr. Robert Russell
Dr. Anna Kreshuk

ABSTRACT

Cryo-electron tomography (cryo-ET) provides unprecedented insights into cellular structures and macromolecular complexes in their native states. However, its interpretation remains challenging due to high noise levels, low contrast, and structural complexity. This thesis presents CryoSiam, a novel self-supervised deep learning framework that addresses these challenges through denoising, semantic segmentation, and particle identification tasks.

Trained entirely on simulated tomograms, CryoSiam leverages self-supervised learning to generate robust voxel- and subtomogram-level embeddings, circumventing the need for annotated real data. Comprehensive ablation studies identified key design choices that optimize performance across tasks. When applied to publicly available real datasets from EMPIAR and the CryoET Data Portal, CryoSiam demonstrated effective generalization to real-world conditions, achieving performance comparable to, and at times exceeding, state-of-the-art supervised methods.

The framework showed versatility in delivering high-quality noise suppression, accurate membrane segmentation without real-data training, and reliable particle identification based on learned embeddings. These results highlight the potential of self-supervised learning to bridge the gap between simulated training environments and real cryo-ET applications.

This thesis also addresses existing limitations, such as reliance on simulated data and challenges in representing structural diversity. Future directions include expanding simulation diversity, exploring semi-supervised approaches, and enhancing computational efficiency. CryoSiam establishes a foundation for advancing cryo-ET analysis, promoting open science, collaboration, and a deeper understanding of cellular architecture at molecular resolution.

ZUSAMMENFASSUNG

Die Kryo-Elektronentomographie (cryo-ET) ermöglicht beispiellose Einblicke in zelluläre Strukturen und makromolekulare Komplexe in ihrem nativen Zustand. Ihre Interpretation bleibt jedoch eine Herausforderung aufgrund hoher Rauschpegel, geringer Kontraste und struktureller Komplexität. In dieser Dissertation wird mit CryoSiam ein neuartiges selbstüberwachtes Deep-Learning-Framework vorgestellt, das diesen Herausforderungen durch Rauschunterdrückung, semantische Segmentierung und Partikelerkennung begegnet.

CryoSiam wird vollständig mit simulierten Tomogrammen trainiert und nutzt selbstüberwachtes Lernen, um robuste Voxel- und Subtomogramm-Embeddings zu erzeugen, wodurch die Notwendigkeit für annotierte reale Daten umgangen wird. Umfassende Ablationsstudien identifizierten zentrale Designentscheidungen, die die Leistung über verschiedene Aufgaben hinweg optimieren. Bei der Anwendung auf öffentlich verfügbare reale Datensätze von EMPIAR und dem CryoET Data Portal zeigte CryoSiam eine effektive Generalisierung auf reale Bedingungen und erreichte dabei Ergebnisse, die mit den derzeit besten überwachten Methoden vergleichbar sind oder diese teilweise übertreffen.

Das Framework erwies sich als vielseitig, indem es hochwertige Rauschunterdrückung, präzise Membransegmentierung ohne Training mit realen Daten sowie zuverlässige Partikelerkennung auf Basis gelernter Embeddings lieferte. Diese Ergebnisse unterstreichen das Potenzial selbstüberwachter Lernansätze, die Lücke zwischen simulierten Trainingsumgebungen und realen cryo-ET-Anwendungen zu schließen.

Diese Dissertation behandelt auch bestehende Einschränkungen wie die Abhängigkeit von simulierten Daten und Herausforderungen bei der Repräsentation struktureller Vielfalt. Künftige Forschungsansätze umfassen die Erweiterung der Simulationsvielfalt, die Erforschung semisupervisierter Ansätze und die Steigerung der Recheneffizienz. CryoSiam schafft eine Grundlage für den Fortschritt in der Analyse von cryo-ET-Daten und fördert offene Wissenschaft, Zusammenarbeit sowie ein tieferes Verständnis der zellulären Architektur auf molekularer Ebene.

ACKNOWLEDGEMENTS

I want to express my heartfelt gratitude to everyone who has supported and motivated me throughout my PhD journey.

First and foremost, I want to thank my supervisor, Judith Zaugg, for her invaluable guidance, encouragement, and unwavering support. Her insightful advice and thoughtful feedback have been instrumental in shaping my research and personal development. I am deeply grateful for her mentorship and the opportunities I have had to grow under her supervision.

I also sincerely thank Julia Mahamid and Anna Kreshuk for their collaboration and support during my PhD. Working closely with their groups at EMBL has been a truly enriching experience. Their expertise and enthusiasm for science have been a great source of inspiration, and I have learned much from our interactions.

I want to express my special thanks to Sonja Gievska, my supervisor during my Bachelor's and Master's studies, for her constant encouragement and support, not only before I embarked on this PhD journey but also throughout it. Her unwavering belief in me has strengthened and motivated me during challenging times.

I want to thank the IT services team at EMBL and Jurij Pecar for ensuring seamless access to computational resources and providing crucial support for the cluster infrastructure. Your assistance has been essential for my research, and I greatly appreciate your help keeping everything running smoothly.

I am incredibly grateful to Liang Xue, who guided me through the initial understanding of cryo-electron tomography data, particularly in the context of the Mycoplasma project. I would also like to thank Rasmus Kjeldsen Jensen for his invaluable support in understanding the Mycoplasma data and for introducing me to experimental work in the lab. His guidance during my first lab experiences significantly impacted my research and personal growth. Additionally, I am deeply thankful to Ricardo Sanchez for our numerous discussions about the challenges and attributes of cryo-electron tomography data. His insights and advice were always greatly appreciated.

I want to acknowledge the members of my thesis advisory committee, Robert Russell, Anna Kreshuk, Julia Mahamid, and Jan Kosinski, for their time, guidance, and valuable feedback throughout my PhD. Your insights and suggestions have contributed significantly to improving the quality of my work.

I am incredibly grateful for the close collaboration with the Mahamid lab, whose expertise in cryo-electron tomography has been integral to my research.

I also want to express my appreciation to the Kreshuk lab for the stimulating deep-learning discussions that helped shape the computational aspects of my work. The Zaugg lab has provided a welcoming and supportive environment throughout my PhD. I thank all my lab mates for their collaboration, discussions, and encouragement, making this journey much more enjoyable. Special thanks to my colleagues from Office 113, particularly Max, Christian, Evi, Charles, Jupa, Kristy, Anna, and Ivan, for the laughter, conversations, and uplifting energy. I am grateful to all of you for creating such a welcoming and supportive environment during my PhD.

To my friends and colleagues at EMBL, thank you for the shared laughter, coffee breaks, and late-night discussions that made my time here memorable. The friendships I have built during my PhD have been a source of joy and support. Special thanks to Kristy, Lena, Alyona, Merve, Efi, Amelia, Manu, Qin, Dorothy, Simona, and Kavan for your kindness and encouragement and for making everyday life at and outside of EMBL brighter.

I thank my family and friends outside of EMBL for their patience and constant encouragement. Your unwavering support has been a source of strength, and I am truly grateful for everything you have done to help me reach this point. A special acknowledgment goes to Martina Toshevska, a long-time friend and collaborator. Beyond academic discussions, her unwavering support and shared conversations throughout our respective PhD journeys have been a source of motivation and strength.

A very special thanks to my mom, who has always supported me unconditionally through every step of this journey. Your strength, love, and encouragement have been my foundation. Thank you to my brother and sister for your unwavering love, support, and belief in me. Your presence in my life has been invaluable, and I am endlessly grateful for everything you do. Ви благодарам!

STATEMENT OF CONTRIBUTIONS

This thesis was conducted at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, from October 2020 to December 2024 and was supervised by Prof. Dr. Judith Zaugg.

Although this research was conducted within a collaborative environment, all results and analyses presented in this thesis were independently produced and conducted by the author unless explicitly stated otherwise. The author was solely responsible for the design and implementation of the methods, the execution of the experiments, the data simulation, the analysis, and the interpretation of the results discussed herein.

The computational analyses and simulations required for this research were performed using the advanced resources and infrastructure provided by the High-Performance Computing (HPC) cluster at EMBL. Access to these resources, including high-performance computing systems and technical support, was critical for managing the computationally intensive components of this work. The author gratefully acknowledges the vital role of these facilities in enabling the successful completion of this research.

Additionally, the computing time provided by the high-performance computers at the NHR Centers NHR@TUD is gratefully acknowledged. This support is funded by the Federal Ministry of Education and Research and the state governments participating based on the resolutions of the GWK for the national high-performance computing at universities.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

2D Two-dimensional

3D Three-dimensional

CNN Convolutional neural network

Cryo-EM Cryo-electron microscopy

Cryo-ET Cryo-electron tomography

Cryo-FIB Cryo-focused ion beam

CTF Contrast transfer function

DED Direct electron detectors

DL Deep learning

DNA Deoxyribonucleic acid

FPN Feature pyramid network

k-NN k-Nearest neighbors

MLP Multi-layer perceptron

PCA Principal component analysis

RGB Red-Green-Blue

RNA Ribonucleic acid

SNR Signal-to-noise ratio

SSL Self-supervised learning

STA Subtomogram averaging

TEM Transmission electron microscope

TM Template matching

UMAP Uniform manifold approximation and projection for dimension reduction

VPP Volta phase plate

WBP Weighted back projection

Part I

FIELD OVERVIEW AND RESEARCH CONTEXT

Understanding the challenges of cryo-electron tomography (cryo-ET) data segmentation necessitates a comprehensive examination of the field's foundations. In this part, cryo-ET is introduced as a transformative technique in structural biology, showcasing its capability to reveal macromolecular structures in their native states. It examines the intricacies of the segmentation problem, reviewing current methodologies and thoughtfully assessing their limitations. Furthermore, this part highlights the emergence of self-supervised learning as an innovative approach, underscoring its potential to address critical challenges in analyzing cryo-ET data.

1 INTRODUCTION

The structural and functional characterization of biological macromolecules is a central objective of structural biology, a discipline that elucidates the molecular mechanisms underlying cellular processes [1]. Such insights are essential for driving progress in fundamental biological research. Established methodologies, including X-ray crystallography and single-particle cryo-electron microscopy (cryo-EM), have successfully resolved high-resolution structures [2, 3]. However, these techniques are inherently limited in their ability to visualize macromolecular complexes within their native cellular contexts [4, 5]. This limitation highlights the pressing need for imaging approaches that integrate high-resolution, three-dimensional (3D) structural data with the preservation of biological context.

Cryo-electron tomography (cryo-ET) meets this need by allowing the visualization of biological specimens in near-native states with sub-nanometer resolution [6, 7]. Its ability to bridge the molecular and cellular scales gap has made cryo-ET an essential tool for exploring complex biological systems [8, 9]. Nevertheless, analyzing cryo-ET data presents unique challenges, including high noise levels, missing data artifacts, and a lack of annotated datasets for segmentation and interpretation [10]. These challenges highlight the necessity for developing advanced computational methods to leverage cryo-ET's potential fully [11].

Recent advancements in deep learning (DL), particularly self-supervised learning (SSL), offer promising solutions to the challenges inherent in cryo-ET data analysis [12]. In this thesis, I leverage SSL to address key challenges in cryo-ET data processing. SSL enables models to learn meaningful data representations from unlabeled data by leveraging pretext tasks, such as predicting missing information or reconstructing spatial structures [13]. Unlike traditional supervised learning approaches, which require extensive labeled datasets, SSL is particularly well-suited for cryo-ET data, where annotations are scarce and expensive to generate [14]. Furthermore, SSL excels in producing generalizable representations that can be fine-tuned for downstream tasks like segmentation, denoising, and particle identification. Through my work, I aim to address the variability and complexity of cryo-ET datasets using SSL [15, 16]. By reducing dependency on annotated data and enabling transferability across tasks and datasets, I demonstrate that SSL represents a transformative approach for advancing cryo-ET analysis.

This chapter explores the role of cryo-ET in structural biology, emphasizing its significance, applications, and the challenges inherent in its data analysis. I provide an overview of cryo-ET principles and their impact on the field in Section 1.1. In Section 1.2, I highlight the critical challenges of cryo-ET data analysis, while in Section 1.3, I introduce the potential of SSL to address these challenges. The chapter concludes with a discussion of the limitations of existing approaches and the specific contributions I make in this thesis to advance cryo-ET analysis.

1.1 BACKGROUND AND IMPORTANCE OF CRYO-ELECTRON TOMOGRAPHY

Cryo-ET represents a transformative advancement in biological imaging, providing 3D visualizations of biological specimens under near-native conditions [7, 17]. Unlike traditional structural biology techniques such as X-ray crystallography and single-particle cryo-EM, cryo-ET preserves the structural integrity of specimens by rapidly freezing them in a vitreous state, halting biological activity and capturing their molecular configuration at the moment of vitrification [18, 19]. By acquiring multiple two-dimensional (2D) projections at varying tilt angles, cryo-ET reconstructs detailed 3D tomograms that retain the biological context of the sample [17].

As illustrated in Figure 1, the ultimate goal of cryo-ET is to achieve detailed structural reconstructions of biological specimens, providing insights into macromolecular organization in their native cellular environments. However, this process is far from straightforward. While cryo-ET can capture high-resolution structural details, such as those seen in the Mycoplasma cell example in Figure 1, reaching this level of detail is highly challenging. The imaging process is hindered by noise, missing data, and sample variability, making robust computational methods essential for tomogram reconstruction and analysis. Addressing these challenges is a central focus of this thesis.

What distinguishes cryo-ET from cryo-EM is its ability to image intact cellular environments rather than isolated particles [20]. By studying biological systems in situ, cryo-ET provides critical contextual information about macromolecular complexes' spatial organization and interactions within the cellular environment [9]. This capability enables the examination of cellular attributes such as organelle morphology, molecular interactions, and the structural arrangement of dynamic networks like the cytoskeleton [21–24]. Such insights are unattainable with traditional methods that cannot preserve and visualize the intricate organization of biological systems in their native states.

Cryo-ET's core principle, tilt-series acquisition, involves incrementally tilting the sample holder within a transmission electron microscope (TEM) to capture a series of 2D projections [25]. These projections are computationally reconstructed into 3D tomograms that reveal the spatial relationships among cellular components [26]. The success of this technique hinges on meticulous sample preparation. Vitrification prevents ice crystal formation, preserving fine structural details, while cryo-focused ion beam (cryo-FIB) milling enables imaging of thicker samples by creating electron-transparent lamellae [19, 27].

The unique ability of cryo-ET to visualize biological specimens in their near-native states has catalyzed transformative discoveries across structural biology [23]. Cryo-ET has revealed the molecular landscapes of organelles such as mitochondria and the Golgi apparatus, illuminated the assembly of viruses within host cells, and captured the architecture of cytoskeletal networks during dynamic cellular processes [20, 28–31]. Compared to traditional techniques, cryo-ET offers distinct advantages: while X-ray crystallography resolves static, crystalline structures at atomic resolution, cryo-ET excels in analyzing heterogeneous and dynamic systems without requiring crystallization [4]. Similarly, single-particle cryo-EM provides high-resolution reconstructions of isolated

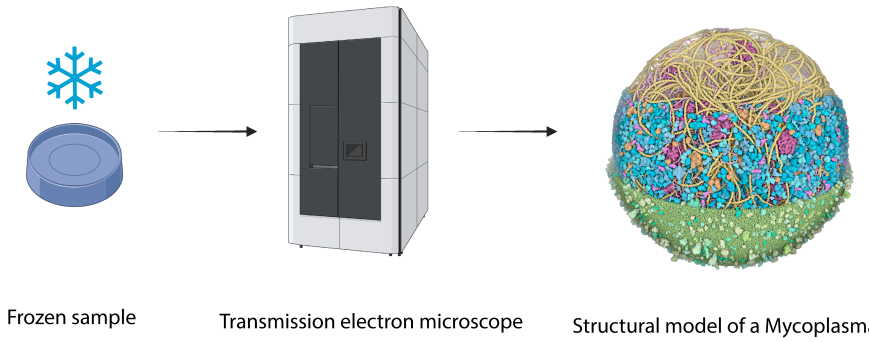

Figure 1: **Workflow of cryo-electron tomography for visualizing macromolecular structures.** The process begins with the vitrification of a biological sample to preserve it in a near-native state, followed by imaging using a transmission electron microscope (TEM). The final goal is to reconstruct high-resolution structural details, such as the organization of macromolecular complexes within a Mycoplasma cell. Achieving such detailed structural insights is challenging due to the complexity and variability of biological specimens, the inherent noise in cryo-ET data, and the need for advanced computational methods for accurate analysis. The Mycoplasma cell illustration is by David Goodsell, and the image was partially created using BioRender.

macromolecules, but lacks the contextual information cryo-ET delivers, capturing the spatial organization of molecules within intact cells [17].

Recent technological innovations have made cryo-ET a cornerstone of structural biology. The development of direct electron detectors has significantly improved signal-to-noise ratio (SNR), enhancing data quality and resolution [32, 33]. Advances in computational reconstruction algorithms, including robust techniques for tilt-series alignment and tomogram reconstruction, have further refined the quality of cryo-ET data [34]. Additionally, workflows integrating cryo-FIB milling and correlative light and electron microscopy (CLEM) have expanded the applicability of cryo-ET to larger and more complex systems [35, 36]. These innovations continue to push the boundaries of cryo-ET, enabling breakthroughs in areas such as structural virology, organelle dynamics, and cellular architecture.

1.2 CHALLENGES IN SEGMENTATION AND ANALYSIS OF CRYO-ET DATA

Although cryo-ET provides detailed 3D visualizations of biological systems in near-native states, analyzing its datasets and extracting biologically meaningful insights remains a significant challenge [37–39]. A central difficulty lies in the complexity of cryo-ET data, characterized by high noise levels, artifacts introduced during data acquisition, and the inherent heterogeneity of biological samples [39–41]. These challenges are compounded by variability in imaging conditions, contrast, and resolution, fully underscoring the pressing need for advanced computational tools to leverage cryo-ET's potential [23, 39].

A significant challenge in cryo-ET data analysis is its inherently low SNR [39, 42]. This limitation arises from the requirement to minimize electron dosage

during imaging, which is critical for preserving the native structure of biological specimens [43]. However, this constraint results in noisy projection images and, consequently, noisy tomograms [6, 20]. The low SNR obscures structural details, complicating the differentiation of macromolecular features from background noise [43].

Technological advancements, such as direct electron detectors (DEDs), have improved SNR by enhancing sensitivity and enabling low-dose imaging [8]. Nevertheless, noise remains a significant barrier, particularly in regions of tomograms where biological structures are faint or overlapping. Addressing this issue necessitates robust computational denoising methods that enhance feature visibility without distorting the underlying data [41, 42, 44, 45].

In addition to noise and artifacts, variability in contrast and resolution presents significant challenges for cryo-ET segmentation. Contrast differences can result from imaging parameters, such as defocus settings and electron dose, as well as physical characteristics of the sample, such as ice thickness [23, 46, 47]. On the other hand, resolution is influenced by factors including the imaging system, the target resolution during acquisition, and the local structural density within the tomogram [38]. Notably, regions within the same tomogram may exhibit varying resolutions due to [48]. For instance, areas closer to the edges of a thick sample often suffer from reduced resolution, complicating the accurate segmentation and identification of features [49]. These inconsistencies necessitate segmentation algorithms capable of adapting to variable imaging conditions and spatial resolution within and across datasets [50].

The "missing wedge" artifact represents another substantial challenge in cryo-ET data analysis, stemming from the limited angular range during tilt-series acquisition [51]. Mechanical constraints and the need to minimize electron exposure typically restrict tilt angles to ±60 or ±70 degrees, preventing the acquisition of a full 180-degree angular range [19]. This incomplete angular coverage creates gaps in the Fourier transform of the data, resulting in anisotropic resolution and artifacts along the tilt axis [52].

These missing wedge artifacts significantly distort the reconstructed 3D volumes, complicating segmentation and interpretation [52]. Computational strategies, such as iterative reconstruction algorithms and advanced regularization techniques, have been employed to address this issue [53, 54]. However, these methods have only partially mitigated the effects of the missing wedge and have yet to achieve universal applicability across the diverse range of cryo-ET datasets [54–56].

The heterogeneity of cryo-ET samples adds another layer of complexity. Variability arises from differences in sample preparation, imaging conditions, and the intrinsic biological diversity of the specimens [57]. For instance, uneven ice thickness across the sample grid can result in regions with varying contrast and resolution, further complicating analysis [23, 47]. Similarly, biological variability, such as differences in macromolecular composition and density, introduces additional challenges for segmentation [58].

This heterogeneity poses particular difficulties for automated segmentation tools, which often struggle to generalize across datasets with diverse structural and imaging characteristics. Features with overlapping density profiles, such

as membrane-bound proteins or cytoskeletal filaments, are incredibly challenging to identify and distinguish [59]. Developing segmentation methods to address these challenges is essential for advancing cryo-ET data analysis.

Segmentation remains one of cryo-ET data analysis's most challenging and critical steps. Despite the development of numerous methods, many approaches rely heavily on supervised learning, which requires extensive ground truth data for training [50, 58, 60]. Generating this ground truth is a labor-intensive and time-consuming process that depends on expert knowledge, presenting a significant bottleneck in the segmentation workflow [58]. The inherent complexity of cryo-ET tomograms, characterized by low SNR and overlapping density profiles, further exacerbates the difficulties of manual annotation, limiting the scalability of traditional methods.

Fully supervised segmentation techniques, such as U-Net-based architectures and convolutional neural networks (CNNs), have demonstrated strong performance in identifying features within cryo-ET tomograms [58]. However, their effectiveness is intrinsically tied to the availability of high-quality annotated datasets [20, 61, 62]. This reliance on annotated data, combined with the variability and heterogeneity inherent to cryo-ET datasets, has spurred the exploration of alternative approaches that reduce dependency on manual labeling while maintaining or even improving segmentation accuracy [37, 63].

Recent advances have focused on automated and self-supervised segmentation methods, which aim to address the limitations of manual annotation by utilizing vast amounts of unlabeled cryo-ET data. Innovations such as DeepETPicker [64], TomoTwin [65], CryoSAM [66] and MiLoPYP [67] have introduced semi-automated workflows that enable faster and more efficient segmentation with minimal human intervention. Moreover, self-supervised and unsupervised learning approaches have demonstrated substantial promise by extracting meaningful representations directly from raw data. For instance, MiLoPYP [67] employs self-supervised pretext tasks to enhance feature extraction, while CryoSAM [66] bridges the gap between 2D segmentation models and 3D volumetric segmentation, offering a novel framework for tomogram analysis.

These advancements address critical gaps in cryo-ET segmentation by enabling models to generalize across datasets with varying noise levels, contrast, and resolutions. By reducing reliance on labor-intensive manual annotations, these methods pave the way for scalable, high-throughput analysis of tomograms [65]. Nevertheless, while significant progress has been made, developing robust, fully automated tools capable of handling cryo-ET data's inherent variability and artifacts remains an ongoing challenge.

Cryo-ET segmentation transforms from manual, supervised workflows to more adaptable and efficient self-supervised and unsupervised approaches. These innovations hold the potential to unlock cryo-ET's full capabilities, enabling faster, more accurate, and less labor-intensive analyses that can meet the demands of modern structural biology.

1.3 IMPORTANCE OF SELF-SUPERVISED LEARNING FOR IMAGE SEGMENTATION

Self-supervised learning (SSL) is an emerging machine learning paradigm that reduces the dependence on labeled datasets by enabling models to learn directly from unlabeled data [68]. SSL utilizes pretext tasks, unlike traditional supervised learning, which relies on large quantities of annotated data for training. These auxiliary tasks are designed to uncover meaningful features inherent in the data [69]. Examples include predicting missing parts of an image [70, 71], identifying spatial relationships [72, 73], or distinguishing between transformed and original samples [74]. By focusing on these tasks, SSL enables models to learn rich representations that can be adapted for downstream tasks like segmentation, classification, or detection, even in scenarios where labeled data is unavailable [75, 76].

Supervised learning, while highly effective across various domains, has a critical limitation due to its reliance on extensive manual annotation, particularly in fields like cryo-ET where labeled data is scarce and expensive to generate [58, 77, 78]. SSL addresses this issue by leveraging the intrinsic structure of data, allowing models to learn directly from unlabeled datasets [79]. This approach is particularly advantageous for cryo-ET, where manual annotation is labor-intensive and requires significant domain expertise to segment complex structures within noisy and artifact-laden tomograms [58].

Cryo-ET data is especially suited to benefit from SSL because of its unique challenges. The heterogeneity of biological samples, variability in imaging conditions, and scarcity of annotated datasets create significant obstacles for supervised approaches [80]. SSL mitigates these challenges by enabling models to extract meaningful features from raw, unlabeled cryo-ET data. For instance, pretext tasks in SSL can guide models to identify structural patterns within noisy tomograms or learn invariant features that generalize across varying imaging conditions [14, 81, 82].

A key strength of SSL in cryo-ET analysis is its ability to generalize across datasets and tasks. Cryo-ET datasets often exhibit substantial variation in noise levels, imaging parameters, and sample preparation methods. Models trained using supervised learning on a specific dataset may overfit the training data and struggle to adapt to new datasets or experimental conditions [83, 84]. SSL, in contrast, focuses on learning transferable representations that capture the underlying structure of the data, making it highly adaptable to diverse datasets and segmentation tasks [75, 76].

SSL's potential to improve segmentation accuracy, accelerate workflows, and make cryo-ET analysis more scalable is transformative. SSL addresses one of the most significant bottlenecks in cryo-ET data analysis by reducing reliance on manual annotations. This thesis explores the application of SSL to cryo-ET segmentation, leveraging its ability to tackle the unique challenges posed by cryo-ET data and enabling more efficient, robust, and generalizable analysis.

1.4 LIMITATIONS OF EXISTING APPROACHES AND THE PROPOSED SOLUTION

Despite significant advancements in cryo-ET, the segmentation and analysis of tomograms continue to face considerable challenges due to limitations in existing methods. Inefficiencies, a lack of generalizability, and the scarcity of annotated data constrain current approaches, creating bottlenecks in cryo-ET workflows and hindering the full realization of its transformative imaging capabilities [20, 50].

Manual segmentation remains a widely used method for analyzing cryo-ET data, relying heavily on the expertise of domain specialists to identify and annotate macromolecular features within noisy and artifact-laden tomograms. However, this process is inherently labor-intensive, time-consuming, and prone to variability due to human error [61]. Such limitations render manual segmentation impractical for large-scale or high-throughput analyses, particularly as the volume of cryo-ET datasets grows.

Template matching offers a degree of automation by employing cross-correlation techniques to detect predefined templates within tomograms. While this approach reduces the reliance on manual annotations, its adaptability is limited. Template matching struggles to accommodate the complexity and heterogeneity of cryo-ET datasets, as it heavily depends on the choice of template [85, 86]. Macromolecules with smaller sizes or lower contrast often evade accurate detection, further diminishing the utility of this method.

Supervised learning approaches provide more robust segmentation capabilities by leveraging large annotated datasets for training [50, 58, 60]. However, generating these datasets is an arduous and resource-intensive process, requiring extensive manual annotation. Moreover, supervised models frequently lack generalizability, excelling only on datasets that closely resemble their training data. This limitation is particularly problematic in cryo-ET, where variability in sample composition, imaging conditions, and noise levels are prevalent.

To address the scarcity of annotated datasets, simulated data, such as the SHREC dataset, provides valuable ground truth for evaluating segmentation methods [87, 88]. These datasets enable controlled experimentation and are reliable benchmarks for developing self-supervised learning frameworks [89]. However, existing simulated datasets often fail to replicate the crowdedness and structural complexity observed in real cryo-ET data [89, 90]. Their simplified macromolecular arrangements limit their applicability for training models that perform well under real-world conditions. Additionally, the limited number of tomograms available from experimental datasets further constrains the development of DL methods [90].

In this thesis, I developed CryoSiam, a novel SSL framework, to address these limitations. CryoSiam leverages SSL to extract meaningful representations from unlabeled cryo-ET data, reducing the dependence on annotated datasets. The framework enhances generalizability across diverse tomograms, even those with varying noise levels and imaging conditions. By focusing on learning transferable features directly from the data, CryoSiam addresses a significant gap in current segmentation approaches.

To support the training and evaluation of this framework, I created CryoETSim, an extensive simulated dataset designed to resemble real cryo-ET data more closely than existing simulated datasets. CryoETSim incorporates key features, such as crowdedness and structural variability, which are often missing in existing datasets. This dataset bridges the gap between simulated and experimental data, providing a robust foundation for developing and testing segmentation models under more realistic conditions [89, 90].

Furthermore, I integrated multiple segmentation tasks within the CryoSiam framework to improve the efficiency and scalability of cryo-ET analysis. These tasks include semantic segmentation, which identifies distinct regions within tomograms; instance segmentation, which differentiates individual macromolecular features in crowded environments; and particle identification, which localizes and characterizes particles within the sample. CryoSiam significantly streamlines segmentation workflows and improves overall accuracy by combining these tasks within a single framework.

Through developing CryoSiam and CryoETSim, I address critical challenges in cryo-ET segmentation, reducing reliance on annotated datasets and enhancing model generalizability. This work demonstrates that self-supervised learning can provide an efficient, accurate, and scalable solution for analyzing cryo-ET data, paving the way for further advancements in structural biology.

1.5 RESEARCH OBJECTIVES AND QUESTIONS

The primary objective of this thesis is to advance the segmentation and analysis of cryo-ET data by addressing critical limitations in existing methods. While cryo-ET offers unparalleled insights into macromolecular structures within their native environments, its transformative potential in structural biology is constrained by several inherent challenges. These include noisy and artifact-laden data, the scarcity of annotated datasets, and the lack of generalizable segmentation models. To overcome these obstacles, I developed CryoSiam, a novel self-supervised learning framework designed to improve segmentation accuracy, reduce reliance on manual annotations, and enhance model generalizability.

CryoSiam aims to provide an efficient and scalable solution for segmenting tomograms by leveraging the power of self-supervised learning to address the bottleneck created by limited labeled data in traditional supervised approaches. CryoSiam ensures robust performance across diverse datasets with varying noise levels and imaging conditions by focusing on generalizable feature extraction. Alongside this framework, I introduced CryoETSim, an extensive simulated dataset designed to replicate the crowdedness and variability observed in real cryo-ET samples. CryoETSim provides a comprehensive foundation for model training and evaluation, bridging the gap between simulation and experimental datasets. CryoSiam and CryoETSim offer a practical and computationally efficient framework suitable for high-throughput cryo-ET workflows, enhancing the quality and scalability of segmentation processes.

This thesis addresses several pressing research questions to evaluate and extend the potential of SSL for cryo-ET segmentation. The first question focuses on understanding how SSL can improve segmentation performance on

cryo-ET data. Specifically, it examines the ability of SSL techniques to extract meaningful representations from noisy and artifact-laden tomograms, thereby enhancing segmentation accuracy while minimizing the need for annotated datasets. Another important question investigates the critical factors influencing the generalizability of segmentation models across diverse cryo-ET datasets. Given the variability in sample preparation, imaging conditions, and noise levels, exploring strategies that improve cross-dataset generalization is essential to ensure robust and consistent results. Finally, the thesis examines whether segmentation models trained solely on simulated data can effectively apply to experimental tomograms without requiring extensive manual annotations. This question focuses on the utility of CryoETSim, which explores how simulated datasets can bridge the gap between simulation and real-world data, enabling models to generalize effectively to experimental tomograms.

Through these objectives and research questions, this thesis aims to solve the challenges associated with cryo-ET segmentation and analysis comprehensively. By developing robust SSL approaches, creating realistic simulated datasets, and addressing critical gaps in generalizability and computational efficiency, I seek to advance cryo-ET analysis workflows and unlock the full potential of this transformative imaging technique.

1.6 THESIS CONTRIBUTIONS AND STRUCTURE

This thesis significantly contributes to cryo-ET data analysis, SSL, and segmentation methodologies. Addressing persistent challenges in segmentation and particle identification introduces innovative tools and approaches that enhance cryo-ET workflows' accuracy, scalability, and efficiency, advancing the boundaries of structural biology.

The first significant contribution is the development of CryoSiam, a novel SSL framework designed to address cryo-ET data analysis's unique challenges. CryoSiam integrates semantic segmentation, instance segmentation, and particle identification into a unified pipeline, offering a comprehensive solution for tomogram analysis. The framework minimizes dependence on annotated datasets by utilizing tailored transformations and contrastive learning techniques while delivering robust, generalizable embeddings. CryoSiam sets a new standard for leveraging SSL to overcome the limitations of existing segmentation methods.

The second key contribution is the creation of CryoETSim, a simulated dataset carefully designed to emulate real cryo-ET data. Comprising 400 tomograms, CryoETSim introduces realistic variations in defocus, sample thickness, and structural diversity, factors often missing in existing datasets. I provided detailed ground truth annotations within CryoETSim, facilitating rigorous evaluation of segmentation and particle identification techniques. CryoETSim is a foundational resource for training and validating advanced segmentation models by bridging the gap between simulated and experimental data domains.

The third contribution lies in the comprehensive evaluation of the CryoSiam framework, which I conducted across key segmentation and particle identification tasks. Through detailed ablation studies, I investigated the impact

of specific design elements, including transformations that work for cryo-ET data, providing critical insights into the framework's methodology. Performance metrics such as DICE score, average precision, and F1 score demonstrated CryoSiam's robustness and accuracy. Therefore, this thesis highlights the transformative potential of SSL, particularly in domains where annotated data is scarce or prohibitively expensive to generate.

Collectively, these contributions address critical gaps in generalizability, annotation efficiency, and scalability in cryo-ET analysis. By providing a more efficient, accurate, and accessible segmentation solution, this research establishes a foundation for future advancements in structural biology, enabling more profound insights into the molecular architecture of complex biological systems.

The remainder of this thesis is organized to guide the reader through the challenges, methodologies, results, and broader implications of this research. Chapter 2 explores the unique characteristics of cryo-ET data and reviews existing segmentation and analysis methods, identifying their strengths and limitations. It also introduces SSL as a promising approach to address critical challenges, including limited annotated datasets and the variability inherent in cryo-ET data.

Building on this foundation, Chapter 3 details the design and theoretical underpinnings of CryoSiam. I describe the construction and design principles behind the CryoETSim dataset, which was developed to resemble real cryo-ET data closely. Additionally, this chapter outlines the training strategies and self-supervised learning techniques applied within CryoSiam, along with the evaluation metrics used to assess its performance.

Chapter 4 presents a comprehensive analysis of CryoSiam's performance on segmentation and particle identification tasks using simulated data. I report the findings from extensive ablation studies, which assess the influence of various design choices, including data transformations, overlapping subtomograms, and loss functions. The results, measured through metrics such as DICE score, average precision, and F1 score, demonstrate the framework's robustness and accuracy across different experimental conditions.

In Chapter 5, the focus shifts to real experimental cryo-ET datasets. This chapter evaluates CryoSiam's effectiveness in practical biological scenarios, showcasing its ability to generalize across datasets, handle noise and artifacts, and adapt to varying imaging conditions. The results highlight CryoSiam's potential as a scalable and efficient solution for cryo-ET workflows in real-world applications.

Chapter 6 reflects on the broader implications of this research, discussing the key contributions and impact of CryoSiam and CryoETSim on cryo-ET analysis. I critically evaluate the limitations of the current approach and propose directions for further improving segmentation methodologies. This chapter also situates the findings within the broader field of structural biology, emphasizing how these advancements contribute to ongoing research efforts.

Finally, Chapter 7 concludes the thesis by summarizing the key findings and contributions. It outlines future opportunities for research, including refining the CryoSiam framework, extending the CryoETSim dataset to include additional structural features, and exploring new applications of self-supervised

learning in cryo-ET workflows. These future directions aim to enhance the scalability and accuracy of cryo-ET data analysis, paving the way for continued breakthroughs in structural biology.

2 FRAMING THE RESEARCH CHALLENGE

The analysis of cryo-ET data presents a transformative opportunity to advance our understanding of macromolecular structures and cellular organization. As discussed in Chapter 1, cryo-ET enables the visualization of biological specimens in near-native states, providing unparalleled insights into their structural and spatial characteristics [91, 92]. However, extracting meaningful information from cryo-ET tomograms remains a difficult challenge due to inherent limitations such as noisy data, missing wedge artifacts, and variability in sample preparation and imaging conditions [44, 57, 93]. These challenges are further aggravated by the scarcity of annotated datasets and the absence of generalizable segmentation methods, which hinder the development of robust computational tools for cryo-ET analysis [50, 78].

This chapter delves into the specific challenges that motivate this research, providing a detailed examination of cryo-ET data's unique characteristics and the shortcomings of existing segmentation approaches. By situating these issues within the broader context of cryo-ET analysis, this chapter underscores the critical need for innovative solutions to address the inherent complexities of cryo-ET data.

A central focus of this discussion is the potential of SSL as a transformative paradigm for cryo-ET segmentation. SSL offers the ability to leverage large datasets of unlabeled data, enabling the development of robust and adaptable models to diverse datasets [94]. By addressing the limitations of traditional supervised approaches, SSL presents a pathway to scalable, efficient, and accurate segmentation methods, aligning with the overarching goals of this research.

The chapter begins by examining the unique characteristics of cryo-ET data, including the factors contributing to noise, artifacts, and variability. It then reviews existing segmentation methods, critically evaluating their strengths, limitations, and applicability to cryo-ET datasets. Finally, it introduces SSL as a promising solution to the challenges of cryo-ET segmentation, laying the foundation for developing the CryoSiam framework presented in subsequent chapters.

2.1 OVERVIEW OF CRYO-ET DATA CHARACTERISTICS

The creation of cryo-ET data begins with the meticulous preparation of biological specimens, a critical step to preserve their native structures and enable meaningful analysis. Unlike traditional imaging methods, cryo-ET requires samples to be frozen rapidly and maintained in a vitreous state to prevent the formation of ice crystals, which could damage or distort the biological structures [18, 23]. This process, known as vitrification, is accomplished by plunging the specimen into liquid ethane cooled by liquid nitrogen [48]. The rapid freezing ensures that water molecules do not rearrange into crystalline ice, in-

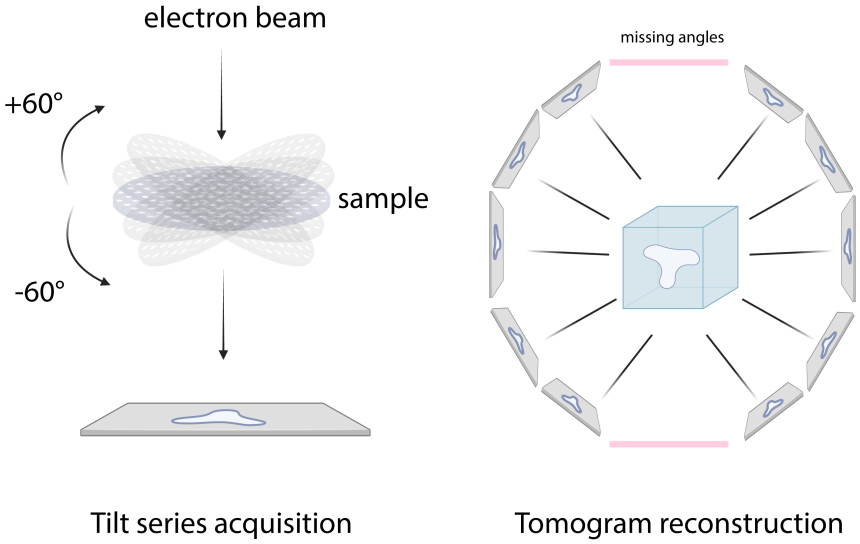

Figure 2: **Tilt-series acquisition and tomogram reconstruction in cryo-ET.** The sample is incrementally tilted along a single axis, typically ranging from -60°to +60°, while an electron beam captures 2D projection images at each angle. These projections form the tilt series, which is then used to reconstruct a 3D tomogram through computational methods. The figure illustrates the critical process of acquiring structural information from multiple perspectives to enable 3D visualization. However, tilting range and electron exposure limitations introduce artifacts such as the missing wedge, posing challenges for accurate reconstruction and analysis. Figure partially created using BioRender.

stead forming an amorphous solid that preserves the sample's ultrastructural integrity [95]. Vitrification effectively locks the specimen in its native state, providing the fidelity necessary for accurate imaging and subsequent analysis [96].

For thicker samples, such as intact cells or tissue sections, focused ion beam (FIB) milling is employed to create thin lamellae suitable for transmission electron microscopy (TEM) [97]. This technique uses a precisely controlled ion beam to remove excess material, leaving an electron-transparent layer that retains the biological context of the original sample [98]. FIB milling is particularly valuable for studying subcellular structures in their native environments, as it allows researchers to target specific regions of interest while preserving spatial relationships within the sample [99].

These specimen preparation methods, including vitrification and FIB milling, are integral to the cryo-ET workflow. They ensure that biological specimens are preserved near-native, free from artifacts introduced by dehydration or chemical fixation [48]. By maintaining structural integrity and minimizing artifacts, these techniques lay the foundation for high-resolution imaging and enable meaningful downstream analysis of cryo-ET data [48, 98].

After sample preparation, tilt-series acquisition is the next critical step in creating cryo-ET data. Figure 2 illustrates the process of tilt-series acquisition

and tomogram reconstruction in cryo-ET. This process involves imaging the sample at multiple tilt angles using a TEM [19]. The specimen, mounted on a specialized stage, is incrementally tilted along a single axis, typically covering a range from approximately -60°to +60°. However, the exact range may vary depending on the experimental setup [25]. At each tilt angle, a 2D projection image is captured, and these images collectively form the tilt series, representing angularly sampled projections of the specimen [100].

Tilt-series acquisition is crucial because it enables the capture of structural information from multiple perspectives. Each projection provides a unique angular view of the sample, allowing for the reconstruction of a 3D representation of the specimen [100]. This process is guided by computed tomography principles, where angular information is essential for resolving structural details in 3D. However, some angular information is inevitably lost due to the limitations of the tilting range and the need to minimize electron exposure to preserve sample integrity. This loss introduces artifacts such as the missing wedge, which affects the resolution and completeness of the reconstructed volume, as discussed in Chapter 1.

Before a tomogram can be reconstructed, the tilt-series data undergoes several preprocessing steps to ensure accuracy and quality. The first step is motion correction, compensating for beam-induced sample movements that can blur the projection images [101]. Algorithms such as MotionCor2 and Unblur are widely employed, offering robust correction capabilities that preserve high-resolution structural details [102, 103]. These tools track the movement of the sample during imaging and adjust the projections accordingly, significantly improving alignment precision and image clarity [102].

Following motion correction, the tilt series is aligned to ensure that all projections are consistently registered. This alignment is achieved by identifying standard features across projections, such as fiducial markers (e.g., gold nanoparticles) embedded in the sample [100]. Software tools like IMOD and Protomo are commonly used for fiducial-based alignment, while markerless alignment algorithms, such as those implemented in Warp or Aretomo, rely on intrinsic sample features for registration [104–108]. These alignment methods are critical for correcting shifts, rotations, and deformations between projections, ensuring that the tilt series accurately reflects the authentic spatial information within the sample [109].

Another essential preprocessing step is correcting contrast transfer function (CTF) modulation. The CTF describes the imaging system's influence on the contrast of features in the projection images, including phase and amplitude alterations introduced by the microscope's optics [110]. CTF effects can degrade the quality of reconstructions by distorting spatial frequencies, making correction crucial for preserving structural accuracy [110]. CTFFIND4, GCTF, and NovaCTF are widely used to estimate and correct for CTF modulation in individual tilt images [111–113]. These corrections compensate for defocus variations across projections, ensuring that the reconstructed tomogram accurately represents the accurate ultrastructural details of the sample [113].

Together, these preprocessing steps of motion correction, tilt-series alignment, and CTF modulation correction are integral to generating high-quality tomograms. By addressing the artifacts introduced during imaging, these pro-

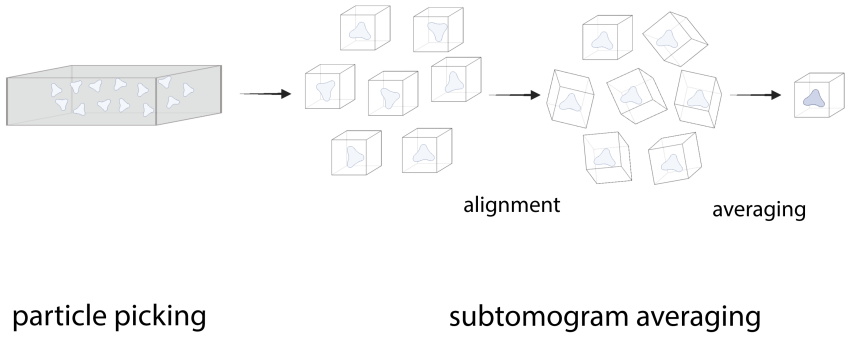

Figure 3: **The subtomogram averaging workflow in cryo-ET.** The process starts with particle picking, where individual subtomograms are extracted from the tomogram. These subtomograms are then aligned to maximize structural consistency across the dataset. Finally, the aligned subtomograms are averaged to reduce noise and enhance resolution, enabling detailed structural insights into macromolecular complexes. Image partially created using BioRender.

cesses ensure that the reconstructed 3D volume reflects the native spatial organization of the biological specimen with minimal distortions [39].

With the tilt series preprocessed and aligned, tomogram reconstruction begins. Weighted back projection (WBP) is the most widely used reconstruction algorithm, projecting the aligned 2D images back along their original angular paths to generate a 3D volume [11]. Tools like Tomo3D and IMOD efficiently implement WBP, but its susceptibility to noise and missing wedge artifacts has led to the development of iterative algorithms, such as the Simultaneous Iterative Reconstruction Technique (SIRT) [39, 114]. While these methods, supported by tools like Aretomo, improve reconstruction accuracy, they are computationally intensive [113].

Integrating tilt-series acquisition, preprocessing, and reconstruction transforms raw 2D projections into high-resolution 3D tomograms, enabling the detailed visualization of macromolecular structures in their native environments. These foundational steps are critical for applying advanced segmentation and particle identification methods, which remain central challenges in cryo-ET analysis.

Subtomogram averaging (STA) is a pivotal computational technique in cryo-ET that enhances the resolution of macromolecular structures by combining information from identical or similar particles within tomograms. The process begins with particle picking, where subtomograms, small 3D volumes containing individual macromolecules or complexes, are extracted from the reconstructed tomogram. Accurate particle picking is essential because only particles of the same type can be effectively aligned and averaged to retrieve high-resolution structures [115, 116].

Figure 3 illustrates the key steps involved in the STA workflow. After picking particles, the subtomograms undergo an iterative alignment process to ensure structural consistency across the dataset. This alignment step compensates for particle orientation and positioning differences within the tomogram. Once aligned, the subtomograms are averaged to reduce noise and amplify the sig-

nal, revealing detailed structural features otherwise obscured due to the low signal-to-noise ratio inherent in cryo-ET data [117].

STA is particularly effective for studying symmetrical or repetitive macromolecular assemblies such as ribosomes, viral capsids, and cytoskeletal filaments. These assemblies benefit from the averaging process, which enhances structural detail by leveraging the consistency among multiple particles [118]. Advanced tools such as Dynamo, RELION, and EMAN2 facilitate STA workflows by providing robust particle alignment and classification algorithms to optimize resolution [116, 119, 120]. These platforms ensure precise orientation and positioning of subtomograms during averaging, which is critical for achieving high-resolution reconstructions.

Despite its effectiveness, STA has limitations. It assumes structural homogeneity among the averaged particles, which may not hold for highly dynamic or heterogeneous assemblies. Additionally, STA requires sufficient subtomograms to achieve meaningful resolution improvements, making it less effective for sparse datasets or rare particle types [121]. Furthermore, the accuracy of the final structure depends heavily on the initial particle-picking step, highlighting the importance of reliable particle detection methods in the cryo-ET pipeline [64].

STA is critical in cryo-ET analysis, complementing segmentation and particle identification workflows. By leveraging STA, researchers can transform noisy, low-resolution tomograms into detailed structural insights, uncovering the intricate organization of biological systems at macromolecular resolution [122].

Although cryo-ET data offers unparalleled insights into the 3D organization of biological specimens, its inherent complexity poses significant challenges for analysis. These challenges arise from technical constraints in data acquisition, physical properties of biological samples, and variability introduced during imaging. Addressing these issues is essential for enabling accurate segmentation and advancing cryo-ET workflows.

Cryo-ET data's low SNR remains one of its most pervasive challenges [123]. Electron dosage is strictly limited to minimize radiation damage, introducing significant noise into the projection images. Thick specimens exacerbate this issue by causing electron scattering, further reducing contrast [124]. Advanced denoising methods are essential to enhance feature visibility while preserving the integrity of structural details [45, 125].

Another critical limitation is the "missing wedge" artifact, caused by the incomplete angular coverage during tilt-series acquisition [52]. This artifact introduces an anisotropic resolution in the reconstructed tomogram, distorting features along the missing angles. Despite iterative reconstruction and regularization techniques, the missing wedge remains a persistent challenge [53, 54, 126].

Biological heterogeneity further complicates analysis. Variability in sample preparation, ice thickness, and imaging conditions leads to inconsistencies in contrast and resolution [57]. Additionally, cellular specimens' crowded, dynamic environments introduce overlapping features, complicating segmentation efforts [127]. These challenges highlight the need for adaptable computational methods that account for biological and technical variability.

The combined effects of noise, missing wedge artifacts, and heterogeneity underscore the complexity of cryo-ET data analysis. These challenges necessitate robust computational approaches to mitigate artifacts, enhance feature visibility, and accommodate diverse datasets. This thesis addresses these challenges by focusing on self-supervised learning and innovative segmentation frameworks, offering scalable and generalizable solutions for cryo-ET workflows.

2.2 EXISTING METHODS FOR SEGMENTATION OF CRYO-ET DATA

Cryo-ET data segmentation is crucial in unlocking the biological insights embedded within tomograms. Segmentation involves identifying and delineating regions of interest within a 3D volume, such as macromolecular complexes, cellular structures, or particles [10]. This process enables detailed analysis and interpretation of biological specimens' structural and spatial organization. Over time, a range of segmentation approaches has been developed to address the unique challenges of cryo-ET data, each offering distinct strengths and limitations [23].

Manual segmentation, one of the earliest methods employed, relies on human expertise to annotate features within tomograms. This approach provides high precision but is time-intensive, laborious, and prone to variability across annotators, making it unsuitable for high-throughput workflows [61]. Template-matching techniques have been widely used to address these limitations and automate particle detection and segmentation. These methods compare regions in the tomogram with predefined templates, calculating cross-correlation scores to identify features of interest [128]. While template matching can achieve high accuracy for well-defined structures, it relies heavily on the availability of representative templates [129]. This reliance makes it less practical for identifying novel or highly variable macromolecular features, and the computational cost of exhaustive angular searches further limits its scalability [38].

Deep learning (DL) revolutionized cryo-ET segmentation by introducing methods capable of leveraging large datasets and advanced computational algorithms to achieve higher accuracy and scalability [50]. Supervised deep learning approaches, such as convolutional neural networks (CNNs) and U-Net architectures, demonstrated significant promise [58, 60]. However, their reliance on extensively annotated datasets remains a critical bottleneck [78]. This dependence on labeled data is particularly challenging in cryo-ET, where manual annotation is labor-intensive and annotated datasets are scarce [78].

More recently, self-supervised learning (SSL) and unsupervised approaches have gained traction as potential solutions to overcome the limitations of annotated data availability [14, 37, 67]. These methods enable models to learn meaningful representations directly from the data, bypassing the need for exhaustive manual labeling [67]. SSL-based frameworks have shown notable success in addressing the variability and heterogeneity of cryo-ET data by focusing on learning generalizable features across diverse datasets [14, 65].

Despite these advancements, the complexity of cryo-ET data continues to pose significant challenges to segmentation methods. Manual approaches of-

fer precision but lack scalability, template matching struggles with adaptability and computational efficiency, and even advanced DL techniques face generalizability and data requirements limitations. This section examines these methods in detail, discussing their strengths and weaknesses to provide a comprehensive understanding of the current landscape of cryo-ET segmentation.

2.2.1 *Manual segmentation*

Manual segmentation is a traditional yet foundational approach in cryo-ET data analysis. This method relies on domain experts to manually annotate features of interest within tomograms, often using visualization tools to trace boundaries and classify regions based on their knowledge of biological structures [130–132]. The process is typically performed slice by slice, requiring meticulous attention to detail to ensure accuracy.

The primary strength of manual segmentation lies in its precision and adaptability. Expert annotators can identify subtle structural features and resolve ambiguities that automated methods might overlook, making it particularly valuable for complex or poorly understood datasets where computational tools may be insufficient. However, this reliance on human expertise introduces significant limitations. Annotating cryo-ET data is labor-intensive and time-consuming, often requiring days or weeks to process a single dataset [78].

Moreover, the subjective nature of manual annotation can lead to inconsistencies, as different experts may interpret the same data in varying ways [61]. Such variability can affect the reproducibility of results and complicate downstream analyses, especially when annotations from multiple sources are combined. Additionally, the scalability of manual segmentation is inherently limited, rendering it impractical for high-throughput studies or the analysis of large-scale datasets [78].

While manual segmentation plays a critical role in cryo-ET analysis, its limitations highlight the pressing need for automated methods that can replicate the precision of expert annotations while offering improved efficiency and consistency. Subsequent sections will explore the evolution of segmentation methodologies, focusing on how template matching and DL-based approaches have sought to overcome these challenges and advance cryo-ET analysis toward greater scalability and accuracy.

2.2.2 *Template matching approaches*

Template matching (TM) is one of cryo-ET's most widely used computational approaches for particle picking. This method involves scanning the tomogram with a predefined template that resembles the target particle to identify regions of high similarity [86, 133]. The process typically includes generating a reference template, performing exhaustive searches across various orientations and positions, and calculating cross-correlation scores to detect potential particle locations [134]. When an accurate template is available, template matching can achieve high specificity, effectively automating particle-picking and identifying particles that closely resemble the reference [135].

Despite its strengths, TM has several critical limitations that restrict its scalability and generalizability. A major challenge is its reliance on an existing template, which requires prior knowledge of the particle's structure [133]. Generating an accurate template for novel or heterogeneous specimens can be difficult or even impossible [20]. Additionally, the method requires computationally intensive angular searches to account for all possible particle orientations within the tomogram [136]. This exhaustive process demands substantial computational resources and time, making it impractical for analyzing large datasets or conducting high-throughput studies [38]. Furthermore, the inherent noise and artifacts in cryo-ET data, such as the missing wedge, can reduce TM accuracy by introducing false positives or missing particles [137].

Several software tools have been developed to implement TM, each offering varying efficiency and functionality. PyTom, a comprehensive toolbox for cryo-ET processing, includes modules for TM, subtomogram alignment, and subtomogram classification, with GPU acceleration to enhance computational performance [138, 139]. PyTME (Python Template Matching Engine) optimizes TM for large datasets, utilizing both CPUs and GPUs for faster processing [140]. STOPGAP, another widely used tool, integrates TM with subtomogram alignment and classification, making it particularly suitable for complex biological specimens [141]. While these tools have advanced particle-picking workflows, their reliance on predefined templates and computational intensity remains a significant limitation [64].

Although TM has been pivotal in developing particle-picking techniques, its constraints have driven the exploration of alternative approaches. Recent advances in DL, particularly in self-supervised and unsupervised learning, aim to overcome the reliance on predefined templates and address the computational inefficiencies of traditional methods [14]. These emerging techniques promise more flexible and scalable solutions for particle picking and segmentation, representing a shift beyond the limitations of TM approaches.

2.2.3 *Supervised deep learning approaches*

Supervised DL has emerged as a transformative approach for segmenting cryo-ET data, automating the identification and delineation of macromolecular structures within tomograms [142]. These methods utilize large annotated datasets to train models capable of recognizing complex patterns and features inherent in cryo-ET images [58, 60].

One prominent example is the work by Liu et al. [142], which introduced a 3D CNN inspired by Fully Convolutional Networks [143] and encoder-decoder architectures [144] for the supervised segmentation of macromolecules in subtomograms. This approach demonstrated significant improvements in segmentation accuracy compared to baseline methods, highlighting the potential of supervised learning for cryo-ET data analysis [142].

Another key contribution is DeepFinder, a deep learning-based pipeline that streamlines the training and application of 3D autoencoder models for particle picking in cryo-ET experiments [60]. This framework provides researchers with tools to accurately identify particles within cryo-ET datasets' dense and intricate landscapes, significantly enhancing efficiency and precision [60].

Similarly, DeePiCt (Deep Picker in Context) incorporates a dual-network approach, combining a 2D CNN for segmenting cellular compartments with a 3D CNN for particle localization and structural segmentation [58]. This integration enhances particle detection and structural analysis accuracy and efficiency in cryo-ET datasets, demonstrating the versatility of supervised deep learning models [58].

Despite significant advances, supervised DL methods face critical challenges, primarily the need for large annotated datasets. Generating such datasets is labor-intensive and time-consuming, requiring domain expertise, particularly in cryo-ET [78]. This scarcity of annotated data limits the development of robust models that can generalize effectively across diverse datasets.

Supervised models are also susceptible to overfitting, especially when trained on a single dataset. By capturing noise or dataset-specific features, these models often struggle to generalize to new, unseen datasets [145]. The inherent heterogeneity of cryo-ET data, including variability in sample preparation, imaging conditions, and structural diversity, further exacerbates these limitations [146].

2.2.4 *Unsupervised and self-supervised deep learning approaches*

Semi-automated and fully automated approaches leveraging unsupervised and self-supervised learning have emerged as transformative alternatives to traditional supervised methods. These advanced methodologies learn directly from the data, eliminating the need for extensive manual annotations and enabling more scalable and adaptable segmentation workflows [65].

A notable example is TomoTwin, a deep metric learning-based method for particle picking in cryo-ET. By pretraining on a dataset of over 120 proteins, TomoTwin can identify and localize macromolecules in tomograms without requiring manual annotations or retraining for new datasets [65]. This capability allows for de novo particle identification, making TomoTwin adaptable and efficient across various biological samples, including those with complex or previously uncharacterized structures [65].

Another key advancement is Vox-UDA (Voxel-wise Unsupervised Domain Adaptation), which addresses the domain shift between simulated and real cryo-ET data. By incorporating a noise generation module to simulate target-like noise in source datasets and leveraging denoised pseudo-labeling, Vox-UDA aligns source and target domains [147]. This method improves segmentation performance on real-world tomograms, reducing the dependency on annotated datasets and enhancing cross-domain generalizability [147].

The CryoSAM framework takes a prompt-based, training-free approach to tomogram segmentation. Using existing 2D foundation models, CryoSAM bridges the gap between 2D and 3D segmentation tasks, allowing users to segment particles of specific categories with minimal input [66]. This flexibility makes CryoSAM particularly valuable for analyzing highly heterogeneous datasets, where manual annotations are often infeasible [66].

MiLoPYP represents a cutting-edge, unsupervised framework explicitly tailored for cryo-ET segmentation. Utilizing contrastive learning, MiLoPYP enables fast molecular pattern mining and precise protein localization with min-

imal manual annotation [67]. This approach excels in detecting and localizing a broad range of targets, from globular and tubular complexes to large membrane proteins, effectively streamlining workflows for high-resolution in situ structural determination [67]. By leveraging SSL and clustering techniques, MiLoPYP efficiently handles crowded and sparse regions within tomograms, offering a robust solution for analyzing diverse cryo-ET datasets [67].

These innovative methods address the key limitations of traditional supervised approaches by reducing reliance on annotated datasets, enabling generalization across diverse tomograms, and accommodating varying noise levels and imaging conditions. Techniques like TomoTwin, Vox-UDA, CryoSAM, and MiLoPYP improve segmentation performance while expanding the adaptability of segmentation pipelines [66].

The versatility of these approaches ensures their applicability to datasets with complex and heterogeneous structural landscapes. By leveraging the intrinsic structure of cryo-ET data, these methods significantly enhance segmentation accuracy and scalability, marking a paradigm shift in cryo-ET data analysis [148].

In summary, semi-automated and fully automated segmentation methods based on unsupervised and self-supervised learning significantly advance cryo-ET analysis. By providing scalable, robust, and efficient workflows, these approaches are helping to unlock more profound insights into the molecular architecture of biological systems, pushing the boundaries of what is achievable with cryo-ET data.

2.3 DENOISING METHODS FOR CRYO-ET DATA

Denoising is a pivotal step in cryo-ET, critical for enhancing the SNR and improving the interpretability of subcellular structures. The inherently noisy nature of cryo-ET data, stemming from the low electron doses required to preserve biological specimens, presents significant challenges for visualization and downstream analysis [125].

Cryo-CARE (Content-Aware Image Restoration) is a deep learning-based denoising approach that employs supervised learning to train neural networks on paired datasets of low-dose noisy images and high-dose counterparts [45, 149]. This method significantly enhances contrast and SNR, facilitating visual interpretation and improving the performance of automated tasks such as dense segmentation [45].

IsoNet expands denoising capabilities by addressing missing wedge artifacts in addition to noise reduction. Using a self-supervised framework, IsoNet compensates for missing angular information, restoring isotropic resolution and enabling the visualization of macromolecular structures with improved information [54]. This dual capability marks a significant advancement in tomogram restoration [54].

DeepDeWedge integrates denoising and missing wedge correction into a unified self-supervised learning framework. Unlike sequential approaches, this method performs both tasks simultaneously without requiring ground truth data, achieving superior performance compared to existing tech-

niques [126]. By addressing these challenges within a single framework, DeepDeWedge streamlines tomogram restoration [126].

CryoSamba offers a volumetric denoising solution using SSL. By employing DL interpolation to average motion-compensated neighboring planes, CryoSamba enhances coherent signals while suppressing high-frequency noise [150]. Operating directly on 3D volumes without needing pre-recorded images or synthetic datasets, CryoSamba improves contrast and SNR, making it invaluable for high-quality structural analysis [150].

Integrating these advanced denoising methods has significantly improved the quality of tomographic reconstructions, addressing critical challenges such as noise and missing wedge artifacts. By enhancing SNR and mitigating acquisition-related artifacts, these techniques enable more precise segmentation and particle identification, advancing cryo-ET workflows and the understanding of molecular architectures [151].

Cryo-CARE, IsoNet, DeepDeWedge, and CryoSamba represent a transformative leap in cryo-ET data processing [150]. These approaches reduce noise, correct artifacts, and provide a robust foundation for downstream tasks such as segmentation and particle picking, driving forward the field of structural biology [42].

2.4 SELF-SUPERVISED LEARNING IN COMPUTER VISION

Self-supervised learning (SSL) has emerged as a transformative DL paradigm, addressing data scarcity and variability challenges by enabling models to learn from raw, unlabeled data through carefully designed pretext tasks [152]. This approach circumvents the need for extensively annotated datasets, making it particularly valuable for fields like cryo-ET, where generating labeled data is labor-intensive, costly, and time-consuming [14].

At its core, SSL leverages the inherent structure of raw data to define auxiliary or "pretext" tasks, guiding models to learn meaningful representations. These tasks exploit intrinsic data features, such as spatial relationships, color distributions, or temporal correlations, enabling models to acquire generalizable embeddings that can be fine-tuned for downstream tasks like classification, detection, or segmentation [69]. For instance, pretext tasks like predicting the relative positions of image patches, reconstructing missing parts of images, or determining image rotations encourage models to focus on invariant features, improving robustness to noise and transformations [69, 153, 154]. By learning structural patterns and invariant properties, SSL-trained models provide a strong foundation for applications requiring a high-level understanding of complex datasets.

Several groundbreaking SSL methods have demonstrated remarkable success in representation learning, advancing the potential for segmentation tasks. SimCLR trains models to maximize agreement between augmented views of the same image by leveraging contrastive loss to learn invariant features across transformations [155]. MoCo introduces a momentum encoder, stabilizing representation learning and reducing batch size requirements, improving performance on large datasets [16, 156]. BYOL eliminates the need for negative samples using a student-teacher architecture, simplifying the training process

while focusing entirely on learning invariant representations [15]. SimSiam builds upon these concepts with an even more streamlined approach, combining elements of SimCLR and BYOL to achieve efficient training without sacrificing quality [157].

DINOv2, a recent advancement, introduces self-distillation with no labels, leveraging a vision transformer (ViT) architecture for SSL [158]. DINOv2 generates high-quality representations that generalize well across datasets and tasks by focusing on feature clustering and density estimation. Its transformer-based design and ability to scale across modalities make it particularly relevant for handling cryo-ET data's complex, heterogeneous nature. These features position DINOv2 as a promising tool for segmentation tasks where robustness and transferability are critical [158].

One of SSL's key strengths is the generalizability of its learned embeddings. Unlike supervised models that often overfit to training data, SSL captures broader patterns and invariant features that are transferable across tasks and datasets [159]. This generalization ability is particularly valuable for cryo-ET, where variability in imaging conditions, noise levels, and sample characteristics presents significant challenges for segmentation [14]. Models trained using SSL on unlabeled datasets can be fine-tuned for segmentation tasks with minimal labeled data, enabling efficient and scalable analysis of cryo-ET tomograms. This adaptability is especially advantageous for addressing the heterogeneity of cryo-ET data, where imaging conditions and structural variability often vary significantly between datasets [57].

In the context of cryo-ET segmentation, SSL has the potential to address challenges like noise, missing wedge artifacts, and limited annotations. Pretext tasks tailored to cryo-ET data, such as reconstructing missing wedge information or predicting spatial relationships within volumetric data, provide models with a robust foundation for downstream analysis. These representations enable segmentation workflows that are more accurate and less dependent on resource-intensive manual annotations [14, 65, 67]. Additionally, SSL has been shown to improve model performance on small or sparse datasets by leveraging the data's inherent structure, further reducing the reliance on annotated datasets [160].

By addressing the unique challenges of cryo-ET, SSL offers a promising pathway for advancing segmentation workflows. Its capacity to extract meaningful, transferable representations from raw data while minimizing dependency on annotations aligns closely with the requirements of cryo-ET analysis. This alignment makes SSL a transformative approach for enabling scalable, accurate, and efficient segmentation of cryo-ET tomograms.

In summary, SSL represents a significant leap forward in cryo-ET segmentation, offering innovative solutions to the field's most pressing challenges. By reducing annotation dependency, improving generalizability, and enhancing scalability, SSL can revolutionize cryo-ET workflows and unlock more profound insights into the molecular architecture of biological systems.

Part II

DESIGNING THE FRAMEWORK: FROM THEORY TO SIMULATED EXPERIMENTS

This part addresses the foundational aspects of the research, encompassing the theoretical framework, the design of the proposed pipeline, and its evaluation using simulated cryo-ET data. Simulated data provides a controlled environment for rigorously testing and refining the methodology. Key components and configurations contributing to effective segmentation performance are identified through ablation studies and detailed analysis. These findings establish the groundwork for subsequent validation of real cryo-ET data.

3

DEVELOPMENT OF THE METHOD

The development of the CryoSiam framework is central to this thesis, addressing multiple critical challenges in cryo-ET data analysis through an integrated, multi-task learning approach. CryoSiam was designed to perform semantic segmentation, instance segmentation, and particle identification, three tasks providing a comprehensive pipeline for interpreting tomograms. Semantic segmentation enables labeling at the voxel level to distinguish between structural components such as membranes, actin filaments, microtubules, and particles. Instance segmentation builds on this, allowing for the identification and separation of individual particle instances. Particle identification focuses on accurately localizing and classifying specific particles, a crucial task in understanding the spatial organization of macromolecular complexes. Together, these tasks address the inherent complexities of cryo-ET data, enabling a detailed and reliable analysis pipeline.

To streamline the development and validation of the CryoSiam framework, a new dataset called CryoETSim was created. This dataset consists of 400 simulated tomograms with structural diversity and realistic variations in imaging conditions, such as defocus levels and sample thickness. These variations replicate the visual modulation and contrast range found in experimental cryo-ET data, providing a more robust and representative training and evaluation environment. CryoETSim also includes comprehensive ground-truth annotations, enabling precise model performance evaluation across multiple tasks. By incorporating such variability and annotations, CryoETSim is an essential resource for developing models capable of handling the complexities of real-world cryo-ET data.

CryoSiam's training incorporates an SSL approach that leverages the structural characteristics of cryo-ET data to build robust feature representations without relying on annotated labels. This training strategy focuses on learning meaningful voxel-level embeddings through SSL, which utilizes augmented views of the same data to maximize similarity in the latent space. The training process reflects variations commonly observed in cryo-ET tomograms by employing transformations such as Gaussian noise, high- and low-pass filtering, and masking. This approach ensures that the CryoSiam framework is initialized with strong, generalized embeddings before fine-tuning for specific downstream tasks, including segmentation and particle identification.

The following sections detail the critical components of the CryoSiam framework. Section 3.1 introduces the design principles and architecture of the model. Section 3.2 discusses the theoretical framework underpinning the method, including the image transformations, downstream tasks, and the contrastive learning strategies utilized. Section 3.3 focuses on creating and utilizing the CryoETSim dataset, while Section 3.4 outlines the training strategy, including optimization techniques and hyperparameter choices. Finally, Section 3.5 describes the evaluation metrics used to assess the performance of CryoSiam across its downstream tasks.

3.1 MODEL DESIGN

The CryoSiam (CRYO-electron tomography Simple Siamese Networks) framework was developed to address critical challenges in cryo-ET data analysis by providing a dual-level solution that combines voxel-level precision segmentation with particle identification for subtomograms. These two capabilities enable detailed structural analysis while distinguishing individual particle identities, making CryoSiam a comprehensive tool for cryo-ET analysis workflows. The overall design of the CryoSiam framework is depicted in Figure 4.

CryoSiam consists of two core components: DenseSimSiam (DENSE simple SIAMese network) and SimSiam (SIMple SIAMese network). The first component, DenseSimSiam, is pre-trained on a self-supervised pretext task and fine-tuned on three downstream tasks: denoising, semantic segmentation, and instance segmentation. Each of these tasks addresses specific challenges in analyzing cryo-ET data. Denoising removes noise from tomograms while preserving structural integrity, ensuring high-quality input for subsequent processing. Semantic segmentation provides voxel-level predictions to identify and separate key structural elements within tomograms, such as membranes, microtubules, actin filaments, and particles. This task offers a detailed understanding of the tomogram's composition, enabling further exploration of its structural components. Building upon this, instance segmentation separates individual particle instances, generating particle centers and their corresponding masks. These masks delineate the boundaries and shapes of each particle with high accuracy, allowing for precise structural characterization.

Once particle masks are generated through instance segmentation, the second component of CryoSiam, SimSiam, is utilized. Like DenseSimSiam, SimSiam is pre-trained on a self-supervised pretext task but specifically designed to process subtomograms. This component learns embeddings that capture detailed features and structural information of subtomograms in a compact, lower-dimensional space. These embeddings are then used to separate particle identities within the embedded space, enabling the framework to distinguish distinct particle classes. This capability is critical for analyzing heterogeneous datasets containing diverse particle types.

DenseSimSiam and SimSiam form a unified framework that integrates voxel-level precision with subtomogram-level classification. This design addresses structural segmentation and particle identification challenges, providing a versatile and robust solution for cryo-ET data analysis. CryoSiam advances the accuracy and applicability of computational approaches in this field by combining segmentation and identification within a single framework. The depiction of the CryoSiam framework in Figure 4 illustrates its structure and highlights the interactions between its components.

SimSiam is an adaptation of an existing SSL framework introduced in the study of Chen et al. [157]. This method is widely recognized for employing a simple architecture and avoiding using negative samples, momentum encoders, or specialized memory banks, making it computationally efficient and effective for representation learning. By leveraging the principles of SimSiam, CryoSiam incorporates its proven capabilities into processing subtomograms. This adaptation ensures that the embeddings generated by SimSiam capture

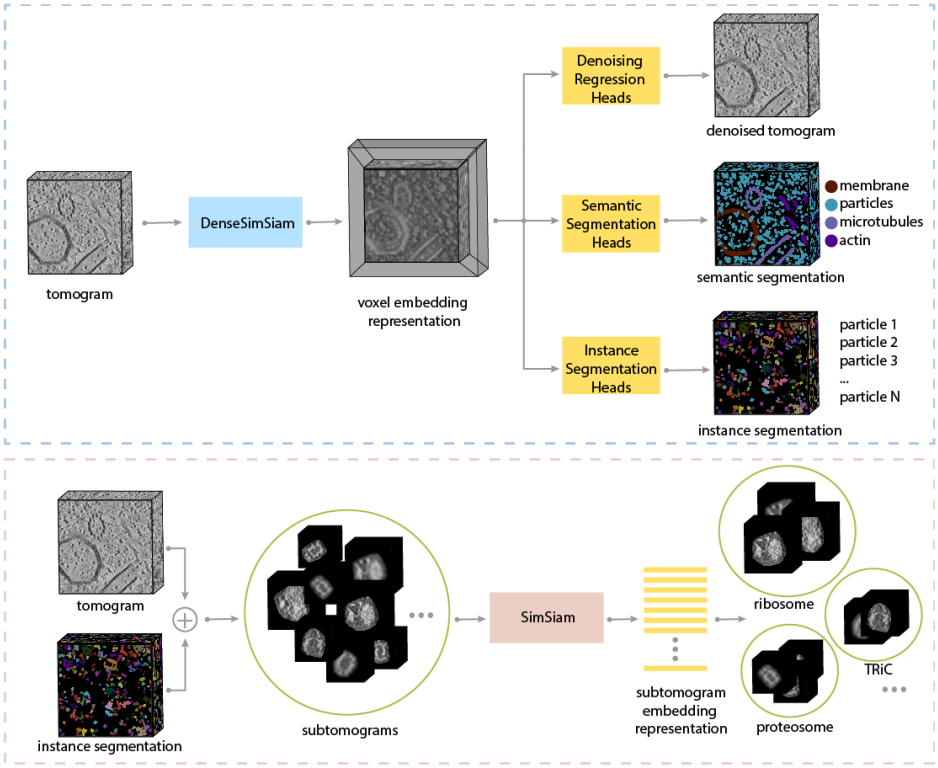

Figure 4: **CryoSiam pipeline for tomogram analysis.** DenseSimSiam processes input tomograms to generate voxel-level embedding representations, which are utilized by task-specific network heads for denoising (producing denoised tomograms), semantic segmentation (segmenting membranes, particles, microtubules, and actin), and instance segmentation (providing individual particle masks). Subtomograms derived and masked from instance segmentation outputs are further processed through the SimSiam network, generating subtomogram embedding representations for downstream tasks such as particle identification.

detailed and meaningful structural features, which are then used to separate particle identities within the embedded space.

While the original SimSiam framework was designed to operate on 2D images, this work modified its architecture and training protocol to support 3D inputs. This extension enables the model to process volumetric data directly, allowing CryoSiam to fully exploit the rich spatial information embedded in tomograms for segmentation and particle identification tasks. Additionally, to adapt the SimSiam component of the CryoSiam framework for cryo-ET data, specific image transformations were introduced to account for the unique characteristics of tomograms.

The transformations applied include the addition of random Gaussian noise, which simulates the inherent noise present in cryo-ET imaging, enabling the model to learn to distinguish meaningful structural information from background noise. Gaussian low-pass filter and Gaussian high-pass filter simulate different imaging conditions that may cause variations in structural information; for example, a particle may appear more blurred under certain conditions (low-pass filter) or reveal more high-resolution details under others

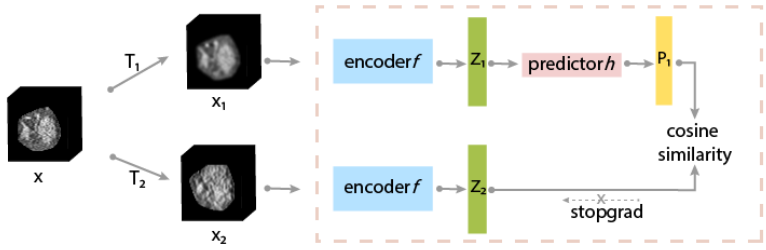

Figure 5: **Detailed architecture of SimSiam.** Subtomograms derived from instance segmentation masks are processed by the SimSiam network. The architecture comprises an encoder (f), a predictor (h), and a stop-gradient operation to ensure stable training. The network is trained in a self-supervised manner by optimizing the cosine similarity between subtomogram embeddings.

(high-pass filter). Affine transformations, including random rotations, cropping, and translations, were applied to introduce variations in orientation and positioning, improving the model's generalization capabilities.

The SimSiam method was initially designed to operate at the image or, in this study, on the subtomogram level, focusing on learning representations for regions rather than individual voxels. To adapt this approach for voxel-level precision, it was necessary to introduce additional components and reimagine the initial design of SimSiam. A conceptually similar method, described in Zhang et al. [161], provided a foundation for this adaptation. This method extends representation learning to the voxel level by inputting two overlapping segments of an image and generating embeddings for every voxel within the input, enabling voxel-level representations.

Building on this idea, the SimSiam framework was extended into DenseSimSiam by introducing an additional controlling step. This step not only allows for the generation of voxel-level embeddings but also ensures consistency and alignment between representations at different levels of granularity. This enhancement enables DenseSimSiam to handle the intricate requirements of cryo-ET data, where precise voxel-level information is critical for segmentation and particle identification tasks.

To further enhance the robustness of voxel-level embeddings, the DenseSimSiam model incorporates a diverse set of transformations during training. These include Gaussian noise, low-pass and high-pass filtering, as previously mentioned. Additionally, a novel masking technique was introduced, wherein specific voxels are masked out in one view while remaining unmasked in the other view. This masking strategy forces the model to rely on the contextual information from neighboring voxels to maintain consistent representations. By learning to predict the missing information, the model becomes more robust to variations and inconsistencies in the data, ensuring that embeddings capture not only the features of individual voxels but also their relationships within the broader neighborhood. This approach significantly improves the model's ability to generalize and adapt to complex cryo-ET datasets.

3.2 THEORETICAL FRAMEWORK

As shown in Figure 5, SimSiam takes two views, x_1 and x_2, of the subtomogram x. Different random augmentations are applied individually as transformations T_1 and T_2 to the respective views. The two augmented views, x_1 and x_2, are fed into an encoder f that utilizes a ResNet architecture [162] to extract important features from each view, followed by a multilayer perceptron (MLP) head to project the feature representations. Thus, the encoder f produces the projected representations as global embeddings $Z_1 \triangleq f(x_1)$ and $Z_2 \triangleq f(x_2)$, with shared weights between the two views.

In addition, a predictor h is used to generate modified representations $P_1 \triangleq h(f(x_1))$ and $P_2 \triangleq h(f(x_2))$. This allows us to minimize the negative cosine similarity between the two representations, defined as:

$$\mathcal{D}(P_1, Z_2) = -\frac{P_1}{\|P_1\|_2} \cdot \frac{Z_2}{\|Z_2\|_2} \tag{1}$$

The predictor h and the stop-gradient operation are used to stop the collapsing problem [157] while minimizing the negative cosine similarity of the image representations:

$$\mathcal{L}_{global} = \frac{1}{2}\mathcal{D}(P_1, stopgrad(Z_2)) + \frac{1}{2}\mathcal{D}(P_2, stopgrad(Z_1)) \tag{2}$$

The first part of the DenseSimSiam architecture is the same as the already described one of SimSiam. In the second part of the model, as shown in Figure 6, the features from the encoder f are fed into a decoder g, which is a basic Feature Pyramid Network (FPN) [163] with a 1x1 3-layer CNN projector in the final layer. This decoder creates dense pixel-level embedding representations $z_1 \triangleq g(f(x_1))$ and $z_2 \triangleq g(f(x_2))$, with shared weights between the two views. A 1x1 2-layer CNN predictor h' generates modified representations $p_1 \triangleq h'(g(f(x_1)))$ and $p_2 \triangleq h'(g(f(x_2)))$. By utilizing the representations from the decoder g and the projected representations from the projector h', the cosine similarity on the dense pixel representations is minimized as follows:

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2} \tag{3}$$

To avoid the collapsing problem while learning the cosine similarity of dense embeddings, the predictor h' and the stop-gradient operation are utilized. The dense loss function is defined as the cosine similarity, which is limited to the pixels located within the overlapping region of x_1 and x_2:

$$\mathcal{L}_{dense} = \frac{1}{2}\mathcal{D}(p_1, stopgrad(z_2)) + \frac{1}{2}\mathcal{D}(p_2, stopgrad(z_1)) \tag{4}$$

A 1x1 3-layer CNN projector is employed in the lower layers of the decoder g resulting in embeddings on different levels, $l_{i_1} \triangleq g(f(x_1))_{\{i\}}$ and $l_{i_2} \triangleq g(f(x_2))_{\{i\}}$, where i denotes the i-th level of the decoder g. These embeddings capture compact information about the image at different resolutions. A 1x1 2-layer CNN predictor h' generates the modified representations

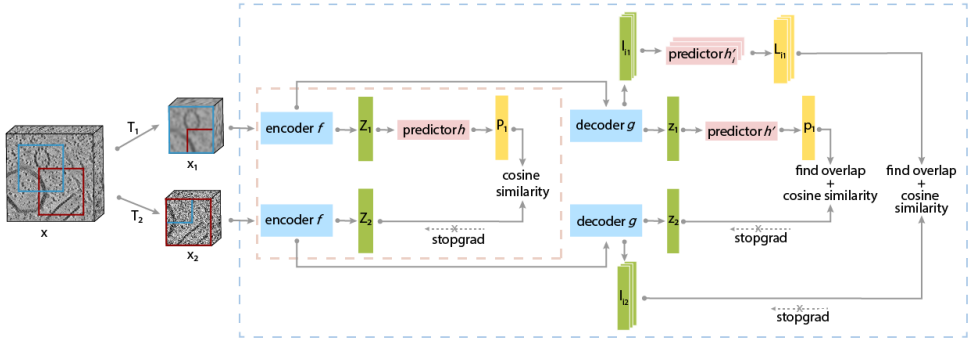

Figure 6: **Detailed architecture of DenseSimSiam.** A cropped input tomogram undergoes transformations to create augmented overlapping views, which are processed by the encoder (f) to generate subtomogram embeddings. The decoder (g) outputs hierarchical embeddings, including level embeddings and voxel embeddings. Predictor networks and stop-gradient operations are employed to prevent training collapse. The model is trained in a self-supervised manner using cosine similarity for subtomogram embeddings, as well as for overlapping voxel and level embeddings.

$L_{i_1} \triangleq h'(g(f(x_1))_{\{i\}})$ and $L_{i_1} \triangleq h'(g(f(x_2))_{\{i\}})$. By utilizing the level embedding representations from the decoder g and the projected representation from the projector h', the cosine similarity on individual levels i is minimized as follows:

$$\mathcal{L}_{level_i} = \frac{1}{2}\mathcal{D}(l_{i_1}, stopgrad(L_{i_2})) + \frac{1}{2}\mathcal{D}(l_{i_2}, stopgrad(L_{i_1})) \tag{5}$$

To obtain the final loss from the levels, the individual losses from the n available levels are summed as follows:

$$\mathcal{L}_{level} = \sum_{i=1}^{n} \mathcal{L}_{level_i} \tag{6}$$

The loss function used for training the global, level and local components simultaneously, while incorporating level information, is defined as follows:

$$\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{dense} + \mathcal{L}_{level} \tag{7}$$

3.2.1 *Image transformations*

Cryo-ET datasets often exhibit significant variations in resolution and quality, even when imaging the same structural components. These differences arise due to variations in sample preparation, imaging conditions, defocus levels, sample thickness, and equipment settings used during data acquisition [20]. Such inconsistencies pose challenges for learning-based methods, as they can introduce biases or reduce the effectiveness of feature extraction. To address these issues, the CryoSiam framework incorporates robust data transformations for the self-supervised training strategies to ensure the learned representations are invariant to these variations while maintaining high accuracy for downstream tasks.

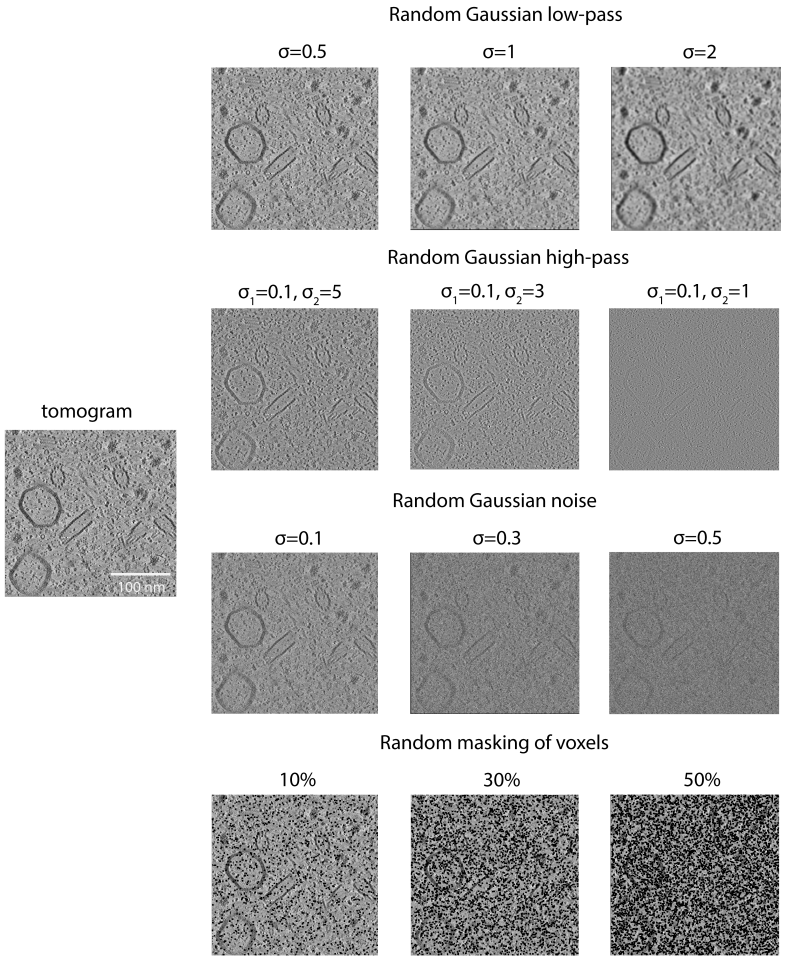

Figure 7: **Illustrations of the image transformations.** The figure demonstrates the transformations applied during self-supervised training to simulate variability in cryo-ET data. Random Gaussian low-pass filtering (σ = 0.5, 1, 2) smooths high-frequency components, while high-pass filtering (σ_1 = 0.1, σ_2 = 5, 3, 1) enhances fine details. Random Gaussian noise (σ = 0.1, 0.3, 0.5) mimics imaging noise, and random voxel masking (10%, 30%, 50%) introduces sparsity, enhancing contextual learning.

The design of transformations used in SSL methods plays a critical role in the learning process, as these transformations shape the model's ability to extract meaningful representations. Transformations must be carefully adapted to the data's specific characteristics to ensure optimal performance. For cryo-ET, the unique properties of tomograms, such as high noise levels and structural heterogeneity, require specialized transformations. Transformations that address these challenges can simulate common variations in cryo-ET data, such as noise, resolution changes, and structural inconsistencies, thereby enhancing the model's generalization capability. Figure 7 illustrates the impact of these transformations, demonstrating how they account for the inherent variability in cryo-ET datasets.

RANDOM GAUSSIAN NOISE Random Gaussian noise is introduced as a transformation to enhance the robustness of voxel embeddings against the

inherent noise present in cryo-ET data. This transformation simulates realistic noise conditions by adding random Gaussian-distributed values, with a sigma value spanning from 0.1 to 0.5, to the voxel intensities. The generated noise is directly added to the tomogram, mimicking the high noise levels characteristic of cryo-ET tomograms. By training the model to process data with varying noise levels, this transformation encourages the learned embeddings to focus on meaningful structural features rather than noise artifacts. This enhancement ensures the embeddings remain invariant to noise variations, improving the model's generalization capabilities across tomograms with differing SNRs.

RANDOM GAUSSIAN LOW-PASS FILTERING The same particle may appear in different tomograms or even within the same tomogram, with varying resolutions due to different imaging conditions. Random Gaussian Low-Pass Filtering addresses this variability by attenuating high-frequency components in the tomograms, effectively smoothing fine details while preserving the broader structural features. This transformation is implemented as a Gaussian filter with sigma values randomly selected from a range of 0.5 to 2, allowing for controlled and varied smoothing of the tomograms. By reducing sensitivity to resolution changes, the transformation ensures the model can concentrate on the core structural features, even when fine details are diminished. By replicating the conditions of lower-resolution imaging, this approach ensures that voxel embeddings maintain robustness and consistency across datasets with different levels of detail.

RANDOM GAUSSIAN HIGH-PASS FILTERING Variations in imaging conditions can result in differences in the visibility of high-frequency details, such as edges and delicate structures, within the same tomogram or across different tomograms. Random Gaussian high-pass filtering enhances these high-frequency components by suppressing low-frequency information, effectively emphasizing fine structural details. This transformation is implemented by subtracting one Gaussian filter, with a sigma value randomly selected from 1 to 1.5, from another Gaussian filter, with a sigma value randomly selected from 0.1 to 0.5. This approach allows for precise enhancement of high-frequency components while introducing controlled variability during training. By emphasizing fine-grained structural features, this transformation directs the model's attention to the detailed characteristics defining macromolecules, improving its ability to differentiate between similar particles or identify subtle structural variations. Incorporating high-pass filtering makes the learned embeddings more attuned to fine-grained details, effectively complementing other transformations and prioritizing broader structural elements.

MASKING-OUT VOXELS TRANSFORMATION A masking-out technique is introduced to enhance the contextual information captured by dense embeddings. This technique serves as an additional transformation applied to the tomogram. In this approach, two transformed views of the tomogram undergo independent voxel masking, with the masking applied to a random percentage of voxels ranging from 10% to 50%. For masked voxels in one view, the dense

embeddings lack individual voxel information, requiring the model to rely on contextual information from the surrounding neighborhood. The loss function ensures that these embeddings remain similar to the corresponding dense embeddings in the second view, which has a different set of masked voxels. This process encourages the dense embeddings to integrate additional neighborhood information, enabling the model to maintain consistency between the two views despite the absence of direct voxel data in some regions.

SUBTOMOGRAM AFFINE AND CROPPING TRANSFORMATIONS In addition to the voxel-level transformations, the SimSiam model incorporates transformations specifically designed for subtomogram embeddings. These transformations, referred to as subtomogram affine and cropping transformations, aim to introduce variability in the spatial and structural context of subtomograms to improve generalization. The affine transformations include random rotations of subtomograms, with angles uniformly selected from 0 to 360 degrees, and translations of up to 10 voxels in any spatial direction. These transformations simulate the variability introduced by different orientations and positions during data acquisition. Random cropping is also applied, where up to 20% of the subtomogram is removed in each instance, with the cropping values selected randomly for every transformation. These augmentations enhance the robustness of the learned embeddings by encouraging the model to focus on core structural features while remaining invariant to positional and orientation-based changes, ultimately improving its performance in downstream particle identification tasks.

3.2.2 *Downstream tasks*

After training DenseSimSiam in a self-supervised manner, the model is fine-tuned to perform three downstream tasks: denoising, semantic segmentation, and instance segmentation. These tasks target specific challenges in cryo-ET data analysis, leveraging the learned voxel embeddings to produce outputs tailored to each task (Figure 4). To achieve this, dedicated CNN heads are integrated into the encoder-decoder architecture. Each task-specific CNN head is designed with a unique output logic and employs a distinct loss function optimized for its objectives, ensuring effective adaptation of the shared features for the desired outcomes.

The denoising task in the CryoSiam framework addresses one of the fundamental challenges in cryo-ET data analysis: the high noise levels inherent in tomograms. This task begins with a noisy tomogram as input and aims to generate a denoised version that retains the original data's structural integrity and fine details. Denoising is a critical preprocessing step, as it enhances the quality of tomograms and facilitates subsequent tasks such as segmentation and particle identification.

To achieve effective denoising, the Mean Squared Error (MSE) loss function, referred to here as $L_{\texttt{denoised}}$, is employed to optimize the model's performance. This loss function quantifies the voxel-wise difference between the predicted denoised tomogram and the corresponding ground truth, which is typically a clean or simulated noise-free tomogram. The loss is defined as:

$$\mathcal{L}_{denoised} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{8}$$

where $\hat{y}_i$ represents the voxel's intensity in the predicted tomogram, y_i represents the intensity in the ground truth, and n is the total number of voxels. By minimizing $\mathcal{L}_{denoised}$, the model is trained to effectively reduce noise while preserving critical structural details within the tomograms.

This approach allows the CryoSiam framework to produce high-quality denoised tomograms that closely resemble the ground truth, ensuring a balance between noise reduction and the preservation of structural features. A well-trained denoising model significantly improves the overall quality of cryo-ET datasets, enabling more accurate downstream analysis and segmentation outcomes.

The semantic segmentation task in the CryoSiam framework is designed to classify each voxel in the tomogram into predefined structural classes while also predicting the distances of voxels from boundaries between these classes. This task provides a detailed voxel-level understanding of the tomogram, segmenting key structural components such as membranes, microtubules, actin filaments, and particles. Semantic segmentation is essential for analyzing the composition and organization of tomographic data, enabling downstream tasks that rely on precise structural delineation.

Three loss functions are employed to train the model for this task, each contributing uniquely to the optimization framework. The first loss function, referred to as $\mathcal{L}_{sem1}$, is the Cross-Entropy (CE) loss function which computes the probability-based classification error for each voxel. It encourages the model to accurately assign semantic class labels to all voxels in the tomogram. This loss is defined as:

$$\mathcal{L}_{sem1} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{9}$$

where N is the number of voxels, C is the number of classes, $y_{i,c}$ is the ground truth probability for class c at voxel i, and $\hat{y}_{i,c}$ is the predicted probability.

The second loss function, $\mathcal{L}_{sem2}$, is the Generalized Dice loss function, which addresses class imbalances commonly observed in cryo-ET data. This loss ensures that smaller or less-represented classes are appropriately weighted during training. It is computed as:

$$\mathcal{L}_{sem2} = 1 - \frac{2\sum_{c=1}^{C} w_c \sum_{i=1}^{N} y_{i,c}\hat{y}_{i,c}}{\sum_{c=1}^{C} w_c \sum_{i=1}^{N}(y_{i,c} + \hat{y}_{i,c})} \tag{10}$$

where $w_c = \frac{1}{\left(\sum_{i=1}^{N} y_{i,c}\right)^2}$ is the class weighting factor to balance the contribution of each class.

The third loss function, $\mathcal{L}_{distance}$, is applied to refine the predicted distances from the background, improving the accuracy of boundary delineation. This loss is defined as:

$$\mathcal{L}_{distance} = \frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2 \tag{11}$$

where $\hat{d}_i$ represents the predicted distance of voxel i from the background, and d_i is the ground truth distance.

The final loss used to train the semantic segmentation task is a combination of these three losses, expressed as:

$$\mathcal{L}_{sem\_total} = \mathcal{L}_{sem1} + \mathcal{L}_{sem2} + \mathcal{L}_{distance} \tag{12}$$

Each of these loss functions contributes uniquely to the segmentation process. The CE loss focuses on voxel-level classification accuracy, ensuring the semantic classes are correctly predicted. Generalized dice loss complements this by addressing class imbalance and ensuring that smaller or less-represented classes are appropriately weighted during training. Finally, MSE loss refines the model's understanding of the spatial context and boundaries within the tomograms, improving the delineation of semantic classes.

Together, these three loss functions create a robust and comprehensive optimization framework for semantic segmentation. This combined approach allows the CryoSiam framework to produce accurate and reliable segmentation results, effectively classifying voxels and creating structural predictions across diverse cryo-ET datasets.

The instance segmentation task in the CryoSiam framework focuses on identifying individual particle instances by predicting their boundaries, foreground regions, and distances from the background. This task aims to generate detailed masks for each particle, enabling precise delineation and localization of structural components. To achieve instance segmentation, the predicted distances and boundaries were used to guide a watershed segmentation algorithm [164]. To further refine the segmentation results, a multi-cut algorithm [165] was applied, ensuring accurate delineation of individual particle instances while minimizing over-segmentation or boundary errors. Instance segmentation is crucial for separating nearby particles and analyzing their spatial relationships within tomograms.

Three distinct loss functions are employed, each addressing a specific aspect of the task. The first loss function, referred to as $\mathcal{L}_{distance}$, is the MSE loss function, which optimizes the predicted distances of voxels from the background. The model refines its understanding of spatial relationships by minimizing the squared differences between the predicted and ground truth distances and effectively delineates particle instances. This loss is defined as:

$$\mathcal{L}_{distance} = \frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2 \tag{13}$$

where $\hat{d}_i$ represents the predicted distance of voxel i from the background, and d_i is the ground truth distance.

The second loss function, $\mathcal{L}_{boundary}$, is a Binary Cross-Entropy (BCE) loss function used to predict the boundaries of individual particles. This loss encourages the model to identify edges, accurately separating adjacent instances. It is computed as:

$$\mathcal{L}_{boundary} = -\frac{1}{N}\sum_{i=1}^{N}\left[b_i \log(\hat{b}_i) + (1 - b_i)\log(1 - \hat{b}_i)\right] \tag{14}$$

where b_i is the ground truth label for whether voxel i lies on a boundary (one for boundary, zero otherwise) and $\hat{b}_i$ is the predicted boundary probability.

The third loss function, $\mathcal{L}_{foreground}$, is another BCE loss function applied to predict the foreground class for each voxel. This loss ensures that the model accurately classifies voxels as belonging to particle instances or the background. It is similarly defined as:

$$\mathcal{L}_{foreground} = -\frac{1}{N}\sum_{i=1}^{N}\left[f_i \log(\hat{f}_i) + (1 - f_i)\log(1 - \hat{f}_i)\right] \tag{15}$$

where f_i is the ground truth label for whether voxel i belongs to the foreground (1 for foreground, 0 for background), and $\hat{f}_i$ is the predicted probability.

The final loss function for instance segmentation, combines these three losses into a unified optimization framework:

$$\mathcal{L}_{instance\_total} = \mathcal{L}_{distance} + \mathcal{L}_{boundary} + \mathcal{L}_{foreground} \tag{16}$$

This comprehensive approach effectively addresses each aspect of the segmentation task. The MSE Loss refines the spatial relationships and distances, enabling precise delineation of particle instances. The BCE Loss for boundaries guides the model to detect particle edges accurately, capturing fine-grained details and ensuring separation between instances. Meanwhile, the BCE Loss for the foreground ensures robust classification of voxels as particles or background.

Together, these loss functions create a robust optimization framework for instance segmentation, enabling the CryoSiam framework to handle the complexities of this task. The model delivers reliable instance segmentation performance across diverse cryo-ET datasets by accurately predicting particle boundaries, foreground regions, and spatial relationships.

3.2.3 *Contrastive learning*

After pretraining a backbone model using SimSiam, we obtain semantically rich embeddings without requiring extensive manual annotations. SimSiam's self-supervised training ensures that embeddings from different augmented views of the same subtomogram remain similar, resulting in robust, invariant feature representations. However, while these representations capture semantic information, they do not explicitly enforce separation between distinct

classes or concepts. We introduce a discriminative signal that refines the embedding space by integrating a Siamese network with a contrastive loss after SimSiam pretraining. This contrastive objective encourages similar examples to form tighter clusters and compels dissimilar examples to lie beyond a specified margin, creating a more structured and task-relevant metric space. In other words, contrastive learning leverages the strong, general-purpose features learned by SimSiam. Further, it improves them by enforcing class-aware separability, ultimately yielding semantically meaningful and discriminative embeddings.

In this approach, pairs of images - positively matched (e.g., belonging to the same class) or negatively matched (belonging to different classes) - are passed through the shared, SimSiam-pretrained backbone. The SimSiam method produces normalized embeddings e_1 and e_2, and binary label y is defined, where $y = 1$ for positive pairs and $y = 0$ for negative pairs. The contrastive loss with a chosen margin $margin > 0$ is given by:

$$\mathcal{L}_{contrastive} = y \cdot d(\mathbf{e}_1, \mathbf{e}_2)^2 \; + \; (1-y) \cdot \max(0, \text{margin} - d(\mathbf{e}_1, \mathbf{e}_2))^2 \tag{17}$$

with

$$d(\mathbf{e}_1, \mathbf{e}_2) = \|\mathbf{e}_1 - \mathbf{e}_2\|_2 \tag{18}$$

This loss encourages embeddings of positive pairs to be pulled closer together, minimizing $d(\mathbf{e}_1, \mathbf{e}_2)$, and enforces a minimal separation (defined by $margin$) for negative pairs. By coupling SimSiam's self-supervised pretraining with the subsequent Siamese contrastive refinement, the resulting embedding space is semantically meaningful and tuned to discriminate between relevant classes or semantic categories with minimal labeled data.

3.3 SIMULATED DATA

In cryo-ET, the lack of annotated datasets presents significant challenges for developing and evaluating computational methods. Manual annotation is a resource-intensive process that demands specialized expertise, making it unsuitable for large-scale data. Simulated data has emerged as a practical solution to overcome these limitations, offering several key advantages for segmentation tasks [87, 88]. Among its benefits, simulated datasets provide precise ground-truth annotations, essential for calculating reliable evaluation metrics for a given task. These metrics enable a consistent and objective assessment of performance. For example, Martinez et al. [90] emphasize the value of realistic synthetic datasets with known ground truth for effectively training deep learning models with the simulated data, enhancing their generalization to experimental data.

The SHREC challenge highlights the critical role of simulated datasets in cryo-ET research [87]. It introduced a publicly accessible dataset comprising ten tomograms of simulated cell-like volumes, each containing twelve distinct types of macromolecular complexes. These simulations incorporated realistic

noise and imaging artifacts, providing a robust benchmark for evaluating localization and classification methods. Ground-truth annotations allowed for objectively comparing various approaches, demonstrating that modern learning-based methods consistently outperformed traditional template matching techniques [88]. This underscores the potential of simulated data to drive advancements in computational method development for cryo-ET.

Using simulated data offers a controlled environment for experimentation, enabling systematic variation of parameters such as noise levels, resolution, and structural complexity. This allows for rigorous evaluation of methods across diverse scenarios, providing insights into their robustness and limitations. Controlled, parameterized simulations have demonstrated their utility in assessing the performance of particle localization and classification approaches, highlighting their value in understanding and optimizing computational techniques [166].

Building on the controlled experimentation capabilities, simulated data also serves a critical function in enabling self-supervised learning. This approach leverages pretext tasks to learn meaningful representations without relying on manual annotations. Simulated datasets provide the flexibility to design and evaluate these tasks effectively, allowing models to extract robust features. For instance, training on single noisy volumes without ground truth has demonstrated how simulated data can support the development of self-supervised methods tailored to cryo-ET in the study of Yang et al. [167].

The dataset generated for this study is named CryoETSim (CRYO-Electron Tomography SIMulated dataset), incorporating key attributes to ensure biological relevance and structural diversity. The simulation process utilized the cryoTomoSim (CTS) software [89]. This approach was chosen to address limitations in currently available datasets, such as SHREC, which do not adequately replicate in-situ tomograms. Specifically, existing datasets lack diversity in contrast, caused by variations in defocus and sample thickness, making them less representative of real-world scenarios.

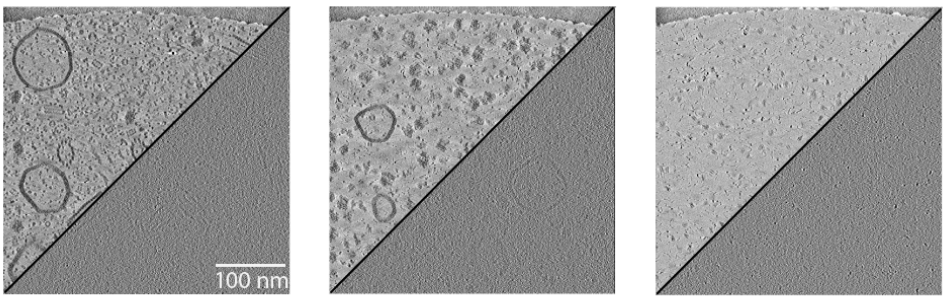

Figure 8: **Representative simulated tomograms from the CryoETSim dataset.** The tomograms correspond to three sample types: general, transcription-translation, and chromatin. The general sample includes diverse structural components, the transcription-translation sample features key molecular machinery involved in transcription and translation, and the chromatin sample contains DNA filaments and nucleosomes. The scale bar represents 100 nm.

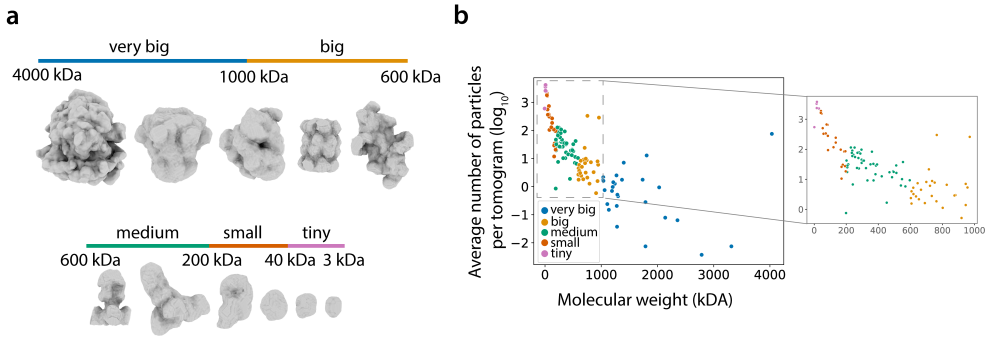

Figure 9: **Particle classification and distribution by molecular weight in CryoETSim general samples.**
(a) Particle classes categorized by molecular weight, ranging from very big (>1000 kDa), big (600–1000 kDa), medium (200–600 kDa), small (40–200 kDa), to tiny (<40 kDa). Representative particles from each category are visualized. (b) Scatter plot showing the average number of particles per tomogram as a function of molecular weight. The inset highlights the distribution for smaller molecular weights, revealing the higher frequency of smaller particles in the dataset.

The CryoETSim dataset comprises 400 tomograms, each representing one of three distinct sample types. The first, known as the general sample, includes approximately 130 PDB-structured entries and structural elements such as membranes, actin filaments, and microtubules, capturing a broad biological context with significant complexity. The second sample, the transcription-translation sample, incorporates key molecular machinery involved in transcription and translation processes, including ribosomes (divided into small and large subunits), RNA polymerase, and their interacting complexes. This sample also includes simulated DNA filaments to enhance biological relevance. The third sample, the chromatin sample, exclusively focuses on chromatin structures, featuring simulated DNA filaments and nucleosomes. These diverse sample types collectively provide a robust dataset for evaluating segmentation techniques across a broad spectrum of biological scenarios, as illustrated in Figure 8.

Each panel in Figure 8 presents a clean tomogram in the top-left corner alongside a noisy tomogram with CTF modulation in the bottom-right corner, illustrating the impact of noise and modulation on contrast and particle visibility. These samples collectively form a comprehensive dataset, providing high-quality data to train SSL methods and evaluate segmentation techniques across diverse biological scenarios.

The general sample in the CryoETSim dataset encompasses PDB structures spanning a wide size range, with molecular weights from 10 kDa to more than 4000 kDa. These particles were categorized into five distinct groups based on molecular weight: very big (>1000 kDa), big (600–1000 kDa), medium (200–600 kDa), small (40–200 kDa), and tiny (<40 kDa) (Figure 9a). To replicate crowded cellular environments with densities reaching up to 80%, particles were added sequentially, starting from the largest and progressing to the smallest. This approach ensured an increasing particle count per tomogram as particle size decreased, closely mimicking natural crowding conditions observed in biolog-

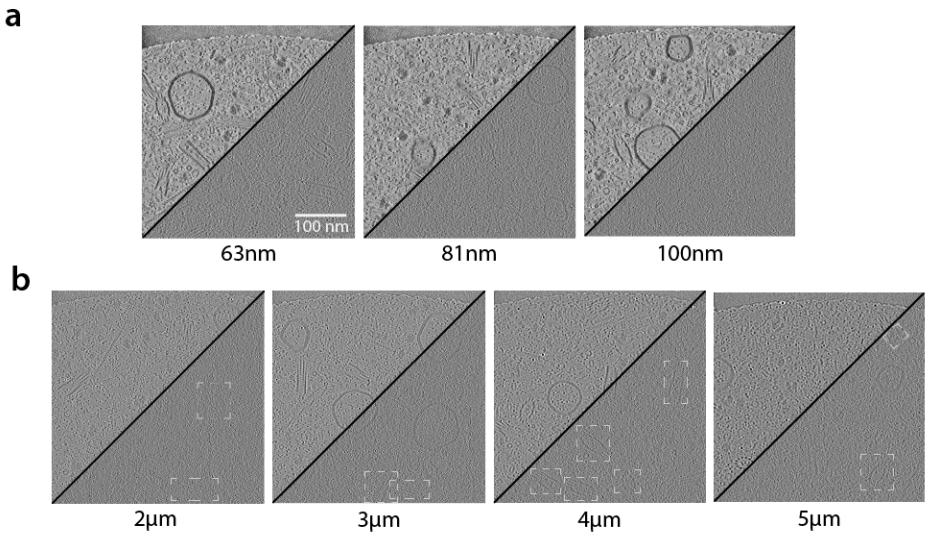

Figure 10: **Impact of thickness and defocus on tomogram visibility and contrast.** (a) Simulated tomograms with increasing thickness (63 nm, 81 nm, 100 nm) demonstrate the reduction in particle visibility and contrast as sample thickness increases. (b) Simulated tomograms with varying defocus levels (2 µm, 3 µm, 4 µm, 5 µm) illustrate the effect of defocus on contrast and structural detail, where higher defocus emphasizes low-frequency information and reduces high-frequency details. The dashed rectangles highlight microtubules, providing a reference for comparing visibility across different defocus levels.

ical samples (Figure 9b). These details highlight the structural diversity in the general sample, making it a valuable resource for testing segmentation methods.

Extending the diversity in the general sample, the DNA filaments in the chromatin and transcription-translation samples were simulated with random lengths to enhance variability and realism. The coordinates of the base pairs were generated using the wormlike chain (WLC) folding model, a computational approach that accurately represents the flexible polymer-like properties of DNA [168]. This method ensured the creation of realistic folding patterns and structural configurations. The resulting filament structures were then converted into PDB files, which served as input for the CTS simulation algorithm. This pipeline enabled the integration of DNA filaments into the tomographic simulations, further enriching the dataset's complexity.

To further enhance the realism of the simulated datasets, variations in sample thickness and defocus levels were incorporated, as illustrated in Figure 10a and Figure 10b, respectively. Figure 10a showcases the influence of sample thickness, varied between 60 nm and 100 nm. For each thickness value, the clean tomogram without CTF modulation is displayed in the top-left corner, while the noisy, CTF-modulated tomogram appears in the bottom-right corner. Higher thickness values reduce contrast, affecting the visibility of smaller and less distinct particles.

Figure 10b highlights the effects of varying defocus values, ranging from 2 µm, which accentuates high-frequency details and produces sharper images, to 5 µm, where low-frequency information prevails, resulting in stronger con-

trast. Each panel in this figure presents the CTF-modulated tomogram in the top-left corner and the noisy tomogram in the bottom-right corner. White dotted rectangles highlight microtubules, whose visibility diminishes at lower defocus due to increased noise and reduced contrast. These controlled variations in thickness and defocus effectively replicate the range of visual modulation and contrast observed in experimental tomograms, adding a layer of authenticity to the simulated dataset.

The CryoETSim dataset was carefully constructed to incorporate key biological and imaging characteristics, ensuring its relevance and utility for developing the segmentation method. This dataset closely replicates the visual and structural diversity observed in experimental cryo-ET data by simulating diverse samples with structural complexity, incorporating realistic noise, CTF modulation, and varying parameters such as defocus and thickness. These features make CryoETSim a robust resource for training and evaluating computational methods, providing a strong foundation for addressing the challenges associated with real-world cryo-ET analysis.

3.4 TRAINING STRATEGY

The CryoSiam training protocol was structured in multiple stages. The self-supervised phase pre-trained the DenseSimSiam and SimSiam models on simulated data to learn generalizable features. This was followed by task-specific training for semantic segmentation, instance segmentation, and particle identification.

The training protocol for DenseSimSiam self-supervised training employed cosine similarity as the loss function, with the model trained for 600 epochs using a cosine annealing scheduler [169]. The scheduler included an initial warmup phase of 5 epochs to gradually increase the learning rate, which was set to a maximum of 0.5. Optimization was performed using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 0.00001 [170]. A batch size of 10 was maintained throughout the training process. This carefully designed training setup ensured stable and efficient optimization, enabling DenseSimSiam to capture meaningful structural features and lay the groundwork for subsequent downstream tasks.

The training process for SimSiam closely followed that of DenseSimSiam, utilizing cosine similarity as the loss function to align feature representations. A cosine annealing scheduler was employed to adjust the learning rate to a maximum of 0.05 dynamically. The scheduler included an initial warmup phase of 10 epochs to gradually increase the learning rate. The training spanned 200 epochs, with a batch size of 10 used throughout the process. Optimization was carried out using SGD with a momentum of 0.9 and a weight decay of 0.00001.

For all downstream tasks within the CryoSiam framework, a one-cycle learning rate scheduler [171] was employed to adjust the learning rate throughout the training process dynamically. This approach allowed the learning rate to increase gradually during the initial training phase, peak at an optimal point, and decrease, enabling faster convergence and improved generalization. A fixed learning rate of 0.001 was used for all tasks, and the models were trained for 200 epochs. The batch sizes varied depending on the specific requirements

of each task, ensuring optimal resource utilization and training efficiency. Additionally, the AdamW optimizer [172] was utilized, combining the benefits of adaptive learning rates with weight decay regularization to prevent overfitting. This combination of the one-cycle learning rate scheduler, AdamW, and task-specific configurations ensured efficient and stable optimization across all downstream tasks.

3.5 EVALUATION METRICS

To comprehensively evaluate the performance of the CryoSiam framework across its downstream tasks, specialized evaluation metrics were employed for semantic segmentation, instance segmentation, and particle identification. Each metric was carefully chosen to align with the objectives and outputs of the respective tasks, ensuring robust and meaningful assessment.

The DICE score, or the Sørensen-Dice coefficient, was used to measure the overlap between predicted segmentation masks and ground truth for semantic segmentation. This metric is particularly well-suited for segmentation tasks with class imbalances, as it directly evaluates the voxel-wise agreement between the two sets. The DICE score is calculated as:

$$\mathrm{DICE} = \frac{2 \times |P \cap G|}{|P| + |G|} \tag{19}$$

where P represents the predicted set of voxels, and G represents the ground truth set. The numerator, $2 \times |P \cap G|$, measures the intersection between the prediction and ground truth. In contrast, the denominator $|P| + |G|$ normalizes the value by the total size of the predicted and ground truth sets. A DICE score of 1 indicates perfect overlap, while a score of 0 signifies no overlap. This metric assessed the framework's accuracy in segmenting the regions corresponding to each semantic class.

For the instance segmentation task of the CryoSiam framework, the mask Average Precision (AP) metric at different Intersection over Union (IoU) thresholds was employed. This metric evaluates the model's ability to accurately localize and separate particle instances by comparing the predicted and ground truth masks.

The calculation of AP involves the following steps. First, the predicted instances are ranked by confidence scores, from highest to lowest. The IoU between the predicted and ground truth masks is computed for each prediction. IoU is defined as:

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{20}$$

where P is the set of voxels in the predicted mask, and G is the set of voxels in the ground truth mask. A match is considered valid if the IoU exceeds a predefined threshold, such as 0.5 (commonly referred to as AP-50).

The precision value is based on the number of true positives (TP), false negatives (FN), and false positives (FP), calculated for given threshold t as:

$$\text{precision}(t) = \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{21}$$

where $TP(t)$ refers to correctly predicted instances, $FP(t)$ refers to incorrectly predicted instances, and $FN(t)$ refers to ground truth instances missed by the model, when the threshold for IoU is t.

The average precision of a single image is then calculated as the mean of the above precision values at each IoU threshold:

$$AP = \frac{1}{n} \sum_{t=1}^{n} \text{precision}(t) \tag{22}$$

where n is the total number of thresholds considered. This formulation effectively summarizes the model's performance across different levels of localization accuracy.

In addition to evaluating AP across multiple thresholds, a specific variant known as AP-50 (AP@0.5) was used to measure performance at a fixed IoU threshold of 0.5. AP-50 evaluates the precision of predictions where the overlap between predicted and ground truth masks is at least 50%. This metric focuses on the model's ability to make confident and accurate predictions that meet one threshold, providing a direct measure of the model's reliability at one fixed IoU threshold of 0.5. Including AP-50 in the evaluation framework ensures a targeted assessment of the CryoSiam framework's ability to achieve a specific overlap threshold, complementing the more comprehensive evaluation provided by the overall AP metric.

For particle identification, the F1 score was employed to evaluate the accuracy of particle localization and classification. The F1 Score is the harmonic mean of precision and recall, balancing these two metrics' trade-offs. It is calculated as:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{23}$$

Here, precision and recall are computed as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{25}$$

Precision reflects the fraction of predicted particles that are correctly identified, providing a measure of prediction reliability by penalizing false positives. Recall, in contrast, measures the fraction of actual particles successfully identified by the model, focusing on prediction completeness and penalizing false negatives. The F1 score combines these two aspects into a single metric, balancing capturing all relevant particles (high recall) and avoiding incorrect predictions (high precision). A high F1 score indicates a substantial trade-off between precision and recall, demonstrating the model's ability to localize and classify

particles accurately while maintaining comprehensive coverage of the ground truth particles.

Together, these metrics provide a comprehensive framework for assessing the CryoSiam framework across its diverse tasks. By focusing on segmentation accuracy, instance detection, and particle-level precision, the evaluation process ensures that the framework meets the demands of cryo-ET data analysis with high reliability and robustness.

4

EXPERIMENTS WITH SIMULATED DATA

The foundation of this research lies in systematically evaluating the proposed CryoSiam framework using simulated cryo-ET data. Simulated datasets provide a controlled and well-annotated environment, enabling detailed assessments of model performance across various tasks such as semantic segmentation, instance segmentation, and particle identification. These experiments serve as a critical step in understanding the strengths and limitations of CryoSiam before transitioning to the complexities of real cryo-ET datasets.

TThis chapter focuses on applying CryoSiam to simulated tomograms, leveraging their clean and noise-free annotations to benchmark key components of the framework. The chapter begins by outlining the experimental setup and the data preparation process, including subtomogram extraction, augmentation strategies, and self-supervised training protocols. Subsequently, ablation studies are presented to dissect the contributions of individual model components and transformations, providing insights into their role in enhancing performance. Finally, comprehensive results for segmentation and particle identification tasks are analyzed, highlighting the model's ability to learn meaningful structural features.

By the end of this chapter, a clear understanding of CryoSiam's performance on simulated data has been established, laying the groundwork for its application to real cryo-ET datasets in subsequent chapters.

4.1 EXPERIMENTAL SETUP

The experiments in this chapter were conducted using the CryoETSim dataset, a comprehensive collection of 400 tomograms representing diverse biological scenarios, including general, transcription-translation, and chromatin-like samples. For training the self-supervised pretext tasks of DenseSimSiam and SimSiam, clean tomograms, free of noise and CTF modulation, were utilized to ensure that the models could learn from high-quality structural data and create robust voxel and subtomogram embeddings. To reflect the diversity and artifacts that typically appear in real cryo-ET data, the transformations applied during training were specifically designed to simulate the data variability. In addition to its role in training, the CryoETSim dataset also provides a robust resource for evaluating segmentation tasks, including semantic and instance segmentation and particle identification, which is central to this study.

The experimental workflow utilized both components of the CryoSiam framework. The first component, DenseSimSiam, was trained on subtomograms extracted from the full tomograms using a sliding window approach. Each subtomogram measured 128 voxels in all three spatial dimensions, with a 50% overlap applied between adjacent subtomograms. This overlapping strategy increased the number of samples available for training and evaluation, preserved spatial continuity, and captured shared structural features across

subtomograms. As a result, the dataset was effectively augmented, producing approximately 25,600 subtomograms for use in the experiments.

The subtomograms were divided into training and validation datasets, with 10% of the subtomograms allocated to the validation set. For the backbone architecture, DenseSimSiam employed a 3D ResNet-10 [162], chosen for its ability to extract hierarchical features from volumetric data effectively. The decoder was implemented as an FPN [163], with the size of the channels set to 128, enabling the model to aggregate multi-scale features and improve the quality of voxel-level embeddings. For the predictors in DenseSimSiam, the design followed the logic from the original SimSiam paper, where a two-layer convolutional neural network was used. The predictor was structured with an encoder-decoder logic: the first layer had the same size as the embedding dimensions, the second layer reduced the size to one-fourth of the embeddings, and the third layer restored it to the original embedding size. The model's global embedding size was configured to 256, while the voxel embedding size was set to 64, ensuring a balance between compact representations and detailed feature learning. Only the first-level embeddings were used during training to compute the loss function, simplifying the optimization process while maintaining effective representation learning. The choice of specific parameters, including the embedding sizes, the architecture, the decoder configuration, and the predictor implementation, was guided by the results of ablation studies, which are described in detail in Section 4.2 of this chapter.

The downstream tasks that follow DenseSimSiam were trained using specialized CNN heads, each designed as a three-layer CNN. These CNN heads were tailored to the specific predictions required by each task, such as denoising, semantic segmentation, and instance segmentation. For the denoising and semantic segmentation tasks, the training and validation datasets were derived from the general sample, with 200 out of 300 tomograms allocated for training and validation, where 10% of the tomograms were used for validation. The remaining 100 tomograms were reserved for testing.

In the DNA filament and chromatin samples, the first 50 tomograms were used for the training and validation datasets, following the same 10% split for validation. In comparison, the remaining 50 tomograms were assigned to the test set. For the instance segmentation task, the same dataset logic was followed; however, only the tomograms from the general sample were used. To create the training datasets for all tasks, subtomograms were extracted with dimensions of 128 voxels in all three spatial dimensions, using a sliding window approach with a 50% overlap. This extraction method ensured comprehensive coverage of the tomographic volume and provided the necessary data for robust model training and evaluation.

During the prediction phase, the model processed the tomograms using a sliding window approach with the same patch size of 128 voxels, consistent with the size used during training. The tomograms were divided into overlapping patches with a 50% overlap, ensuring that all regions of the tomogram were covered. The predictions for each patch were then combined to form the final prediction for the entire tomogram. This approach preserved spatial continuity and allowed the model to produce accurate voxel-level predictions for the full tomogram.

The second part of the CryoSiam framework, known as SimSiam, was explicitly designed to handle subtomograms of individual particles across different classes. This approach addressed the inherent imbalance in the dataset, where smaller particles appeared more frequently in the tomograms than medium-sized or large particles. By extracting up to 10 subtomograms per class from each tomogram, the dataset achieved a more balanced representation of particles with varying sizes, ensuring that the model could learn robust features across different structural categories.

The separation of training and testing datasets followed the same logic as previous tasks, using only the general sample tomograms. The first 200 tomograms were allocated to the training and validation dataset, with 10% designated for validation, while the remaining 100 tomograms comprised the test dataset. After processing, the training and validation dataset consisted of 175,371 subtomograms, and the test dataset included 88,585 subtomograms. This carefully structured dataset facilitated practical training and evaluation for particle identification tasks, enabling the model to generalize well across diverse particle types.

The same dataset separation was applied for the contrastive learning phase of the SimSiam, ensuring consistency in data utilization. A carefully designed strategy was employed to create positive and negative pairs for training. Positive pairs were formed using subtomograms belonging to the same particle class, maintaining structural consistency within the pairs. For negative pairs, a size-based separation of simulated particle classes was utilized. These particle classes, categorized by size, were already provided in the CryoETSim dataset, simplifying the pairing process.

For instance, if the anchor subtomogram represented a big particle, the negative sample was selected from a different big particle class. This approach avoided using particles from different classes as negative samples, as such distinctions present a relatively simple problem for the model, potentially limiting the quality of the learned embeddings. By focusing on challenging size-based negative pairs within the same category, the training process encouraged the model to learn more meaningful and discriminative structural representations, ultimately improving its ability to generalize across particle types.

Two different classifiers were utilized to perform classification after generating embeddings with the SimSiam framework: a k-nearest neighbors (k-NN) classifier and a multi-layer perceptron (MLP) classifier. The k-NN classifier was configured with the number of nearest neighbors set to 1, providing a simple and practical approach to classification based on the similarity of embeddings. Alternatively, the MLP classifier offered a more complex and learnable classification mechanism. The MLP was implemented as a three-layer network, where the first layer had a dimension of 256, the second layer reduced the dimension to 128, and the third layer output predictions corresponding to the number of particle classes 134. This flexible setup allowed for a comparative evaluation of simple versus trainable classifiers in the context of particle identification tasks.

This experimental setup provides a well-structured foundation for evaluating the CryoSiam framework across various tasks, including segmentation and particle identification. By leveraging the carefully designed CryoETSim

dataset, integrating robust training strategies, and innovative contrastive learning approaches, the setup ensures that the framework is rigorously tested under realistic conditions. The subsequent sections present the results and insights derived from these experiments, offering a detailed analysis of the framework's performance across multiple downstream tasks.

4.2 ABLATION STUDIES

Ablation studies were conducted to systematically evaluate the impact of individual components and design choices within the CryoSiam framework on its performance with simulated tomograms. These experiments aimed to identify critical factors that influence the framework's effectiveness when applied to the clean, well-annotated CryoETSim dataset. By isolating specific elements, such as transformations, architectural components, and embedding configurations, the studies provide valuable insights into how each contributes to the framework's performance across tasks like denoising, semantic segmentation, and instance segmentation.

The experiments first evaluated the proposed transformations, including random Gaussian noise, low-pass filtering, high-pass filtering, and voxel masking-out transformations. These transformations were tested individually to determine their contribution to the robustness and generalization of the dense embeddings generated by DenseSimSiam. Next, the impact of the dense embedding size on voxel-level representations was explored, with experiments assessing how different sizes influenced the quality of learned features. Similarly, the global embedding size was investigated to understand how the compactness or expansion of these embeddings affected downstream tasks.

Finally, studies examined how the number of layer embeddings included in the final loss computation influenced model performance. By exploring whether including embeddings beyond the first layer provided benefits or introduced unnecessary complexity, these experiments sought to identify the optimal balance for practical training.

To assess each component's impact on the dense embeddings' quality, an experiment was conducted where a simple three-layer CNN segmentation head was added to the framework. This task aimed to classify each voxel into one of the predefined semantic classes: background, membrane, particle, microtubules, actin, or DNA filaments. The convolutional layers in the segmentation head were configured with a kernel size of 1×1, allowing the network to exclusively utilize information from the individual voxel embeddings without incorporating spatial context.

A reduced dataset was used for this experiment, comprising 40 tomograms from the general sample, 10 from the DNA filament sample, and 10 from the chromatin sample. This dataset was split into training and validation sets, with 10% of the data allocated for validation. This data was used to train DenseSimSiam, after which the whole model was frozen, and only the simple CNN head was trained. For this training phase, the dataset was separated into 48 tomograms for training and 12 tomograms for testing. The DICE score was used as the evaluation metric to measure the segmentation performance

Classes	None	Low-pass	Low-pass, High-pass	Low-pass, High-pass, Gaussian Noise
background	0.835	0.614	**0.900**	0.873
membrane	0.000	0.000	0.000	**0.255**
microtubules	0.000	0.000	0.000	0.000
particles	0.525	0.562	0.599	**0.604**
actin	0.000	0.016	0.000	**0.027**
dna filaments	0.003	0.299	**0.374**	0.328
all	0.302	0.286	0.378	**0.416**

Table 1: **Ablation study results for the proposed transformations.** The table compares the segmentation performance across different structural classes (background, membrane, microtubules, particles, actin, and DNA filaments) under various transformation settings: no transformations (None), low-pass filtering, low-pass combined with high-pass filtering, and the addition of Gaussian noise. The results highlight the importance of combining multiple transformations (low-pass, high-pass filtering, and Gaussian noise) to achieve robust and generalizable segmentation performance.

on the test data, providing a robust assessment of how well the learned dense embeddings supported voxel-level classification.

These ablation studies collectively provide a detailed analysis of the CryoSiam framework, quantifying the effects of each design choice and offering practical insights into optimizing its configuration. The following subsections detail each experiment's specific design, results, and interpretations.

4.2.1 *Image transformations*

The ablation study on image transformations aimed to evaluate the impact of the different transformation combinations on the performance of the CryoSiam framework. Table 1 reports the DICE scores for semantic segmentation across various classes under four settings: no transformations (None), Gaussian low-pass filtering (Low-pass), a combination of low-pass and high-pass filtering (Low-pass, High-pass), and a final combination that additionally included random Gaussian noise (Low-pass, High-pass, Gaussian Noise).

When no transformations were applied (column "None"), the segmentation performance showed limitations, with a relatively high DICE score (0.835) for the background class but close to zero for other structural classes such as membranes, actin, and DNA filaments. This result highlights the difficulty of learning meaningful embeddings without augmentation strategies, especially for more complex structural components.

The inclusion of low-pass filtering alone (column "Low-pass") resulted in a slight drop in overall performance (the DICE score for "all" dropped from 0.302 to 0.286). While the low-pass filter marginally improved the segmentation of actin (0.016) and DNA filaments (0.299), the performance for the background

class degraded significantly. These results suggest that while low-pass filtering can simulate lower-resolution imaging conditions, it may not sufficiently enhance the model's generalization ability across all classes.

Combining low-pass and high-pass filtering (column "Low-pass, High-pass") improved the performance across multiple classes. Notably, the DICE score for the background class increased to 0.900, and the scores for particles and DNA filaments improved to 0.599 and 0.374, respectively. This result indicates that emphasizing broader structural features (low-pass) and fine-grained details (high-pass) enables the model to capture complementary information, improving segmentation accuracy.

The best performance was achieved when low-pass, high-pass filtering, and random Gaussian noise were combined (column "Low-pass, High-pass, Gaussian Noise"). The overall DICE score for all classes increased to 0.416, with notable improvements in previously underperforming classes. For instance, the segmentation accuracy for membranes reached 0.255, while particles improved further to 0.604. Similarly, actin and DNA filaments achieved scores of 0.027 and 0.328, respectively.

The results also highlight the inherent difficulty of the segmentation task due to significant class imbalance in the dataset. While the particles and background classes contain many voxels for training, classes such as membrane and microtubules are far less represented, making them particularly challenging to predict. This imbalance sometimes leads to DICE scores of 0.000, as the model fails to identify these smaller classes entirely. Notably, this study aimed not to develop the best-performing segmentation model but to systematically investigate the influence of different components and transformations on the segmentation performance. For this reason, no attempts were made to improve this part by experimenting with alternative loss functions or modifying the network architecture.

Another key challenge lies in predicting the actin class, which is exceptionally difficult because its structure requires more contextual information beyond single voxel embeddings or close neighborhoods. The reliance on 1×1 kernels in this ablation study limits the model's ability to incorporate broader spatial relationships, which are critical for detecting elongated or sparse structures like actin. These results emphasize the complexity of the problem while providing valuable insights into how individual design choices affect segmentation outcomes under these challenging conditions.

The results demonstrate that combining the proposed transformations is essential for achieving robust and generalizable voxel embeddings. These transformations address the variability in cryo-ET data, enhancing the model's ability to segment structural components across diverse tomographic volumes accurately.

4.2.2 *Masking out of voxels*

This ablation study evaluates the effect of the proposed transformation of masking out voxels on the segmentation performance. The transformation involves randomly masking out a specific percentage of voxels during training to encourage the model to rely on contextual information from neighboring

Classes	0%	10%	25%	50%	75%	90%
background	0.860	0.856	0.832	**0.870**	0.531	0.147
membrane	**0.236**	0.000	0.203	0.000	0.161	0.017
microtubules	0.166	0.087	0.000	0.000	**0.235**	0.013
particles	0.589	0.532	0.542	**0.615**	0.484	0.406
actin	0.000	0.000	0.000	0.000	**0.030**	0.013
dna filaments	0.351	0.302	0.252	**0.384**	0.132	0.306
all	**0.431**	0.363	0.376	0.379	0.286	0.153

Table 2: **Ablation study results for the masking out of voxels transformation.** The table presents DICE scores for varying masking percentages (0%, 10%, 25%, 50%, 75%, and 90%). While 0% masking achieves the highest overall score, 50% masking balances robustness and accuracy, particularly improving the segmentation of particles and DNA filaments.

voxels rather than individual voxel values. Table 2 reports the DICE scores for varying masking percentages: 0% (no masking), 10%, 25%, 50%, 75%, and 90%.

The results show that the 0% masking setting, where no voxels are masked, achieves the highest overall DICE score of 0.431. These results suggest that when no masking is applied, the model performs best regarding raw segmentation accuracy because it can rely directly on the voxel-level information. However, while 0% masking achieves the highest score overall, it does not force the model to learn robust contextual features, which may limit generalization in more challenging conditions.

A masking percentage of 50% achieves the next-best overall performance with a DICE score of 0.379, striking a balance between robustness and accuracy. At this masking level, the particles class achieves its highest score of 0.615, and the DNA filaments class improves to 0.384, demonstrating that moderate masking helps the model learn to incorporate neighborhood information while retaining sufficient structural detail. The background class also performs strongly at 0.870, indicating that the model effectively segments dominant classes under this masking condition.

At lower masking percentages (10% and 25%), the performance decreases slightly compared to the 0% baseline, with overall DICE scores of 0.363 and 0.376, respectively. This suggests minimal masking does not sufficiently challenge the model to learn beyond direct voxel-level information.

Conversely, higher masking percentages (75% and 90%) lead to a significant drop in segmentation performance, with overall DICE scores declining to 0.286 and 0.153, respectively. For example, at 75% masking, the microtubules class achieves its highest score of 0.235, indicating that some structural features remain detectable even with considerable masking. However, excessive masking at these levels removes too much information, causing the model to struggle with identifying fine-grained structures like membranes and actin, where spatial context is already limited.

In summary, while 0% masking achieves the best overall segmentation performance, the 50% masking level represents a meaningful trade-off. It chal-

Classes	8-d	16-d	32-d	64-d
background	0.841	0.808	0.000	**0.887**
membrane	0.000	0.000	0.137	**0.252**
microtubules	0.000	**0.240**	0.010	0.210
particles	0.567	0.556	0.526	**0.620**
actin	0.000	0.033	0.000	**0.056**
dna filaments	0.254	0.202	0.301	**0.411**
all	0.349	0.367	0.175	**0.466**

Table 3: **Ablation study results for the size of voxel embeddings**. The table presents DICE scores for different voxel embedding sizes (8-d, 16-d, 32-d, and 64-d). The 64-d embeddings achieve the best overall performance, balancing representational richness and generalization across structural classes.

lenges the model to learn robust, context-aware embeddings without excessively degrading structural information. It is particularly valuable for improving the segmentation of smaller or more complex classes, such as particles and DNA filaments.

4.2.3 *Voxel embeddings size*

This ablation study evaluates the effect of varying voxel embedding sizes on the segmentation performance. The voxel embedding sizes tested were 8-d, 16-d, 32-d, and 64-d, with the results presented in Table 3.

The results show that the 64-d embedding size achieves the best overall performance with a DICE score of 0.466, outperforming the other embedding sizes across most classes. Notably, the background class achieves its highest score of 0.887, while the particles and DNA filaments classes reach 0.620 and 0.411, respectively. The segmentation of smaller or more complex classes, such as membrane and actin, also improves slightly, with scores of 0.252 and 0.056.

The overall performance is suboptimal when using smaller embedding sizes, such as 8-d and 16-d, with DICE scores of 0.349 and 0.367, respectively. While these smaller embeddings perform reasonably well on the background and particle classes, they fail to capture sufficient feature detail for more challenging classes, such as membrane and microtubules, which remain at or near zero. These results suggest that smaller embeddings lack the representational capacity to encode the structural complexity of the tomograms.

Increasing the embedding size to 32-d results in a decline in overall performance, with the DICE score dropping to 0.175. This reduction can be attributed to overfitting or difficulty in optimizing for the background class, where the DICE score drops to 0.000, indicating that the segmentation heads failed to generalize when trained with this size.

The 64-d embedding size strikes the optimal balance between feature richness and model stability. This size allows the model to encode sufficient structural detail for complex classes like particles and DNA filaments while maintaining strong performance on dominant classes like the background. The results demonstrate that increasing the embedding size improves the represen-

Classes	128-d	256-d	512-d	1,024-d
background	0.000	**0.904**	0.859	0.888
membrane	**0.184**	0.000	0.000	0.000
microtubules	**0.007**	0.000	0.000	0.000
particles	0.603	**0.674**	0.594	0.635
actin	**0.044**	0.000	0.000	0.000
dna filaments	0.298	**0.537**	0.288	0.393
all	0.205	**0.416**	0.364	0.389

Table 4: **Ablation study results for the size of global embeddings.** The table presents DICE scores for global embedding sizes of 128-d, 256-d, 512-d, and 1,024-d. The 256-d embeddings achieved the highest overall performance, balancing structural feature representation and model stability.

tational power of the voxel embeddings, but only up to a certain point, beyond which performance degrades.

In conclusion, the 64-d embedding size was selected as the optimal configuration because it achieves the highest overall DICE score and consistently improves segmentation performance across multiple structural classes.

4.2.4 *Global embeddings size*

An ablation study was conducted for the optimal size of the global embeddings in the DenseSimSiam model by evaluating the segmentation performance across four different embedding dimensions: 128-d, 256-d, 512-d, and 1,024-d. The results, presented in Table 4, highlight the importance of selecting an embedding size that balances representational capacity and model generalization.

The 256-d global embedding size achieved the best overall performance with a DICE score 0.416. This dimension consistently outperformed the other sizes across several classes. Notably, the background class reached a high DICE score of 0.904, while the particles and DNA filaments classes achieved 0.674 and 0.537, respectively. This result indicates that a 256-dimensional embedding provides sufficient capacity to encode meaningful structural features without overfitting or introducing instability.

The overall performance was significantly lower at the smaller embedding size of 128-d, with a DICE score of 0.205. While the particle class retained reasonable performance at 0.603, most other classes, such as background and microtubules, exhibited near-zero scores. Therefore, the 128-d embeddings lack the representational power to capture complex structural information, particularly for dominant or sparsely represented classes.

Increasing the embedding size to 512-d led to a drop in overall performance, with the DICE score decreasing to 0.364. Although the particles class retained moderate performance at 0.594, other classes, such as microtubules and membranes, scored 0.000, indicating instability. Similar trends are observed for the 1,024-d embeddings, where the overall DICE score slightly improved to 0.389. However, individual class performance, particularly for background

Classes	Dense	Global	L2	L4	L8
background	0.000	0.701	**0.891**	0.879	0.000
membrane	0.020	**0.270**	0.000	0.000	0.000
microtubules	0.095	0.202	**0.342**	0.000	0.256
particles	0.422	0.613	0.616	**0.619**	0.413
actin	0.015	0.039	**0.113**	0.000	0.000
dna filaments	0.212	0.313	**0.428**	0.308	0.292
all	0.132	0.402	**0.452**	0.375	0.159

Table 5: **Ablation study results for the influence of different embedding losses.** The table shows DICE scores for using only voxel-level embeddings (Dense) loss, adding global embeddings (Global), and adding level embeddings (L2, L4, and L8) loss with levels from pooling factors of 2, 4, and 8, respectively. Including the L2-level loss to the dense and global losses achieves the highest overall performance, balancing fine-scale detail and global context.

and particles, failed to match the results of the 256-d configuration. Therefore, excessively large embeddings may introduce unnecessary complexity, making optimization more challenging without providing meaningful gains.

In conclusion, the 256-d global embedding size emerged as the optimal configuration, achieving the best trade-off between representational capacity and model stability. This size allows the DenseSimSiam model to encode sufficient structural detail across various classes while avoiding the pitfalls of underfitting with smaller embeddings and overfitting with huge ones.

4.2.5 *Influence of the different losses*

This ablation study evaluates the impact of including losses at different levels of the embedding hierarchy within the DenseSimSiam framework. The configurations include the use of voxel-level dense embeddings loss (Dense), global embeddings loss (Global), and losses applied to L2, L4, and L8-level embeddings, which correspond to pooling factors of the FPN decoder of 2, 4, and 8, respectively. The results, summarized in Table 5, demonstrate that combining multiple levels of embeddings leads to better segmentation performance.

Using only the voxel-level dense embedding loss (Dense) results in the lowest overall performance, with a DICE score of 0.132. This configuration focuses solely on local voxel-level information without incorporating any broader context, which limits the model's ability to segment larger or more complex structures. Consequently, most classes exhibit poor scores, highlighting the importance of including global information for learning meaningful voxel embeddings.

Adding the global embeddings loss (Global) improves the performance significantly, achieving an overall DICE score of 0.402. The inclusion of global information enables the model to capture high-level contextual features better, as seen in the strong performance for the background class (0.701) and notable improvements for particles (0.613). However, the microtubules (0.202) and actin (0.039) classes remain challenging to segment accurately. These re-

sults highlight the benefit of incorporating global information during training to improve the voxel embeggings.

Incorporating the loss from the level embeddings at the L2-level (layer with pooling factor of 2) achieves the best overall performance with a DICE score of 0.452. This configuration effectively balances local detail and global context, making notable improvements across most classes. The background class achieves its highest score (0.891), while particles and DNA filaments also benefit, reaching DICE scores of 0.616 and 0.428, respectively. By incorporating more context than voxel-level embeddings alone, the L2-level loss allows the model to capture fine-grained details and broader structural information in the voxel embeddings, resulting in superior performance across diverse structural components.

On the other hand, including losses at the L4-level and L8-level results in performance degradation. The L4-level embedding loss (with a pooling factor of 4) yields an overall DICE score of 0.375. While the particle class reaches its highest score (0.619), other classes, such as background (0.879) and DNA filaments (0.308), show noticeable declines. Adding the loss from the L8-level embeddings decreases the overall DICE score to 0.159. Most classes, including background, membrane, and actin, exhibit near-zero scores. This deterioration occurs because incorporating the loss from the L8-level embeddings introduces excessive global information focus, which biases the voxel embeddings and causes the model to lose critical fine-scale detail required for precise voxel-level segmentation.

In conclusion, the results demonstrate that including the L2-level loss achieves the highest overall performance, with a DICE score of 0.452. This level effectively balances broader contextual information and fine-grained spatial detail, leading to improvements for dominant classes such as background (0.891) and structural components like particles (0.616) and DNA filaments (0.428). In contrast, adding losses at higher pooling levels, such as L4 and L8, results in performance drops, with lower overall DICE scores (0.375 and 0.159, respectively). These findings suggest that while adding level embedding losses introduces a broader context, adding level embedding loss from higher pooling levels reduces the model's ability to retain fine-grained spatial details in the voxel embeddings.

4.3 RESULTS

This section presents the results of the experiments conducted using the CryoSiam framework on the CryoETSim dataset. The findings are structured into two main parts: (1) an analysis of the learned embeddings, showcasing their quality and separability using dimensionality reduction techniques such as Principal Component Analysis (PCA) [173] and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [174] visualizations, and (2) the evaluation of downstream tasks, including semantic segmentation, instance segmentation, and particle identification.

The embedding quality is critical for understanding how well the self-supervised training captures structural patterns and encodes meaningful features. PCA and UMAP visualizations intuitively assess how different

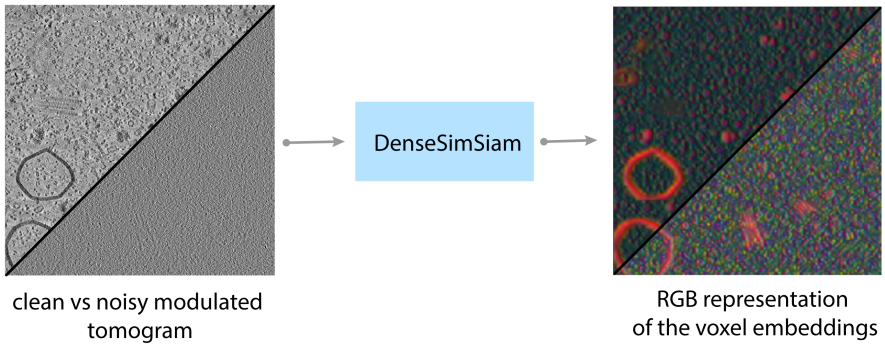

Figure 11: **Voxel embedding generation and visualization**. Clean and noisy tomograms are processed by the DenseSimSiam model to generate voxel embeddings. The embeddings are projected into three principal components using PCA and visualized as RGB images, highlighting structural features within the tomograms.

particle classes or structural components are represented in the learned embedding space. These visualizations set the stage for evaluating how effectively the framework leverages these embeddings for downstream tasks. The subsequent subsections detail the results for denoising, semantic segmentation, instance segmentation, and particle identification tasks, presenting both quantitative metrics and qualitative visualizations. Each task is analyzed in the context of its performance in various structural classes, highlighting strengths, challenges, and opportunities for further refinement.

The quality of voxel-level embeddings generated by DenseSimSiam was assessed using a PCA-based RGB representation. This visualization maps the first three principal components of the voxel embeddings into the RGB color space, offering an intuitive view of how structural features are encoded.

Figure 11 illustrates the PCA visualization for clean and noisy tomograms processed by DenseSimSiam. The left panel shows the input tomograms, while the right panel presents the corresponding RGB representations of the learned voxel embeddings. The embeddings encode distinct structural components such as particles, membranes, and other features, as evidenced by their distinct colors. Notably, regions corresponding to membranes and particles are highlighted, demonstrating the model's ability to extract meaningful features from volumetric data.

The PCA visualization highlights DenseSimSiam's robustness to noise. The embeddings maintain clear distinctions between structural components regardless of whether the input tomogram is clean or noisy. This invariance is achieved through data transformations during self-supervised training, which enhances the model's ability to generalize across different imaging conditions.

The subtomogram embeddings generated by the SimSiam framework were analyzed to assess their ability to capture meaningful structural and molecular features. Subtomograms, representing individual particles, were input into the SimSiam model, which produced embeddings in a high-dimensional space. UMAP was employed to reduce the embedding dimensionality to two

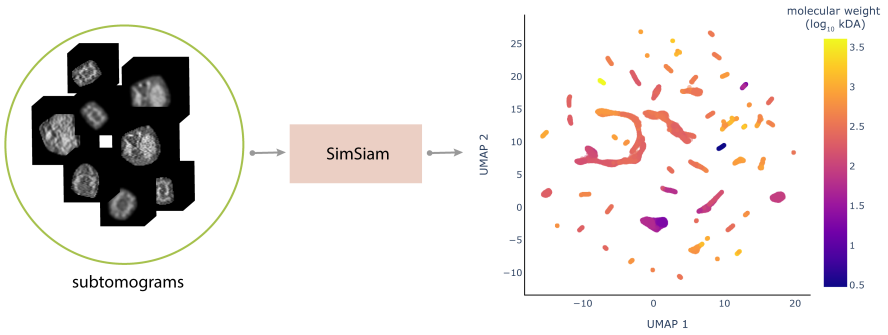

Figure 12: **Subtomogram embeddings and UMAP visualization.** Subtomograms are processed through the SimSiam model to generate embeddings, which are projected into a two-dimensional space using UMAP. The embeddings are color-coded based on molecular weight ($\log_{10}$ scale in kDa), highlighting the clustering of particles according to size and structural similarity. This visualization demonstrates the ability of the SimSiam framework to encode meaningful features that reflect the underlying molecular properties.

dimensions to visualize these embeddings, enabling intuitive visualization of clustering patterns.

The resulting UMAP visualization (Figure 12) demonstrates the separation and clustering of particles based on their molecular weights, which are represented on a $\log_{10}$ scale in kilodaltons (kDa). The embeddings are color-coded according to molecular weight, revealing distinct regions in the UMAP plot corresponding to different particle classes. This clustering highlights the model's capacity to encode subtomogram features that reflect structural similarities and underlying molecular properties.

The SimSiam model ensures robust representation learning by capturing and distinguishing these features, facilitating downstream tasks such as particle identification and classification. The effectiveness of the embeddings in differentiating particle types and sizes underscores the utility of the self-supervised learning approach in analyzing cryo-ET data. This analysis further validates the ability of the CryoSiam framework to generalize across diverse particle types, enabling accurate structural interpretation from tomographic data.

4.3.1 *Denoising results*

The first downstream task in the CryoSiam framework is tomogram denoising, which serves as a foundational step for improving the quality of input data. This task aims to reduce noise in tomograms while restoring structural details and enhancing contrast lost due to CTF modulation. Leveraging the voxel embeddings generated by DenseSimSiam, the denoising module effectively processes the high noise levels and imaging artifacts inherent in cryo-ET data.

To train the denoising model, the input was a noisy and CTF-modulated tomogram, while the output was the corresponding clean tomogram, free from noise and CTF effects. This training configuration enables the model to perform noise suppression, and addresses contrast degradation caused by CTF

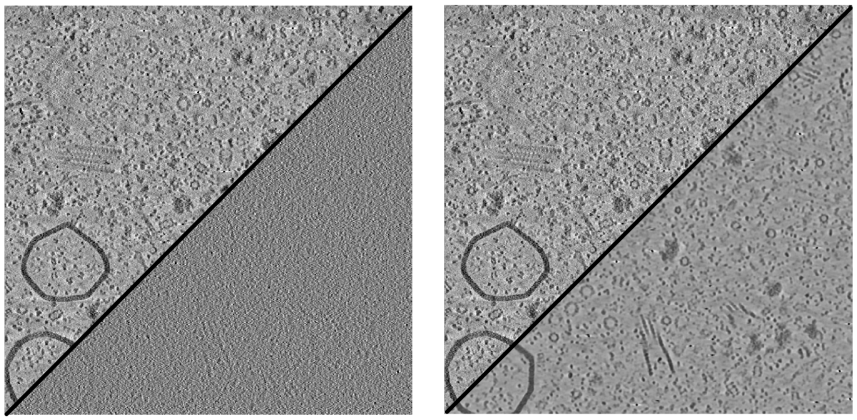

Figure 13: **Denoising of the simulated tomograms**. The left panel shows a clean tomogram (top-left) alongside its noise-modulated counterpart (bottom-right), illustrating the degradation introduced by noise and CTF effects. The right panel presents the clean tomogram (top-left) compared to the CryoSiam-denoised tomogram (bottom-right), highlighting the restoration of contrast and structural details.

modulation, thereby enhancing the visibility of structural details in the tomograms. This dual objective of noise reduction and contrast improvement ensures that the tomograms are suitable for embedding generation and subsequent segmentation tasks.

Figure 20 presents qualitative results from the denoising task, showcasing a comparison between noisy, CTF-modulated input tomograms and their denoised counterparts. The denoised outputs exhibit significantly enhanced clarity, with structural features such as membranes, particles, and filaments becoming more distinguishable. These visual improvements highlight the model's ability to retain key structural features while reducing noise and correcting contrast distortions.

As the first downstream task, denoising sets the stage for robust and reliable performance in subsequent tasks by ensuring high-quality tomograms well-suited for embedding generation and segmentation. This critical step ensures that downstream tasks, such as semantic segmentation and particle identification, operate on less noisy and more interpretable tomograms.

4.3.2 *Semantic segmentation results*

The downstream semantic segmentation task aimed to classify tomogram voxels into five distinct structural classes: background, membranes, microtubules, particles, and actin (Figure 14a). An additional filament class was included for the transcription-translation and chromatin samples to represent DNA and RNA filaments. These predictions are achieved by leveraging the voxel embeddings generated by the DenseSimSiam framework, which encodes fine-grained structural and contextual information. The embeddings are processed by task-

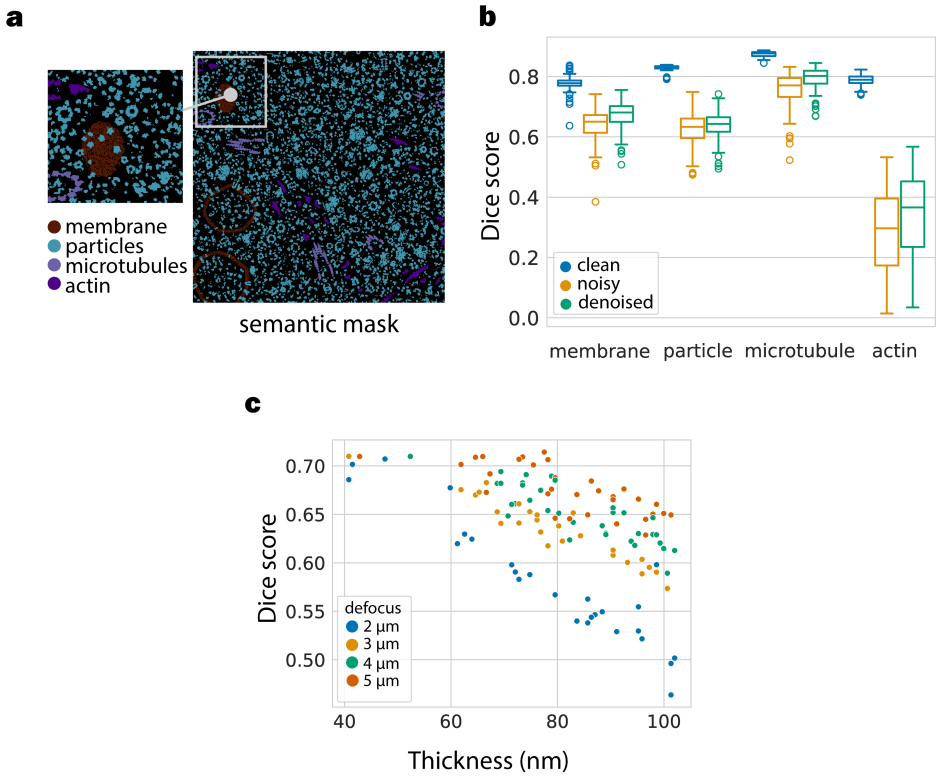

Figure 14: **Semantic segmentation performance under different noise levels and imaging conditions.**
a) Example semantic mask highlighting the segmentation of membranes (brown), particles (blue), microtubules (violet), and actin (purple). b) DICE scores for each class (membrane, particle, microtubule, actin) across individual test tomograms, evaluated on models trained with clean, noisy, and denoised tomograms. This panel illustrates the variability in segmentation performance across tomograms and classes. c) Average DICE scores across all semantic classes plotted against sample thickness for various defocus settings (2 µm, 3 µm, 4 µm, 5 µm), revealing the impact of thickness and defocus on segmentation accuracy.

specific CNN heads and designed to optimize voxel-level predictions through cross-entropy, generalized DICE, and MSE losses.

For training the semantic segmentation models, the encoder of the DenseSimSiam framework was frozen, ensuring that the pre-trained voxel embeddings remained unchanged. The decoder and task-specific CNN heads were then trained to map these embeddings to accurate voxel-level predictions. This setup allowed the framework to fine-tune the downstream components without altering the foundational representations established during the self-supervised pretraining phase.

The DICE scores for semantic segmentation were evaluated for models trained on clean, noisy, and denoised tomograms, with separate models dedicated to each dataset type. The denoised tomograms were generated using the denoising subtask described earlier to reduce noise and recover contrast lost due to CTF modulation. As summarized in Table 6, the model trained on clean tomograms achieved the highest overall DICE score of 0.836,

Classes	clean	noisy	denoised
background	0.944	0.894	0.896
membrane	0.778	0.639	0.671
microtubules	0.874	0.754	0.793
particles	0.822	0.619	0.636
actin	0.788	0.280	0.342
dna filaments	0.703	0.332	0.350
all	0.836	0.635	0.664

Table 6: **DICE scores for semantic segmentation models on clean, noisy, and denoised tomograms.** Each dataset type was used to train a separate model, with denoised tomograms generated using the denoising subtask. The clean data achieves the highest overall performance, while denoising improves segmentation accuracy relative to noisy data.

significantly outperforming the models trained on noisy and denoised data, which achieved scores of 0.635 and 0.664, respectively. Although denoising improved performance compared to noisy data, the results indicate that it did not fully recover the segmentation accuracy achieved with clean data, likely due to the challenges of restoring fine structural details lost during the imaging process.

Class-specific analysis reveals distinct trends in segmentation performance under varying conditions. The background class consistently achieved the highest accuracy across all datasets, with DICE scores ranging from 0.944 for clean tomograms to 0.896 for denoised data. This stability reflects the abundance and simplicity of background voxels, which lack intricate structural features. In contrast, structural classes such as membranes and microtubules demonstrated moderate robustness to noise. For instance, microtubules achieved a DICE score of 0.754 on noisy data, improving to 0.793 with denoised tomograms. At the same time, membranes exhibited greater sensitivity, with scores dropping from 0.778 for clean tomograms to 0.639 for noisy tomograms and recovering slightly to 0.671 after denoising. These results highlight the model's ability to leverage repetitive structural features in microtubules but also underscore the challenges of segmenting thin and complex structures like membranes under degraded conditions.

Particles, which form a central focus of this task, were significantly affected by noise, with DICE scores decreasing from 0.822 on clean tomograms to 0.619 on noisy tomograms. Although denoising improved performance to 0.636, particles' complex interaction and overlap with other structural components limited recovery. More intricate classes, such as actin and DNA filaments, exhibited the most significant sensitivity to noise and modulation effects. Actin segmentation dropped sharply from 0.788 on clean tomograms to 0.280 on noisy tomograms, recovering marginally to 0.342 with denoised data. Similarly, DNA filaments in transcription-translation and chromatin samples followed a similar trend, with DICE scores decreasing to 0.332 for noisy tomograms and improving slightly to 0.350 after denoising. The lower representation of actin and DNA filaments in the dataset, combined with their thin and

intertwined structural features, likely contributed to these challenges, as these structures require both high contextual information and fine voxel-level details for accurate segmentation.

The impact of noise, denoising, and imaging conditions on segmentation performance is further illustrated in Figure 14. Panel b displays DICE scores for each semantic class across individual test tomograms, highlighting the variability in segmentation performance. For example, while membrane and microtubule classes exhibit relatively consistent DICE scores, the particle and actin classes show a wider range of variability, particularly for noisy and denoised datasets. This variability underscores the influence of individual tomogram characteristics, such as structural complexity or imaging conditions, on segmentation accuracy.

Panel c examines the relationship between DICE scores, sample thickness, and defocus levels. Here, the DICE scores averaged across all semantic classes are plotted against sample thickness for various defocus settings. The results show a clear negative correlation between sample thickness and segmentation performance, with thicker samples leading to lower DICE scores. Additionally, higher defocus values exacerbate this decline, particularly for samples exceeding 80 nm in thickness. These trends indicate that increased sample thickness and reduced contrast at higher defocus levels significantly hinder the model's ability to segment structural components accurately.

The semantic segmentation results demonstrate the capability of the DenseSimSiam framework to adapt to varying noise levels and imaging conditions while achieving reliable segmentation of structural components. High performance on clean data highlights the model's potential in optimal scenarios. In contrast, noisy and denoised data results emphasize the importance of preprocessing steps like denoising to mitigate noise-induced challenges. Variability in class-specific performance reflects the differing complexities of structural features, with actin posing significant challenges due to its filamentous structure and reliance on fine-grained contextual information. The observed correlations between sample thickness, defocus, and segmentation accuracy further illustrate the sensitivity of segmentation performance to tomographic properties. These findings underscore the need for tailored strategies to enhance segmentation robustness across diverse experimental conditions and structural complexities.

4.3.3 *Instance segmentation results*

The instance segmentation task in the CryoSiam framework aimed to identify and segment individual particle instances within tomograms by leveraging boundary prediction, distance from boundaries, and instance labels. The workflow involved generating initial segmentations using the watershed algorithm, guided by the predicted distances and boundaries, and refining these segmentations with a multi-cut algorithm to ensure precise delineation of individual particles (Figure 15a). This combination of methods proved critical for separating nearby particles and generating coherent masks for downstream particle identification.

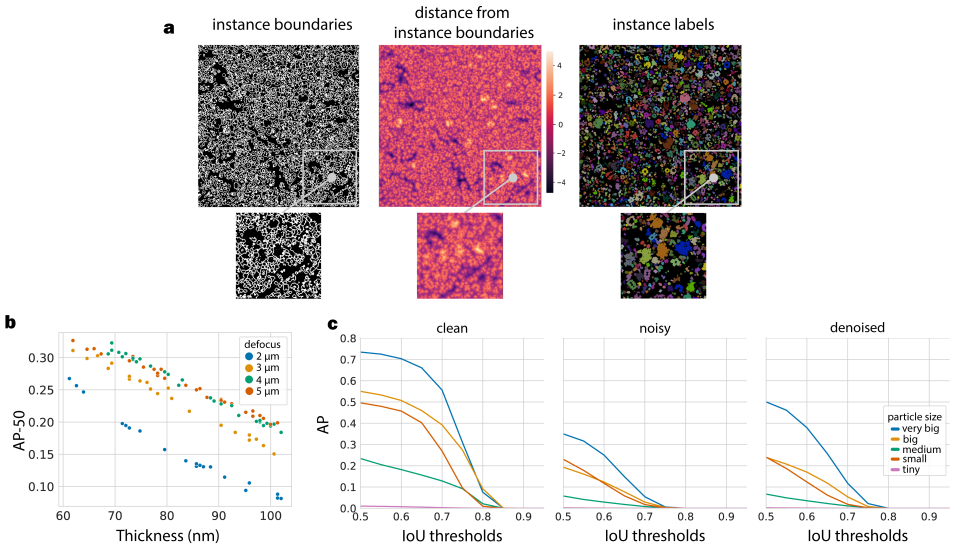

Figure 15: **Instance segmentation performance under different noise levels and imaging conditions.**
(a) Example outputs for instance segmentation showing predicted boundaries, distance from boundaries, and instance labels. (b) AP-50 scores for instance segmentation across tomograms of varying thickness and defocus levels, illustrating the decline in performance with increasing thickness and defocus. (c) AP scores at varying IoU thresholds, separated by particle sizes, for models trained on clean, noisy, and denoised tomograms. Larger particles exhibit higher AP scores, while smaller particles are more affected by noise and CTF modulation.

The performance of the instance segmentation task was evaluated using the AP metric at varying IoU thresholds, providing a quantitative measure of the model's ability to segment and delineate individual particles accurately under different data conditions. Table 7 summarizes the results for models trained on clean, noisy, and denoised tomograms. The model trained on clean tomograms achieved the highest AP-50 score of 0.437, reflecting its capacity to segment particles with minimal interference from imaging artifacts. In contrast, the noisy model yielded a much lower AP-50 score of 0.229, demonstrating the detrimental impact of noise and CTF modulation on the model's ability to differentiate particle boundaries and foreground regions. The denoised model performed slightly better than the noisy model, with an AP-50 score of 0.249, indicating that the denoising step partially mitigated the effects of noise but was insufficient to restore segmentation accuracy to the level achieved with clean tomograms.

The decline in AP scores as the IoU threshold increased further underscores the difficulty of achieving high segmentation precision under stringent overlap requirements. This trend reveals that even for clean tomograms, where particle boundaries are more clearly defined, higher IoU thresholds demand precise alignment between the predicted and ground truth masks, making minor segmentation errors more penalizing. These errors become more pronounced for noisy and denoised data due to the blurring of particle boundaries and the loss of fine structural details, which are critical for achieving high IoU matches.

AP	clean	noisy	denoised
AP-50	0.437	0.229	0.249
AP-55	0.417	0.180	0.196
AP-60	0.393	0.121	0.133
AP-65	0.346	0.063	0.071
AP-70	0.241	0.020	0.024
AP-75	0.099	0.003	0.003
AP-80	0.016	0.000	0.000

Table 7: **AP scores for instance segmentation at varying IoU thresholds.** The table compares AP scores for models trained on clean, noisy, and denoised tomograms. Results highlight the significant impact of noise and CTF modulation on segmentation performance, with the clean tomograms achieving the highest AP scores across all IoU thresholds. The denoised tomograms improve performance relative to noisy tomograms but do not fully recover the performance seen with clean data.

The inability of the denoised model to match the clean-data performance highlights a key limitation: while denoising can enhance contrast and reduce noise, it often results in a loss of high-frequency details necessary for accurate particle boundary delineation. For instance, denoising may smooth out small, intricate structures or merge nearby features, making it more challenging for the model to identify individual instances correctly. This finding suggests that while denoising is beneficial as a preprocessing step, it cannot fully compensate for the absence of clean data, emphasizing the importance of high-quality training datasets.

The impact of tomogram thickness and defocus on instance segmentation performance from noisy data is illustrated in Figure 15b, highlighting the interplay between sample properties and segmentation accuracy. Across all datasets, clean, noisy, and denoised, increasing tomogram thickness led to a consistent decline in AP-50 scores. This trend underscores the inherent difficulty of segmenting particles in thicker tomograms, where overlapping structures, occlusions, and cumulative electron scattering reduce the visibility of individual particles. In thicker samples, structural details become obscured, making it more challenging for the model to delineate particle boundaries and accurately segment individual instances.

The challenges posed by increased thickness are worsened by higher defocus levels, which further reduce image contrast and clarity. In noisy and denoised tomograms, the combination of high defocus and increased thickness significantly degrades segmentation performance. This degradation occurs because noise and denoising artifacts further obscure particle boundaries and fine structural details, which are difficult to discern in thick samples. The noisy dataset demonstrates a sharper decline in AP-50 scores with increasing thickness and defocus, as noise amplifies the difficulty of resolving individual particles in cluttered and low-contrast regions. The results highlight the critical role of imaging conditions in determining segmentation performance.

Figure 15c closely examines AP scores across various IoU thresholds, categorized by particle size. The particles were classified into five categories based on the number of voxels comprising each instance: very big (>6000 voxels), big (2000–6000 voxels), medium (900–2000 voxels), small (150–900 voxels), and tiny (<150 voxels). This classification highlights the varying segmentation challenges associated with different particle sizes. The results reveal a distinct trend: larger particles, such as those in the very big and big classes, consistently achieve higher AP scores. This performance demonstrates the model's capability to segment well-defined and prominent structural features effectively. Larger particles typically have clear boundaries and strong contrast, making them less susceptible to noise and structural distortions.

In contrast, smaller particles, particularly those classified as small and tiny, exhibit significantly lower AP scores, especially in noisy datasets. These results underscore the challenges of segmenting smaller particles, as their structural details are less pronounced and more sensitive to noise and distortion. Noise introduces additional artifacts that obscure boundaries and reduce the clarity of these particles, making segmentation particularly difficult.

Medium-sized particles present a unique challenge due to their more elongated shapes, often resulting in reduced contrast and less defined boundaries. These characteristics make segmentation for this size category more complex than for the more compact larger particles. The denoised dataset shows moderate improvements for medium-sized particles, indicating that the denoising process successfully enhances contrast and suppresses noise for these intermediate structures. However, the contrast and elongated morphology limitations still hinder optimal segmentation performance.

The denoised dataset achieves slightly better AP scores than the noisy dataset for larger particles. This suggests that the benefits of denoising are more pronounced for particles with inherently stronger structural features, where noise suppression and contrast enhancement can aid segmentation. However, these benefits do not fully translate to smaller particles, where the loss of detail from denoising counterbalances its advantages.

These findings emphasize the importance of developing methods that preserve fine structural details and enhance contrast during preprocessing and denoising. Such approaches would significantly improve the segmentation performance for small and elongated particles, ensuring that the embeddings effectively capture the nuanced features required for accurate identification and delineation across diverse particle sizes and imaging conditions.

These findings underscore the critical role of clean, noise-free data in achieving high-performance instance segmentation. Models trained on such data consistently outperformed those trained on noisy or denoised tomograms, demonstrating the value of preserving structural integrity and detail during training. The segmentation performance was susceptible to imaging artifacts, sample properties, and particle size, highlighting the task's inherent complexity. While denoising partially alleviated the impact of noise, accurately segmenting small and intricate particles remains particularly challenging, especially in thick or highly defocused tomograms where contrast and structural clarity are diminished. Integrating watershed segmentation with multi-cut refinement proved essential for generating reliable and coherent segmentations,

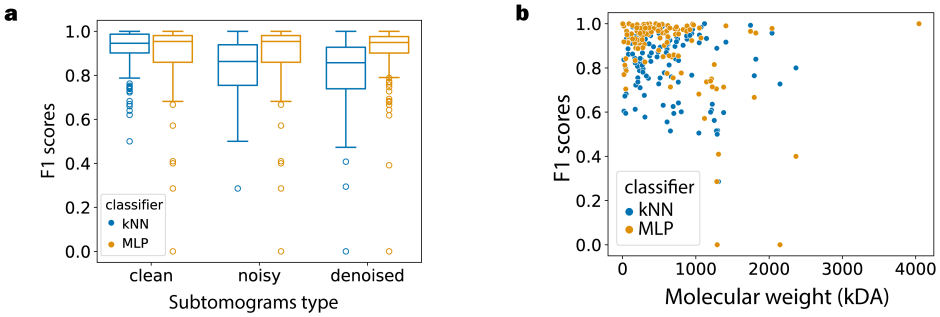

Figure 16: **F1 scores for particle identification across tomogram types and molecular weight distributions.**
(a) F1 score comparisons between k-NN and MLP classifiers for clean, noisy, and denoised tomograms, showing the influence of tomogram quality on classification performance. The MLP consistently outperforms k-NN, particularly for noisy data. (b) F1 scores are plotted against the molecular weight of particles (log scale), illustrating classifier performance across different particle sizes. Larger particles yield higher F1 scores for both classifiers, with MLP showing better performance for medium and small particles.

providing a strong foundation for robust downstream particle identification and analysis.

4.3.4 *Particle identification results*

The particle identification task involved classifying subtomograms into their respective molecular weight-based particle categories. This evaluation was conducted using embeddings generated by the SimSiam framework, followed by classification using either a k-NN classifier or an MLP classifier. Both classifiers were tested across datasets derived from clean, noisy, and denoised tomograms to assess the robustness of the embeddings and classification pipeline under varying conditions.

Figure 16a compares the F1 scores achieved by the k-NN and MLP classifiers for each tomogram type. Models trained on clean tomograms yielded the highest F1 scores, mainly when using the MLP classifier, which consistently outperformed k-NN. Both classifiers exhibited reduced performance for noisy and denoised datasets, though the MLP classifier demonstrated greater resilience to noise. This suggests that the MLP classifier can leverage the learned subtomogram embeddings more effectively, particularly in challenging conditions.

Figure 16b shows the F1 scores against the molecular weight of particles, focusing on noisy data. Larger particles consistently achieved higher F1 scores for both classifiers, likely due to their distinct structural features. In contrast, smaller particles exhibited significantly lower F1 scores, reflecting the challenge of identifying particles with reduced contrast and subtle structural details. The k-NN classifier, configured with $k = 1$, was particularly limited for medium and small particles, relying on simple nearest-neighbor logic. This

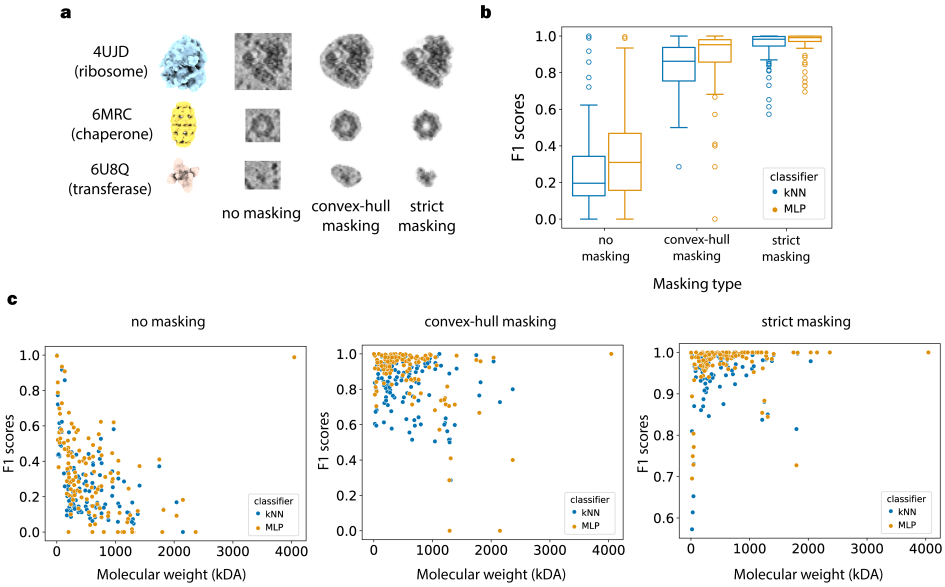

Figure 17: **Impact of masking strategies on particle identification performance.** (a) Examples of particle representations under different masking strategies: no masking, convex-hull masking, and strict masking, highlighting how masking influences structural details visible to the classifiers. (b) F1 score distributions for k-NN and MLP classifiers across masking types, with strict masking yielding the best performance. Convex-hull masking balances structural preservation and noise reduction, while no masking leads to lower scores due to higher background interference. (c) F1 scores versus molecular weight for each masking type, showing improved classification performance across all masking strategies, with strict masking resulting in the most consistent accuracy across particle sizes.

suggests that the embeddings from different particle types may not have been well-separated in the embedding space for k-NN to perform effectively. By contrast, the MLP classifier, with its ability to learn more complex decision boundaries, leveraged nuanced patterns in the embeddings to achieve better classification performance.

Figure 17a visually compares representative particle subtomograms under three masking strategies: no masking, convex-hull masking, and strict masking. Under the no masking condition, the particles are embedded within the original tomographic volume, including surrounding background noise and neighboring structures. This lack of isolation makes it challenging to focus solely on the particle features, as noise and adjacent elements can interfere with the embeddings, particularly for smaller particles where noise dominates the subtomogram.

In the convex-hull masking condition, the particles are enclosed within a geometric mask that roughly captures their shape, providing a more precise representation of the particle while still including some surrounding voxels. This approach reduces background noise while preserving structural context, offering a balanced view that retains sufficient information for identification without introducing excessive noise. The effects of this masking are pronounced for smaller particles, where the SNR is higher, and the suppression of background noise helps highlight the structural features.

The strict masking condition tightly isolates the particles, removing all surrounding background voxels and focusing entirely on the particle's core structure. While this strategy maximally suppresses noise and extraneous elements, it may also exclude contextual information that could be important for identifying complex or elongated particles. The lack of surrounding context in strict masking is especially noticeable for smaller particles, where structural features alone may not be sufficient for reliable identification.

These visualizations demonstrate the trade-offs inherent in each masking strategy. No masking provides the least structural clarity, convex-hull masking strikes a balance between clarity and context, and strict masking emphasizes particle features at the cost of removing structural context. The effects of these strategies are pronounced for smaller particles, where the suppression of noise and the inclusion or exclusion of structural context can significantly impact particle representation.

Figure 17b shows the F1 score distributions for k-NN and MLP classifiers under no masking, convex-hull masking, and strict masking strategies. Strict masking achieved the highest F1 scores across all particle types, with the MLP classifier outperforming k-NN due to its ability to learn complex decision boundaries. Convex-hull masking provided intermediate performance, balancing structural preservation and noise suppression. In contrast, the no masking strategy resulted in the lowest scores, particularly for k-NN, as background noise and surrounding structures interfered with the embeddings. These results underscore the importance of masking strategies, particularly strict masking, for enhancing particle identification in noisy or complex subtomograms.

Figure 17c explores the relationship between F1 scores and molecular weight for the three masking strategies: no masking, convex-hull masking, and strict masking. Across all strategies, larger particles consistently achieved higher F1 scores, reflecting their stronger structural signals and greater resilience to background interference. Strict masking provided the highest and most stable classification performance for these particles, as it eliminated background regions and allowed the classifier to focus exclusively on particle features.

For smaller particles, the impact of masking strategies was more pronounced. Strict masking significantly improved F1 scores by isolating structural features that are otherwise challenging to identify due to the crowding of tomograms. The densely packed environment typical of real-world tomograms introduces overlapping and closely neighboring particles, making smaller structures harder to classify accurately. Convex-hull masking demonstrated moderate performance, particularly for medium-sized particles, where retaining some background context does not significantly affect performance. However, the no-masking strategy resulted in the poorest outcomes for smaller particles, as embeddings were heavily influenced by surrounding structures, reducing classification accuracy.

These results highlight the importance of masking in particle identification, particularly for addressing the challenges posed by crowded tomograms. Strict masking is the most effective strategy, enabling more consistent and accurate classification across particle sizes.

In particle identification, achieving robust embeddings is critical for enabling classifiers that rely on distance metrics to distinguish particle types,

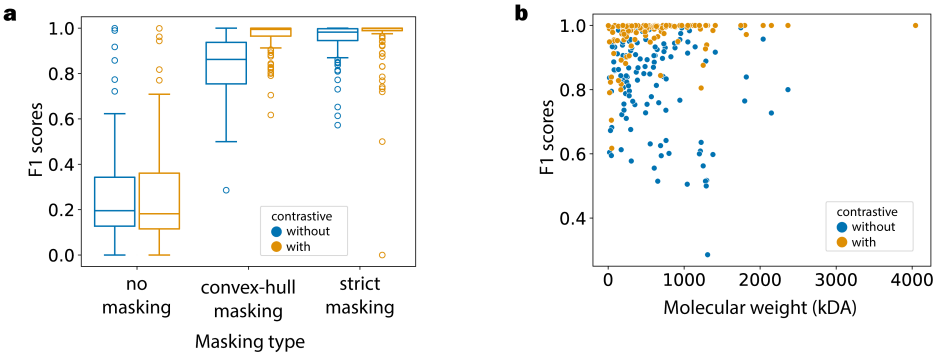

Figure 18: **Impact of contrastive learning on particle identification performance.** (a) Boxplots showing the distribution of F1 scores for k-NN classification across all particles. Contrastive learning (orange) results in a tighter distribution and higher median scores compared to the model without contrastive learning (blue). (b) Scatter plot illustrating the relationship between F1 scores and molecular weight (kDa) for each particle class. Larger particles exhibit consistently high F1 scores, with contrastive learning further improving the separation of smaller and medium-sized particles.

such as k-NN or other distance-based approaches, including nearest centroid classifiers and clustering algorithms. These methods are particularly advantageous in scenarios where fully supervised models, like MLPs, may not be practical due to the limited availability of labeled cryo-ET data. To ensure that the embedding space supports accurate particle classification, two experiments were conducted with the SimSiam model.

In the first experiment, SimSiam was trained solely using self-supervised pretraining. The resulting embeddings were then classified using k-NN with cosine distance, where $k = 1$. The SimSiam model was additionally trained with contrastive learning in the second experiment. Positive and negative pairs were generated, and the model was optimized using an L2 distance-based contrastive loss function. The embeddings from this approach were classified using k-NN with Euclidean distance.

Figure 18a highlights the F1 score distributions for k-NN classification under different masking strategies, comparing models trained with and without contrastive learning. Notable improvements are observed with convex-hull and strict masking, where the model trained with contrastive learning achieves higher median F1 scores and tighter distributions. However, for no masking, the difference between models with and without contrastive learning is minimal, as the absence of masking allows significant interference from nearby particles, introducing noise that undermines the advantages of the refined embedding space. These results indicate that contrastive learning effectively enhances the embedding space for particle classification, mainly when masking strategies are employed to reduce background noise.

Figure 18b further investigates the relationship between F1 scores and molecular weight, comparing the effects of contrastive learning. Both approaches yield high F1 scores for larger particles, reflecting the ease of distinguishing particles with prominent structural features and distinct separation in the embedding space. However, the difference becomes apparent for smaller and

medium-sized particles, where the embeddings generated with contrastive learning result in noticeably higher F1 scores. This improvement highlights the ability of contrastive learning to capture better subtle differences in detailed or overlapping features, which are more prevalent in smaller particles. The embeddings without contrastive learning struggle to adequately separate these challenging cases, leading to lower F1 scores, particularly in the low-molecular-weight range.

These results underscore the value of contrastive learning in enhancing embeddings for distance-based classifiers like k-NN. By focusing on pairwise similarities and differences during training, the contrastive approach enables better separation of particle types, particularly for smaller and more challenging classes. This improvement is crucial for real-world applications where supervised models may not always be feasible.

4.4 ANALYSIS AND INSIGHTS

The experiments and ablation studies conducted in this chapter offer valuable insights into the performance of the CryoSiam framework and its design choices. Through systematic evaluation, several key takeaways have emerged, shaping the development of the final model.

The ablation studies underscored the importance of specific transformations and architectural components in achieving robust segmentation and particle identification. Transformations such as Gaussian noise, low-pass filtering, and voxel masking played a critical role in simulating the variability and artifacts observed in real cryo-ET data. Among these, voxel masking significantly enhanced the quality of voxel embeddings, allowing the model to generalize more effectively across diverse tasks. Similarly, the evaluation of embedding sizes highlighted the need to balance compactness and detail. While smaller voxel and global embeddings improved computational efficiency, they occasionally sacrificed fine-grained structural information. Conversely, excessively large embeddings introduced redundancy without tangible performance gains. These findings informed the selection of embedding sizes in the final model, optimizing computational efficiency and predictive accuracy.

Including multi-scale context through level embeddings proved essential for segmentation tasks. The studies on level loss functions revealed that combining dense, global, and low-pooling-level losses, such as those from the L2 level, offered the best performance. This approach balanced fine structural details with broader contextual information, avoiding the biases and performance degradation observed at higher pooling levels, such as L4 and L8. These insights guided the decision to retain dense and L2-level embeddings in the final model, enabling robust segmentation across varying structural scales.

The evaluation of CryoSiam across downstream tasks highlighted both its strengths and limitations. While the model excelled in segmenting well-defined and dominant classes such as membranes and particles, it faced challenges with underrepresented or structurally complex classes like actin and microtubules. These challenges were attributed to the limited diversity of simulated training data for these classes. To address this limitation, future model iterations could benefit from expanded datasets that simulate a wider

range of structural variations, including filament types and morphologies. Additionally, the inherent difficulty of segmenting small particles, where features often overlap or are obscured, emphasized the need for refined particle identification strategies. This challenge informed the use of contrastive learning to improve the separation of particles in the embedding space.

This chapter's iterative analysis and refinement were pivotal in shaping the CryoSiam framework. The final model design synthesizes findings from ablation studies and task-specific evaluations. Transformations addressing data variability and noise were retained to enhance generalization while embedding sizes were carefully tuned to balance computational efficiency and structural detail. Integrating multi-scale context through dense and low-level embeddings further solidified the model's capability to handle complex tasks effectively.

This chapter's analyses demonstrate CryoSiam's effectiveness in handling simulated cryo-ET datasets while identifying areas for improvement. The insights gained provide a strong foundation for transitioning to real data, which will be explored in the subsequent chapter. They also underscore the potential of the CryoSiam framework in advancing cryo-ET data analysis.

Part III

TRANSLATING SIMULATIONS TO REALITY: CHALLENGES AND ACHIEVEMENTS

Building upon the insights gained from experiments on simulated data, this part transitions to applying the developed model to real cryo-ET datasets. It explores the challenges of transferring models trained on synthetic data to real-world scenarios and presents strategies for fine-tuning and adaptation. The results obtained from real data experiments are compared with existing methods, highlighting the practical implications and limitations of the approach. This part concludes by discussing the broader impact of the research and potential directions for future work.

5 TRANSFER TO REAL CRYO-ET DATA

This chapter marks the transition from evaluating the CryoSiam framework on simulated cryo-ET datasets to its application on real-world data. While simulated datasets offer a controlled environment to test and refine the model rigorously, real cryo-ET data presents unique challenges, including variability in imaging conditions, sample heterogeneity, and incomplete annotations. These challenges provide an opportunity to assess the robustness and adaptability of the CryoSiam framework in practical scenarios.

The chapter introduces the real cryo-ET datasets used for evaluation, highlighting their diversity and biological significance. It then details the methodology employed to adapt and apply the framework to these datasets, leveraging insights from the simulated data experiments. The performance of CryoSiam is evaluated across key tasks such as tomogram denoising, semantic segmentation, instance segmentation, and particle identification. Comparisons with established methods further illustrate the strengths and limitations of the approach. Finally, the chapter reflects on the outcomes of these experiments, identifying areas where the framework succeeds and areas where further refinement is needed.

This chapter demonstrates the practical utility of CryoSiam in advancing cryo-ET analysis by bridging the gap between controlled simulations and real-world applications. It also provides a foundation for future work to optimize the framework for increasingly complex and varied datasets.

5.1 DATASET OVERVIEW

This thesis leverages datasets from publicly available repositories, notably the Electron Microscopy Public Image Archive (EMPIAR) [175, 176] and the CryoET Data Portal [177]. These platforms serve as open-access hubs for cryo-ET data, hosting a diverse array of datasets, some of which include partially annotated ground truths. Such resources are indispensable for advancing computational methods in segmentation, denoising, and particle identification, particularly when transitioning from simulated to real-world data.

EMPIAR, a widely recognized repository in structural biology, has evolved to host over a thousand entries comprising more than 2 petabytes of raw cryo-EM and cryo-ET data [176]. It facilitates reproducibility and benchmarking by providing tools for dataset visualization, search, and download. Complementing this is the CryoET Data Portal, an initiative by the Chan Zuckerberg Initiative, which aggregates cryo-ET datasets contributed by researchers globally [177]. Beyond raw tomograms, this portal offers annotations, metadata, and pre-computed results from established algorithms like CryoCARE [45] and MemBrainSeg [59], providing vital references for evaluating new methods.

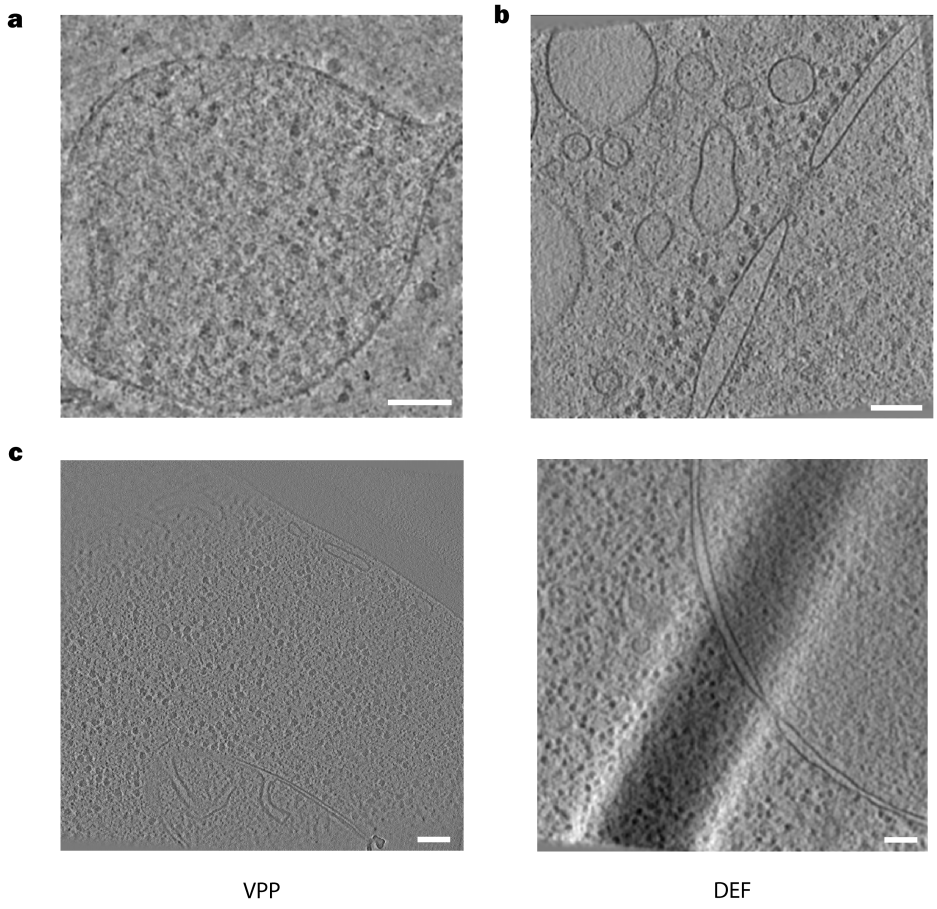

Figure 19: **Representative tomograms from publicly available real cryo-ET datasets.** (a) EMPIAR-10499: Cellular architecture of Mycoplasma pneumoniae treated with chloramphenicol. (b) EMPIAR-11756: Native ultrastructure of Chlamydomonas reinhardtii cells imaged with a Volta phase plate. (c) EMPIAR-10988: Schizosaccharomyces pombe cells imaged with a Volta phase plate (VPP, left) and defocus (DEF, right) conditions. Scale bars represent 100 nm.

The EMPIAR-10499 dataset offers a detailed resource for studying the cellular structures of Mycoplasma pneumoniae and its interactions with the antibiotic chloramphenicol [178]. This dataset includes tilt-series data captured using a Titan Krios microscope with a K2 Summit direct electron detector at 300 kV. Imaging was performed with dose-symmetric tilting, spanning angles from -60°to +60°in 3°increments. The pixel size is 1.7005 Å. Gold fiducials were embedded in the samples to aid alignment during tilt-series processing.

Reconstructed tomograms from this dataset, accessible via CryoET Data Portal ID 10003, highlight key structural features, including membranes, ribosomes, and other cytoplasmic components. Ground truth annotations of 70S ribosomes bound to chloramphenicol were created using a combination of template matching and manual curation, followed by subtomogram averaging to refine structural details. While these annotations provide a critical benchmark for particle identification and segmentation workflows, they remain incomplete, as not all ribosomes may have been identified. This limitation un-

derscores the importance of complementary datasets for achieving a broader perspective on cryo-ET analysis.

Figure 19a illustrates a representative tomogram and a selected slice highlighting the structural features of Mycoplasma pneumoniae cells. A Gaussian filter was applied to the tomogram slice to enhance visibility, addressing the inherently low contrast and high SNR produced by the imaging conditions of this dataset.

The EMPIAR-11756 dataset provides in situ cryo-ET data of Chlamydomonas reinhardtii, a single-celled organism extensively studied for its biological and structural characteristics [179]. Captured using a Titan Krios microscope equipped with a Falcon IV direct electron detector and operated at 300 kV, the imaging utilized a Volta phase plate (VPP) to enhance contrast and improve the visualization of structural features at low defocus. With a pixel size of 1.96 Å, the dataset offers high-resolution insights into the native ultrastructure of Chlamydomonas reinhardtii cells, highlighting their intricate cellular organization.

Annotations of various cellular components and macromolecular complexes are provided via the CryoET Data Portal under dataset ID 10301. These ground truths include ribosomes, rubisco complexes, microtubules, nucleosomes, and F1-Fo ATP synthase complexes. Additionally, denoised tomograms processed with CryoCARE [45] and membrane segmentations performed with MemBrainSeg [59] are available, supporting diverse computational analyses. While these annotations represent significant strides in structural characterization, other cellular features remain unannotated, leaving gaps in the dataset's comprehensiveness. This limitation reflects the dataset's targeted annotation strategy, emphasizing specific, well-defined components while excluding less distinct structures.

These resources provide a foundation for benchmarking novel computational methods, including segmentation and particle identification workflows. Figure 19b illustrates the dataset with a representative tomogram and a selected slice, showcasing the structural features of Chlamydomonas reinhardtii cells in their native state.

The EMPIAR-10988 dataset provides cryo-electron tomography (cryo-ET) data of Schizosaccharomyces pombe (fission yeast) cells prepared via cryo-focused ion beam (cryo-FIB) milling [58]. The tilt series were acquired using a Titan Krios microscope at 300 kV, equipped with a K2 direct detector camera and a Volta phase plate (VPP) to enhance contrast. Data were collected with 2° increments over a ±50°tilt range, with each tilt recorded at a pixel size of 3.37 Å.

Annotated ground truths in the EMPIAR-10988 dataset include a range of cellular and macromolecular structures, such as membranes, ribosomes, and fatty acid synthase (FAS). These annotations were generated through a multi-step process involving template matching and manual curation, followed by iterative predictions with the DL-based DeepPiCt method [58]. Subsequent rounds of manual curation further refined the results, which were then enhanced through subtomogram averaging to improve structural clarity. While these annotations provide a valuable benchmark for segmentation and particle identification workflows, they remain incomplete, focusing primarily on spe-

Dataset	Condition	Organism	Membrane GT	Particles
EMPIAR-10988	VPP	S. pombe	yes	ribosome FAS
EMPIAR-10988	DEF	S. pombe	yes	ribosome FAS
EMPIAR-10499	DEF	M. pneumoniae	no	ribosome
EMPIAR-11756	VPP	C. reinhardtii	no	ribosome nucleosome RubisCO F1-Fo

Table 8: **Summary of datasets utilized for real data evaluation.** The table includes the dataset IDs, imaging conditions (VPP: Volta Phase Plate; DEF: Defocus), the organism imaged, the availability of membrane ground truth annotations (Membrane GT), and the types of particles annotated in the tomograms. These datasets represent a diverse range of biological systems and imaging setups, providing a robust foundation for benchmarking CryoSiam's performance.

cific, readily identifiable structures such as ribosomes and FAS. Other macromolecular complexes and cellular structures in the tomograms are not annotated or determined, limiting the dataset's utility for comprehensive evaluations. This gap underscores the need for broader annotation efforts to enable more exhaustive benchmarking of computational methods in cryo-ET.

Figure 19c presents a representative tomogram and a corresponding slice from this dataset, emphasizing the diversity of cellular structures captured in the imaging. A Gaussian filter was applied to the visualized slice to improve contrast and highlight structural features, addressing the inherent challenges of low contrast in cryo-ET data.

By combining datasets from EMPIAR and the CryoET Data Portal, this work aligns with community-driven initiatives promoting open science. These datasets, summarized in Table 8, illustrate the diversity of imaging conditions, organisms, and annotations available for real cryo-ET data. For example, the EMPIAR-10988 dataset provides ground truth annotations for membrane segmentation and includes particles such as ribosomes and fatty acid synthase (FAS). In contrast, the EMPIAR-10499 and EMPIAR-11756 datasets lack membrane annotations but feature a variety of particles, including nucleosomes, RubisCO complexes, and F1-Fo ATP synthase. The organisms imaged range from Schizosaccharomyces pombe to Chlamydomonas reinhardtii and Mycoplasma pneumoniae, further reflecting the dataset variability.

This heterogeneity highlights the challenges of analyzing real cryo-ET data, such as inconsistent annotations and differences in biological samples and imaging conditions. The methodologies developed in this thesis aim to address these challenges by employing self-supervised learning approaches that do not rely on ground truth annotations for training, ensuring adaptability to diverse real-world datasets.

5.2 METHODOLOGY FOR REAL DATA EVALUATION

As described in earlier chapters, CryoSiam's evaluation of real cryo-ET datasets leverages the framework's segmentation, denoising, and particle identification modules. These modules were applied directly to real tomograms without additional fine-tuning, providing insight into the generalizability of models trained exclusively on simulated data. This section details the preprocessing steps, prediction generation process, and evaluation criteria used to assess CryoSiam's performance.

Minimal preprocessing was performed to prepare real tomograms for CryoSiam's pipeline. After tomograms were reconstructed using WBP, intensity normalization was applied to handle extreme pixel values. The normalization process involved clipping the intensity values at the 0.1 and 99.9 percentiles, effectively removing outlier pixels with abnormally high or low intensities. This approach ensured consistency with the preprocessing steps applied to simulated data, maintaining comparability between simulated and real tomogram inputs.

Denoised tomograms generated by the CryoSiam framework were utilized as inputs for downstream tasks. These tomograms were combined with lamella predictions to focus the analysis on regions of interest within the tomogram, avoiding artifacts and noise from extraneous areas. Prior experiments on simulated data informed the decision to use denoised tomograms, where denoising consistently enhanced segmentation and particle identification results.

CCryoSiam predictions were generated using the trained models directly applied to the real datasets. No fine-tuning was conducted to adapt the models to the real data, allowing an evaluation of the framework's robustness. Semantic segmentation predictions were made for structural classes, such as membranes and particles, while instance segmentation tasks used CryoSiam's distance and boundary predictions combined with watershed-based partitioning. For particle identification, subtomogram embeddings generated by SimSiam were clustered using a two-stage clustering technique. The first stage involves applying K-means to separate the particles broadly, and spectral clustering was used on individual K-means clusters for further separation.

The evaluation of CryoSiam on real data was inherently challenging due to the lack of comprehensive ground truth annotations for the datasets. Instead of relying on standard evaluation metrics, the analysis focused on qualitative comparisons and the consistency of predictions. In cases where partial ground truth or established methods were available, comparisons were performed to assess the alignment of CryoSiam's predictions with expected structural features.

Visual inspection of predictions across datasets for segmentation tasks highlighted CryoSiam's ability to adapt to different imaging conditions and biological structures. Predicted particle identities from the clusters were visualized and compared with similar structures from the simulated data with known identities to assess the approach's reliability. This qualitative and comparative framework provided valuable insights into CryoSiam's strengths and limita-

tions in handling real cryo-ET data, offering a preliminary validation of its applicability to experimental datasets.

5.3 RESULTS AND COMPARISONS

The results presented in this section showcase CryoSiam's application to real cryo-ET datasets, highlighting its performance across denoising, segmentation, and particle identification tasks. By leveraging models trained exclusively on simulated data, CryoSiam demonstrates its capability to generalize to real-world tomograms, overcoming challenges such as noise, structural variability, and incomplete ground truth annotations.

Key comparisons between CryoSiam's predictions and available references include manually curated ground truths and predictions from established methods like MemBrainSeg and CryoCARE. These comparisons emphasize the framework's ability to perform competitively despite the absence of fine-tuning on real datasets. Qualitative and quantitative analyses are presented where possible, providing insights into CryoSiam's performance across diverse imaging conditions and biological contexts.

This section underscores CryoSiam's potential to bridge the gap between simulated and real data. It offers a robust tool for cryo-ET analysis that can adapt to the complexities of experimental tomograms.

5.3.1 *Tomogram denoising*

Tomogram denoising is essential in cryo-ET to enhance visualization, facilitate manual inspection, and improve downstream analyses. However, the absence of ground truth and noise-free tomograms poses a significant challenge for developing and evaluating denoising methods. Consequently, qualitative assessments, primarily through visual inspection, are often employed to determine denoising performance.

Traditional approaches have utilized noise-to-noise frameworks, notably Noise2Noise [149], which train models using only pairs of noisy images. This concept has been adapted in cryo-ET by reconstructing tomograms from even and odd movie frames separately, creating two images of the same structure with uncorrelated noise. CryoCARE [45] exemplifies this adaptation, leveraging these paired noisy inputs to train a denoising model without requiring clean targets.

The CryoSiam framework's denoising capabilities were evaluated on real cryo-ET datasets, with comparisons against Gaussian filtering and CryoCARE. Given the lack of ground truth data for real tomograms, the evaluation focuses on visual inspection to assess noise suppression and contrast enhancement improvements across various datasets. The clean tomograms are known for the simulated data used during training, providing an apparent reference unaffected by noise or CTF modulation. The CryoSiam denoising model was trained on these simulated tomograms to learn the mapping from noisy to clean data. This section explores how well the model, trained entirely on simulated data, performs when directly applied to real cryo-ET datasets, assessing

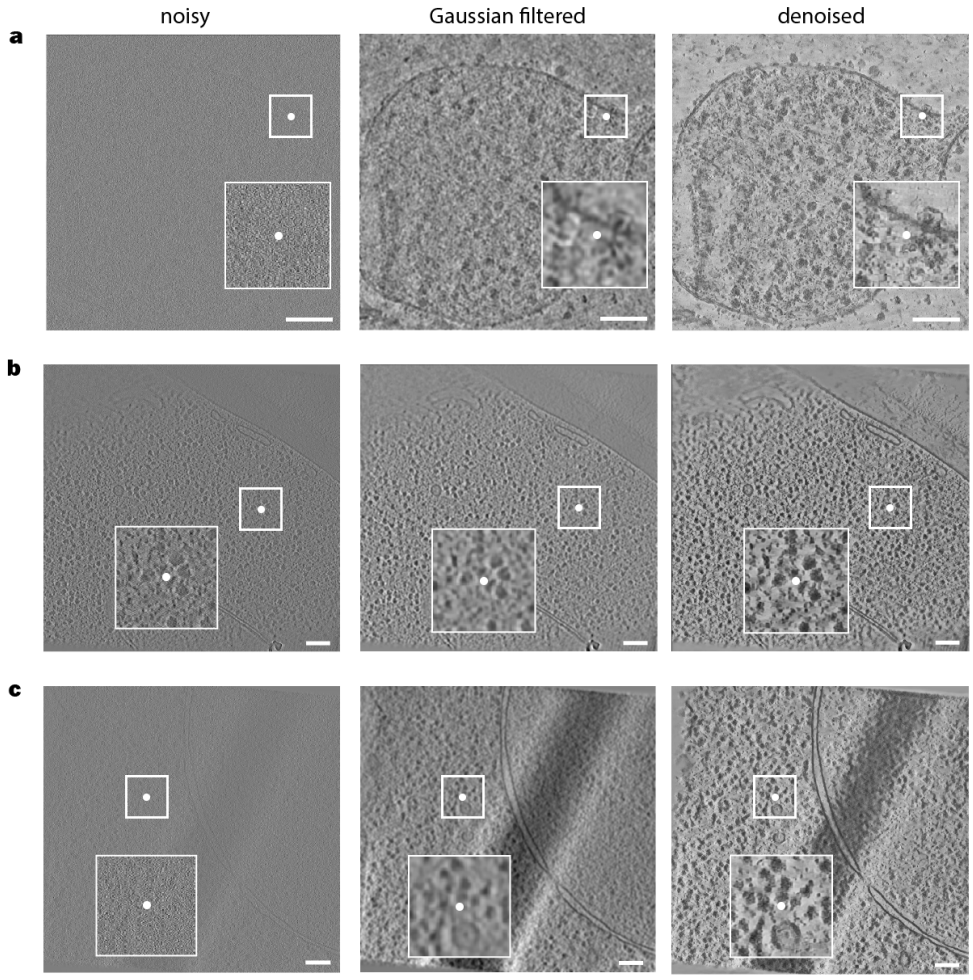

Figure 20: **Denoising results for cryo-ET tomograms using CryoSiam.** The first column shows the original noisy tomograms, the second column presents Gaussian-filtered tomograms for noise smoothing, and the third column displays the denoised outputs from the CryoSiam pipeline. Rows correspond to different datasets: (a) EMPIAR-10499, (b) EMPIAR-10988 with Volta phase plate (VPP), and (c) EMPIAR-10988 with defocus (DEF). CryoSiam denoising significantly enhances structural clarity and contrast compared to Gaussian filtering, as highlighted in the magnified insets. Scale bars represent 100 nm.

its ability to generalize and provide reliable denoising across diverse imaging conditions.

Figure 20 presents denoising results for EMPIAR-10499 and EMPIAR-10988 (VPP and DEF). The first column displays noisy tomograms with low contrast and high noise levels, characteristic of cryo-ET data. Gaussian-filtered tomograms in the second column show reduced noise and improved contrast but at the expense of more minor structural details, which are often blurred or lost. CryoSiam-denoised tomograms in the third column exhibit substantial noise suppression while preserving fine structural features. Magnified insets demonstrate CryoSiam's ability to maintain details of smaller structural components obscured in noisy or Gaussian-filtered tomograms.

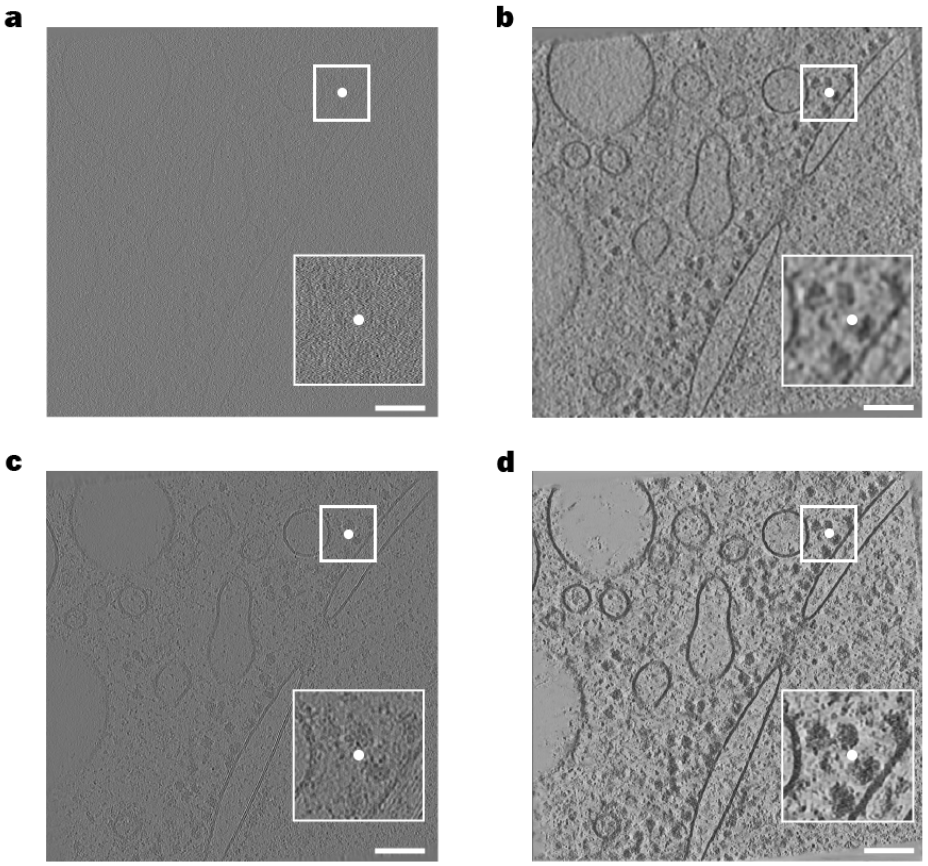

Figure 21: **Comparative denoising results for EMPIAR-11756 tomograms.** (a) Noisy tomogram, (b) Gaussian-filtered tomogram, (c) CryoCARE-denoised tomogram, and (d) CryoSiam-denoised tomogram. Magnified insets illustrate detailed improvements in structural clarity and contrast, particularly in areas affected by CTF modulation due to imaging defocus. While CryoCARE reduces noise effectively, CryoSiam further enhances contrast and restores details, addressing CTF artifacts. Scale bars represent 100 nm.

Significantly, these datasets differ in voxel resolution. EMPIAR-10499 has a pixel size of 6.8 Å, similar to the simulated data for training CryoSiam. In contrast, EMPIAR-10988 (VPP and DEF) has a voxel size of 13.48 Å, reflecting a higher binning level than the simulated training data. Despite this variation, CryoSiam effectively denoises both datasets, highlighting its robustness to differences in voxel resolution. Furthermore, the EMPIAR-10988 dataset includes tomograms acquired with a Volta Phase Plate (VPP), a condition not simulated during training. Despite this challenge, CryoSiam demonstrates robust performance, effectively denoising and preserving structural details under VPP and DEF imaging conditions. Additionally, the samples in the real datasets can be significantly thicker than those in the simulated training data, further emphasizing the method's generalizability and adaptability to diverse real-world tomographic datasets.

For EMPIAR-11756, Figure 21 presents a comparative analysis of denoising methods, showcasing Gaussian-filtered tomograms (b), CryoCARE-denoised tomograms (c), and CryoSiam-denoised tomograms (d). The CryoCARE-denoised tomograms, provided as part of the EMPIAR dataset, effectively suppress noise while preserving structural details. However, they retain artifacts associated with CTF modulation from defocus imaging conditions, reducing contrast in critical areas. In contrast, CryoSiam demonstrates superior performance by reducing noise and restoring contrast lost due to CTF effects, as evidenced in the magnified insets. These insets highlight CryoSiam's ability to enhance the visibility of intricate cellular structures, including ribosomes, microtubules, and membranes, with greater structural fidelity than other approaches. Gaussian filtering, on the other hand, reduces noise but significantly blurs finer structural features, compromising clarity. The results underline CryoSiam's capability to produce tomograms with improved clarity, which is crucial for downstream analysis in cryo-ET.

CryoSiam consistently demonstrates significant denoising improvements across all evaluated datasets, particularly under challenging defocus imaging conditions. While Gaussian filtering reduces rudimentary noise, it severely blurs fine structural features, losing important details. CryoCARE, though effective in suppressing noise, retains artifacts introduced by CTF modulation, diminishing contrast and obscuring structural clarity. By first employing SSL to generate robust representations and then fine-tuning for the supervised downstream task of denoising, CryoSiam effectively overcomes these limitations. This approach enables the production of tomograms with significantly reduced noise and restored contrast, preserving structural fidelity even under challenging imaging conditions.

These results position CryoSiam as a highly effective denoising approach for cryo-ET datasets, surpassing the performance of traditional Gaussian filtering and advanced methods like CryoCARE. CryoSiam provides a reliable foundation for downstream applications, including segmentation, particle identification, and structural analysis. It significantly enhances contrast and maintains fine structural details, addressing critical challenges in the cryo-ET workflow.

5.3.2 *Semantic segmentation*

Semantic segmentation within the CryoSiam framework was evaluated to classify voxels into structural categories, such as membranes, microtubules, actin filaments, and general particles, across real cryo-ET datasets. This evaluation assessed the model's ability to generalize from simulated data to real-world tomograms without requiring additional fine-tuning. The framework combines voxel embeddings learned through SSL with a downstream semantic segmentation task trained exclusively on simulated tomograms. This dual approach leverages CryoSiam's ability to learn rich feature representations and apply them to the segmentation of real data, even in challenging imaging conditions characterized by noise and low contrast.

To evaluate membrane segmentation specifically, CryoSiam was compared with MemBrainSeg [59], a supervised approach designed explicitly for membrane segmentation using real cryo-ET datasets for training. While Mem-

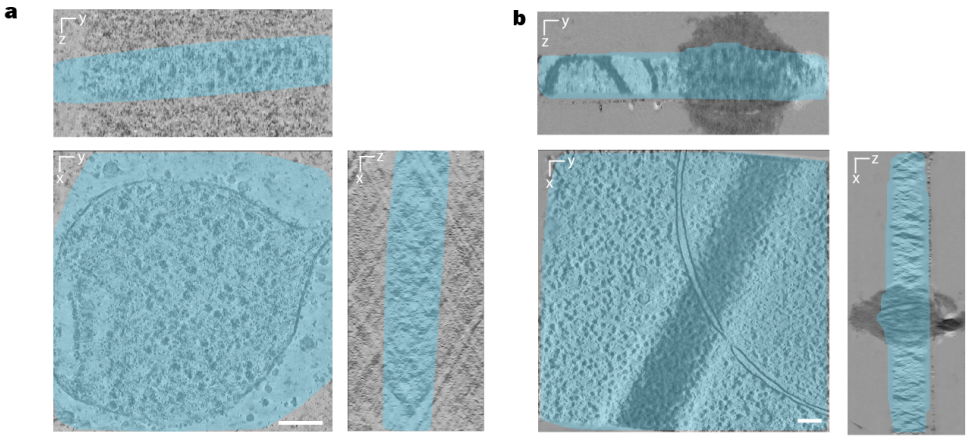

Figure 22: **Lamella prediction results on real cryo-ET data.**
(a) Predicted lamella region overlaid on a representative tomogram from EMPIAR-10499. (b) Predicted lamella region overlaid on a representative tomogram from EMPIAR-10988 (DEF). The blue shading highlights the lamella region, effectively delineating areas with meaningful structural information from surrounding noise and artifacts. These predictions enable downstream segmentation tasks to focus exclusively on biologically relevant regions, improving overall accuracy and reducing the influence of structural noise. Scale bars represent 100 nm.

BrainSeg benefits from being tailored to real data, the comparison highlights CryoSiam's capacity to achieve meaningful segmentation results without direct training on real tomograms. This evaluation explores the generalizability of CryoSiam and its ability to segment membranes and other structural components across diverse imaging conditions.

To address the challenge of tomograms including regions outside the sample, which often contain structural noise due to the approximations made during reconstruction, CryoSiam was trained on an additional downstream task for lamella prediction using simulated data. These outside regions arise because it is difficult to precisely estimate the sample's thickness during imaging, leading to tomograms encompassing more area than necessary to ensure full sample inclusion. The lamella prediction task allows CryoSiam to focus on areas containing structural information while excluding regions of noise and artifacts.

For datasets like EMPIAR-10499, which contain entire bacterial cells of Mycoplasma pneumoniae, the lamella prediction task adapts to locate and predict the region where the bacterium is situated within the tomogram. Instead of identifying lamella as in cryo-FIB milled samples, the task highlights the bacterial cell boundaries, distinguishing them from surrounding noise and empty regions. In contrast, for datasets like EMPIAR-10988 (DEF), where cryo-FIB milling produces defined lamella, the prediction task effectively identifies regions of the tomogram containing meaningful structural information, such as cellular components within the lamella.

The predicted lamella regions are illustrated in Figure 22 for representative tomograms from EMPIAR-10499 (Figure 22a) and EMPIAR-10988 (DEF, Figure 22b). The blue shading highlights the lamella regions, effectively delin-

eating meaningful structural content from surrounding noise. For EMPIAR-10499, this corresponds to the location of the bacterial cell, while for EMPIAR-10988, it identifies the lamella containing cellular structures. This step is critical for real data, as cryo-ET data contains structural noise in regions outside the sample. CryoSiam ensures that subsequent segmentation tasks are applied only to biologically relevant areas by isolating the lamella regions, improving accuracy and computational efficiency.

These results demonstrate that the lamella prediction task identifies relevant sample areas across datasets with varying imaging conditions and adapts to different structural contexts, such as whole bacterial cells or cryo-FIB milled lamella.

The performance of CryoSiam in semantic segmentation was evaluated on real datasets by combining denoised tomograms and lamella predictions as inputs to the segmentation task. The segmentation results were compared against ground truth annotations for membrane where available and predictions from MemBrainSeg [59], a supervised method for membrane prediction trained explicitly on real cryo-ET data. Figure 23 and Figure 24 illustrate these comparisons, highlighting CryoSiam's ability to generalize to diverse datasets and imaging conditions, even when trained exclusively on simulated data.

CryoSiam's membrane segmentation was critically evaluated using the EMPIAR-10988 dataset under both VPP and DEF imaging conditions, where manually curated ground truth annotations are available. This dataset benchmarked CryoSiam's performance against MemBrainSeg predictions and ground truth annotations. CryoSiam accurately segmented membrane structures, closely approximating the ground truth while detecting additional structural features, such as membrane-associated particles. This detection of particles resulted in "holes" within the membrane segmentation, a feature not observed in MemBrainSeg predictions, which primarily focus on continuous membranes. While MemBrainSeg exhibited advantages in delineating finer membrane boundaries, it occasionally over-segmented or introduced artifacts in regions without clear structural support. CryoSiam's robustness in avoiding such artifacts demonstrates its capability to generalize from simulated training data to real cryo-ET datasets.

For EMPIAR-10499 and EMPIAR-11756, where no ground truth annotations are available, the evaluation was limited to a qualitative comparison of CryoSiam and MemBrainSeg predictions. CryoSiam successfully segmented membranes in these datasets without real data supervision. In EMPIAR-10499, which contains the complete bacterial cell of Mycoplasma pneumoniae, CryoSiam captured the bacterium's smooth and continuous membrane structure, while MemBrainSeg showed more fragmented boundaries in certain areas. For EMPIAR-11756, characterized by a wider variety of cellular components and macromolecular complexes, CryoSiam effectively highlighted membrane boundaries and avoided over-segmentation, providing segmentation results comparable to MemBrainSeg.

Figure 23 showcases three-dimensional membrane and particle segmentation visualizations. Panel (a) displays results for the EMPIAR-10988 DEF dataset, including ground truth annotations (gray), MemBrainSeg predictions (yellow), and CryoSiam predictions (blue). CryoSiam demonstrates compet-

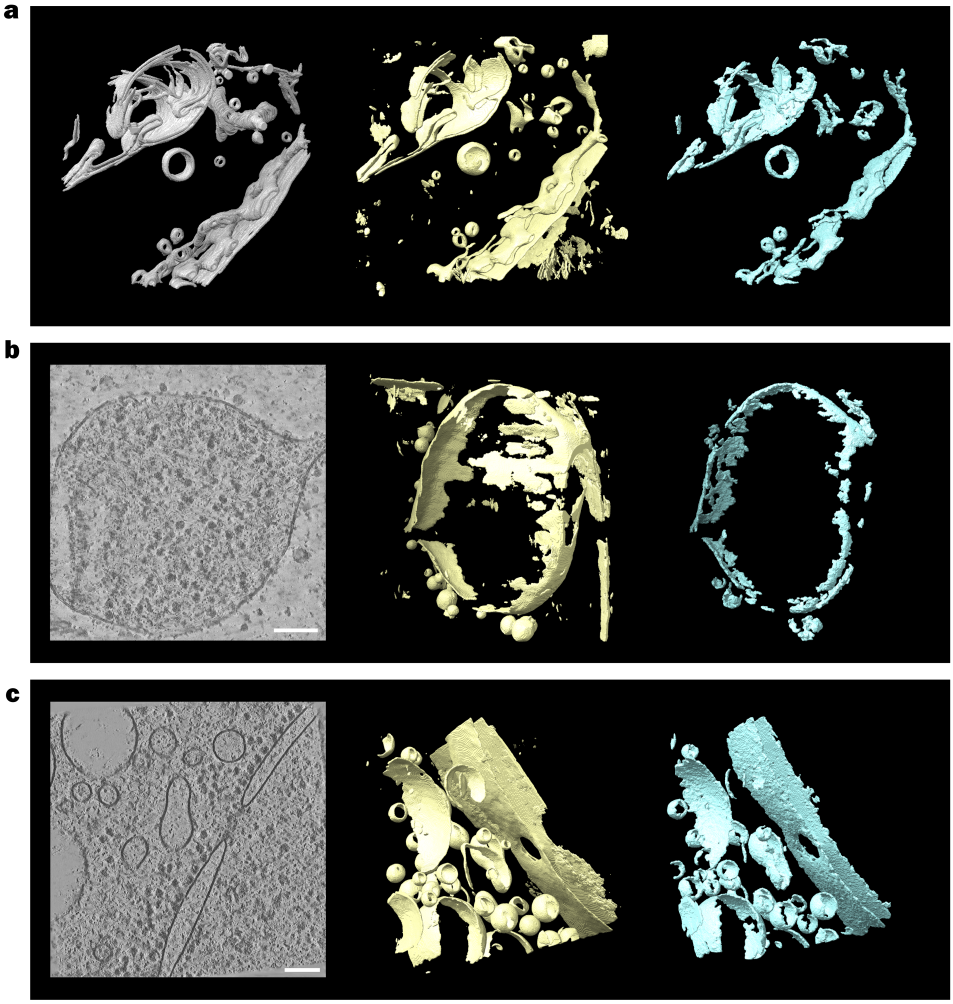

Figure 23: **Semantic segmentation results for membrane segmentation across three datasets.**
(a) Ground truth segmentation (gray), MemBrainSeg predictions (yellow), and CryoSiam predictions (blue) for a tomogram from EMPIAR-10988 DEF. (b) MemBrainSeg predictions (yellow) and CryoSiam predictions (blue) for a tomogram from EMPIAR-10499 and a 2D slice from the tomogram (first column). (c) MemBrainSeg predictions (yellow) and CryoSiam predictions (blue) for a tomogram from EMPIAR-11756 and a 2D slice from the tomogram (first column). The visualizations highlight segmentation results across different imaging conditions and biological samples, showcasing the ability of CryoSiam to capture membrane structures even without explicit training on real data. Scale bars represent 100 nm.

itive membrane segmentation accuracy, capturing complex membrane structures while highlighting associated particles. Panels b and c show segmentation results for EMPIAR-10499 and EMPIAR-11756, respectively. CryoSiam maintains high consistency in membrane and particle segmentation for these datasets despite the absence of real-data training or ground truth annotations.

Figure 24 provides an overlay of segmentation results on tomographic slices, highlighting CryoSiam's ability to segment membranes and localize particles across datasets, including EMPIAR-10988 DEF (panels a and b), EMPIAR-

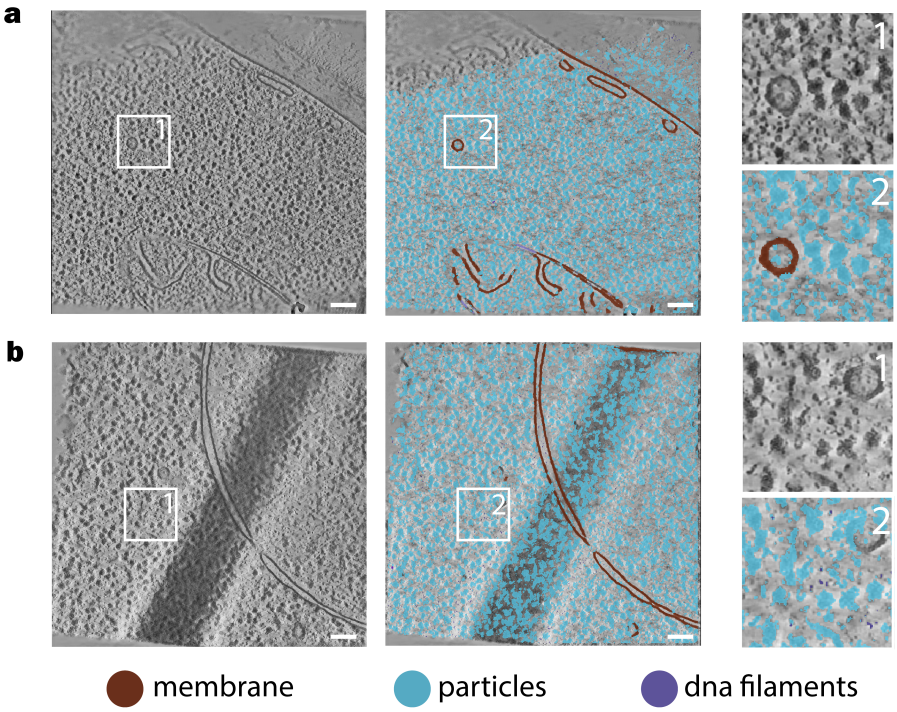

Figure 24: **Segmentation results across three datasets using CryoSiam.** (a) EMPIAR-10988 VPP with magnified areas (1: tomogram, 2: segmentation) highlighting fine structural details. (b) EMPIAR-10988 DEF segmentation showcasing robust performance despite imaging variability. (c) EMPIAR-10499 segmentation highlighting smooth and continuous membranes. (d) EMPIAR-11756 segmentation illustrating effective performance under distinct imaging conditions. Membranes are depicted in brown, particles in blue, and DNA filaments in purple. Scale bars: 100 nm.

10499 (panel c), and EMPIAR-11756 (panel d). Membranes are represented in brown, particles are highlighted in blue and predicted DNA/RNA filaments are shown in purple. The results demonstrate CryoSiam's robustness in performing membrane segmentation directly on real tomograms using models trained solely on simulated data.

CryoSiam effectively segments membrane structures while identifying membrane-associated particles, a challenging task in cryo-ET analysis. In panel a of Figure 24, CryoSiam successfully resolves fine membrane details for EMPIAR-10988 DEF, while the magnified insets illustrate its ability to detect associated particles accurately. This demonstrates the model's capacity to generalize to real tomograms with complex structures and imaging conditions. Panel b further emphasizes CryoSiam's ability to handle continuous and detailed membrane segmentation, capturing nuanced structural elements.

In panel c, the results for EMPIAR-10499 show CryoSiam's capacity to segment the entire Mycoplasma pneumoniae bacterial cell. CryoSiam captures the bacterial membrane smoothly and continuously, avoiding fragmentation and successfully identifying associated particles within the cell. These results underline CryoSiam's effectiveness in segmenting complete cellular structures, even in areas with high structural variability or overlapping features.

Panel d showcases the segmentation results for EMPIAR-11756, characterized by unique imaging conditions and various biological structures. CryoSiam achieves consistent membrane segmentation and particle localization, accurately delineating membrane boundaries even in regions with overlapping structures or noise. The magnified insets in this panel highlight CryoSiam's ability to maintain detail and clarity in challenging regions, demonstrating its generalizability to diverse datasets and imaging conditions.

An important observation arises from comparing the segmentation results between defocus imaging (panels b and c) and Volta phase plate (VPP) imaging (panel a). In the defocus data, CryoSiam predicts DNA/RNA filaments, represented in purple, that are not interacting with other particles, reflecting their distinct structural features. However, these filaments are not detected in the VPP data (panel a), likely due to the different visual appearance of such sensitive features under VPP imaging conditions, with which CryoSiam was not explicitly trained. These results suggest that contrast differences between VPP and defocus imaging influence the detection of fine structural details, pointing to potential avenues for improving CryoSiam's adaptability to diverse imaging modalities.

Overall, the segmentation results across all panels in Figure 24 underscore CryoSiam's strength in tackling complex segmentation scenarios. CryoSiam demonstrates that models trained on simulated data can be directly applied to real tomograms by accurately predicting membranes, membrane-associated particles, and DNA/RNA filaments, achieving robust and detailed results. This capability is particularly significant for cryo-ET analysis, where ground truth annotations are scarce, and segmentation tasks must contend with high noise levels and structural complexity.

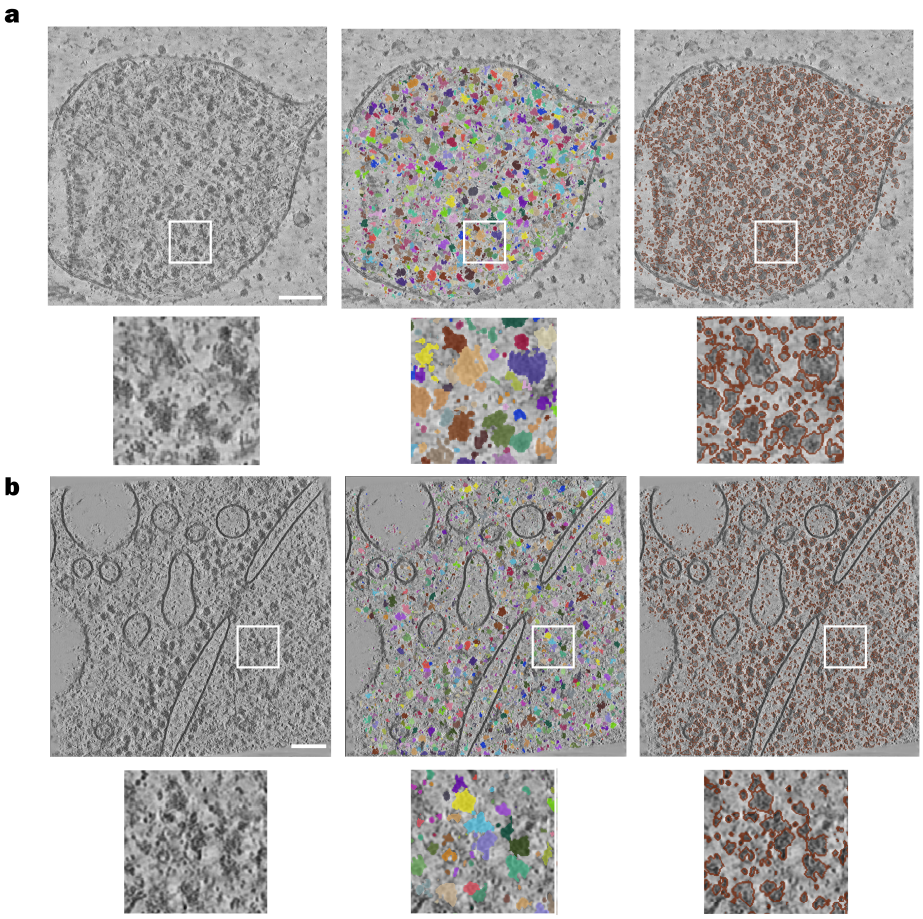

Figure 25: **Instance segmentation results for EMPIAR-10499 and EMPIAR-11756 datasets.**
(a) Instance segmentation applied to a tomogram from EMPIAR-10499. The first column shows the raw tomogram, the second column presents the segmented instances in random colors for each particle, and the third column highlights instance boundaries in brown for clearer visualization of particle separation. (b) Instance segmentation results for a tomogram from EMPIAR-11756, following the same layout as (a). Magnified insets further emphasize CryoSiam's capability to delineate individual particles within dense cellular environments, demonstrating effective separation of overlapping and closely packed structures. Scale bars represent 100 nm.

5.3.3 *Instance segmentation*

Instance segmentation in cryo-ET involves identifying and separating individual macromolecular complexes within noisy, dense cellular environments. This task is critical for downstream structural analysis, enabling the extraction of distinct particles for further processing. The CryoSiam framework includes an instance segmentation module trained on simulated data, which aims to generalize to real cryo-ET tomograms without additional fine-tuning. Applying this model to real datasets demonstrates the framework's capability to handle the complexity of biological samples under diverse imaging conditions.

Figure 25 showcases segmentation results for two real datasets: EMPIAR-10499 and EMPIAR-11756. The leftmost column in each panel shows a raw

tomographic slice from the dataset, which serves as the CryoSiam instance segmentation model input. The middle column displays the segmentation output, where each detected particle instance is assigned a unique color, demonstrating the successful differentiation of individual macromolecular complexes. The rightmost column provides a boundary visualization, highlighting particle outlines in brown to emphasize the accuracy of boundary predictions.

In EMPIAR-10499, the instance segmentation model effectively detects individual ribosomal particles within the dense cytoplasm of Mycoplasma pneumoniae cells. This dataset presents a challenging scenario due to the tomogram's high density of particles and significant structural noise. Despite these challenges, the CryoSiam model accurately identifies particle boundaries, successfully separating closely packed ribosomes. The magnified insets demonstrate the model's ability to resolve ribosomes in areas where noise partially obscures particle features.

For EMPIAR-11756, the instance segmentation task is more complex due to various macromolecular complexes, including ribosomes, Rubisco complexes, nucleosomes, and F1-Fo ATP synthase complexes. Despite size, shape, and orientation differences, the CryoSiam model differentiates these structures into distinct instances. The boundary visualization highlights the robustness of the predictions, particularly in densely populated regions where overlapping particles are common.

One key observation from these results is that the CryoSiam instance segmentation model can accurately segment particles even in highly crowded regions and under different imaging conditions. In both datasets, the model performs well at identifying and separating individual particles without significant over-segmentation or merging errors, which is critical for downstream analysis tasks.

Unlike TM methods, which require predefined particle templates, the CryoSiam framework utilizes learned voxel embeddings to distinguish particle boundaries and sizes based on structural features. This data-driven approach allows the model to generalize across diverse datasets without requiring prior knowledge of particle shapes or orientations.

CryoSiam can perform robust instance segmentation on real cryo-ET datasets, effectively separating individual particles even in challenging scenarios with noise, structural variability, and high particle density. These results validate the instance segmentation model's generalizability and potential for broad application in cryo-ET workflows.

5.3.4 *Particle identification*

The particle identification task in this thesis focused on evaluating CryoSiam's ability to identify distinct macromolecular complexes within real cryo-ET datasets. For this purpose, the EMPIAR-10499 dataset, which contains Mycoplasma pneumoniae cells treated with chloramphenicol, was selected due to the relatively simple architecture of the bacterial cell compared to more complex eukaryotic cells present in other datasets. The choice of this dataset allowed for a more controlled and interpretable evaluation of CryoSiam's

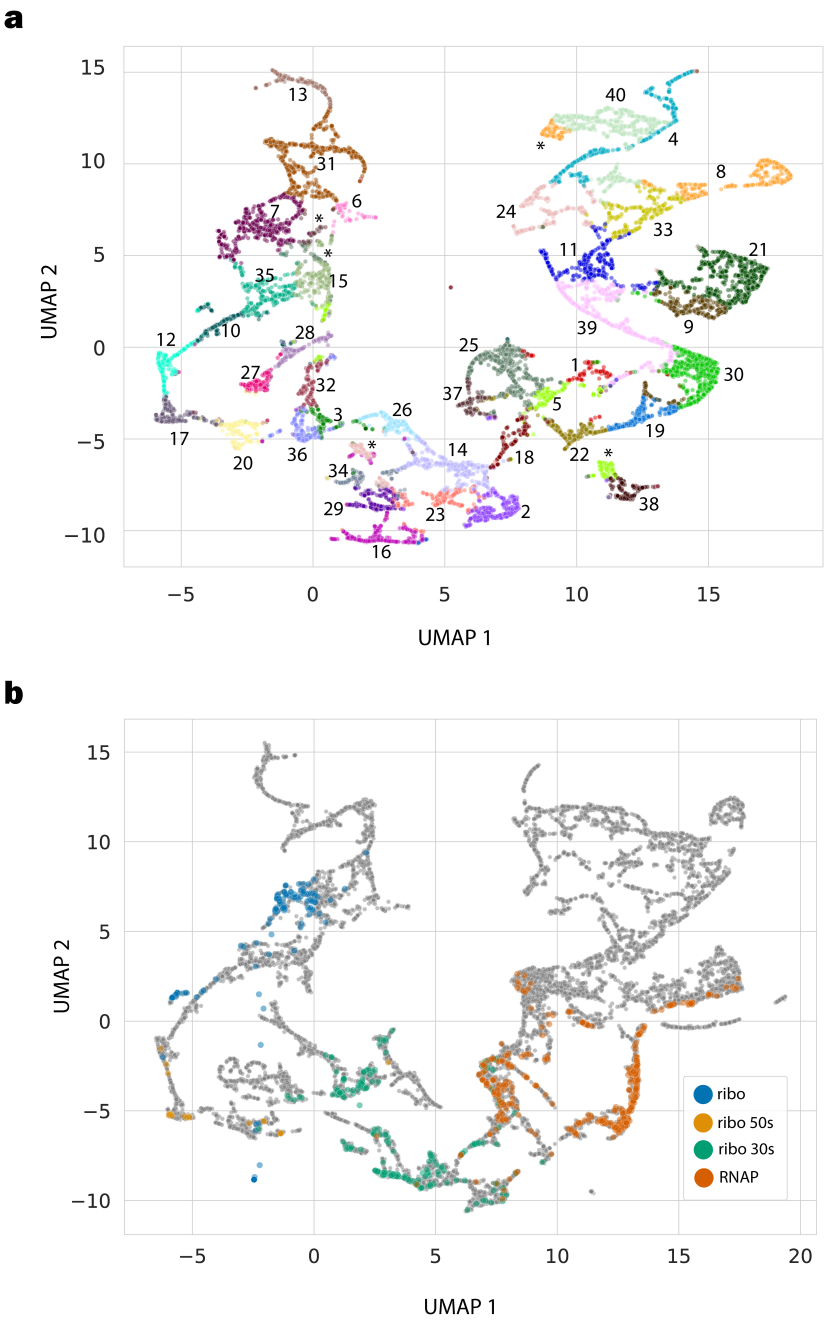

Figure 26: **UMAP projection of subtomogram embeddings and comparison with simulated ground truth.**
(a) UMAP projection of embeddings from EMPIAR-10499, with K-means clusters shown in different colors. Labels indicate cluster IDs, and repeated clusters are marked with an asterisk (*). (b) Simulated ground truth embeddings for ribosomes (blue), 50S subunits (orange), and 30S subunits (green) projected into the same space, showing alignment with real data clusters and indicating separation based on particle identities.

particle identification capabilities, reducing confounding factors such as high cellular complexity or extensive heterogeneity in particle types.

The identification process began with instance segmentation applied to denoised tomograms from EMPIAR-10499. Denoising was employed to enhance

contrast and improve the visibility of structural details, aligning with previous findings that demonstrated its positive impact on segmentation performance. After segmentation, identified instances within the tomograms were used to generate subtomograms, each containing a single particle. This step ensured that the subsequent particle identification task was focused on individual macromolecular complexes rather than overlapping or partially captured structures within the tomograms.

Once the subtomograms were extracted, they were processed through the SimSiam model to generate embeddings for each subtomogram. The use of SimSiam enabled the creation of meaningful and distinguishable embeddings for different particle types without requiring labeled training data from real tomograms. These embeddings served as the foundation for particle identification, comparing extracted features with reference particles and clustering similar particles within the embedding space.

The initial stage of particle identification focused on clustering the subtomogram embeddings derived from the instance segmentation outputs. K-means clustering was employed to provide an initial separation of the subtomograms into distinct groups. To visualize the clustering results in a 2D space, UMAP was applied, as shown in Figure 26. The UMAP projection reveals a clear separation between the clusters identified by K-means, with different colors representing distinct clusters. However, some clusters appear in multiple areas within the UMAP space, indicated by an asterisk (*), signifying overlapping clusters. Given the simplicity of K-means as a clustering algorithm, this overlap is expected. The method was primarily used to achieve an initial separation, allowing further refinement in subsequent steps. The total number of clusters chosen for K-means was 40, balancing initial separation with computational feasibility.

Following the initial clustering, the next goal was to assess whether the UMAP space exhibits separation based on known particle identities. Subtomograms from simulated ground truth data were projected into the same UMAP space. The simulated data used for this evaluation was from the transcription demonstration sample, containing ribosome particles, the large ribosomal subunit (50S), the small ribosomal subunit (30S), and RNA polymerase (RNAP). These particles have distinct structural features and sizes, making them ideal for evaluating how well the embeddings preserve meaningful biological differences.

The UMAP projection in Figure 26b shows distinct regions for different particle types. Ribosomes in blue form a well-defined cluster, demonstrating that the embeddings can differentiate ribosomes from other particles. Additionally, the large ribosomal subunit (50S), depicted in orange, maps to a different area in the UMAP space than the small ribosomal subunit (30S), shown in green. This separation by particle size aligns with expectations, as the size and complexity of the particles are key factors influencing their embeddings.

Interestingly, the RNA polymerase particles, smaller than ribosomes but closer to the 30S subunit, also map to a distinct region in the UMAP space. This separation indicates that the embeddings capture not only size differences but also structural variations that distinguish different types of particles. The distinct regions in the UMAP for ribosomes, ribosomal subunits, and RNA

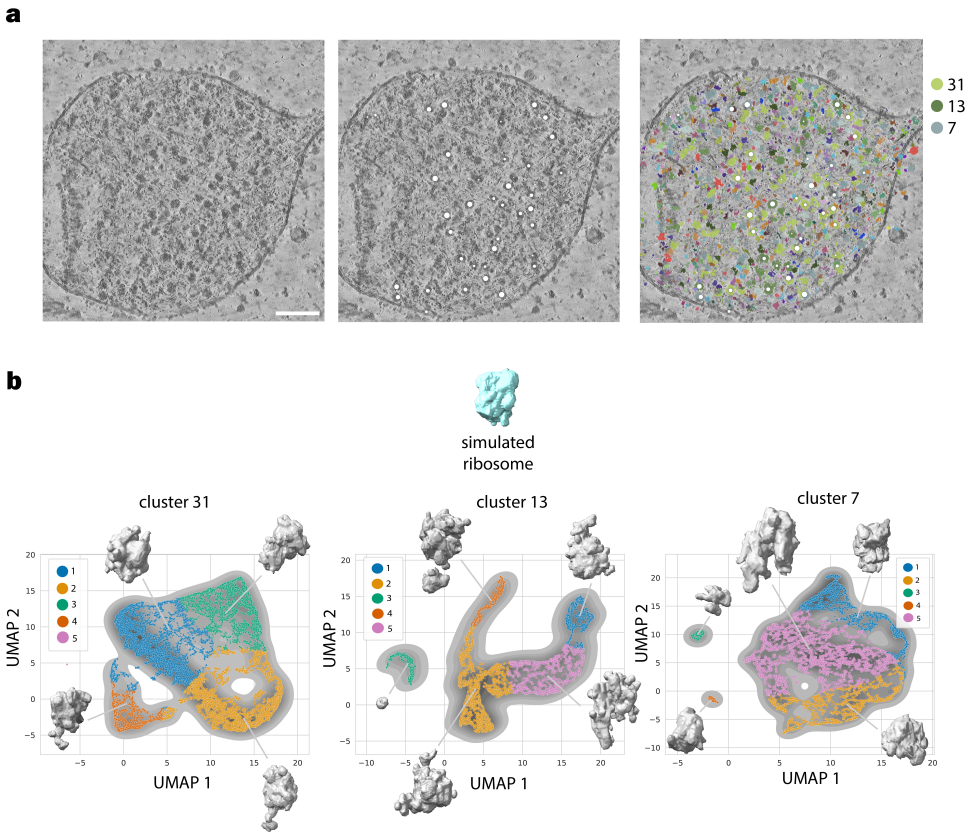

Figure 27: **Particle identification results for EMPIAR-10499 for ribosome clusters** (a) Visualization of the tomogram (left), ribosome ground truth particle centers mapped onto the tomogram (middle), and the identified clusters from the first-stage K-means clustering overlaid on the tomogram (right). The clusters most closely correspond to the ribosome locations are clusters 31, 13, and 7, indicated by different colors. (b) UMAP projections for clusters 31, 13, and 7 were further refined using spectral clustering. Subclusters reveal additional separations within each cluster. Denoised subtomograms corresponding to selected particles are shown as surface visualizations, highlighting structural differences within each subcluster. One ribosome example from the simulated data is shown in blue above the UMAP plots for reference.

polymerase suggest that the embeddings contain meaningful information for particle identification despite being trained solely on simulated data.

These observations pointed to specific K-means clusters that warranted further investigation. A second clustering stage was performed using spectral clustering to refine and address K-means' limitations. This method was expected to handle complex, non-linear boundaries in the embedding space and provide a more accurate separation of particle types.

The particle identification process revealed unexpected cluster sizes in the UMAP projections from Figure 26. Given that ribosomes are among the largest particles expected in Mycoplasma pneumoniae, larger clusters were particularly intriguing. To investigate this further, the provided ground truth annotations of ribosome centers were mapped onto the tomogram alongside the cluster identities from the initial K-means clustering (Frigure 27a). This mapping

identified clusters 31, 13, and 7 as capturing ribosome particle centers in most cases. These clusters, however, also included additional densities around the ribosomes, suggesting that ribosomes might be interacting with other molecular components within the bacterial cell.

The subsequent analysis, shown in Figure 27, explores these clusters in more detail. Panel (b) provides an in-depth investigation of clusters 31, 13, and 7, using spectral clustering to refine them into five sub-clusters each. A simulated ribosome was included to assess size and shape similarities for comparison. The spectral clustering revealed important insights into the complex structures present in these clusters.

Cluster 31 predominantly captured ribosomes in interaction with surrounding densities, indicating the presence of ribosome complexes. The sub-clusters revealed that ribosomes rarely appeared as isolated particles but were frequently bound to other cellular components, suggesting ongoing cellular processes or interactions within the cell.

Similar observations were made in cluster 13. In this second stage of clustering, sub-cluster 2 captured ribosomes near smaller particles resembling RNA polymerase. This observation aligns with the biological context of active transcription-translation coupling in Mycoplasma pneumoniae, where ribosomes and RNA polymerase form complexes during gene expression.

Cluster 7 provided additional insights. While this cluster also included ribosomes, two sub-clusters, labeled 3 and 4, captured distinct densities that did not resemble ribosomes. These sub-clusters, well-separated from other ribosome-related clusters, highlight the importance of applying spectral clustering for refined particle identification. Separating these unique sub-clusters indicates that CryoSiam can capture diverse structural elements within the tomogram, including previously unidentified or less well-characterized particles.

Combining K-means clustering for initial grouping and spectral clustering for refined separation proved effective in identifying complex molecular assemblies in real tomograms. The analysis underscores CryoSiam's ability to detect ribosomes, ribosome-associated complexes, and distinct particles, highlighting its potential for uncovering new insights in cryo-ET datasets.

Interestingly, the results from Figure 26 highlighted distinct areas in the UMAP projection that corresponded to the known particles from the simulated dataset, specifically, the ribosome 50S, ribosome 30S, and RNA polymerase. We further selected several clusters from the K-means clustering results to investigate these observations for a second-stage spectral clustering analysis. The goal was to determine whether the embeddings could reveal finer separations within these clusters, particularly for complex particles like ribosomes and RNA polymerase. Figure 28 presents the results of this secondary clustering.

Panel (a) illustrates the UMAP projections for clusters 17, 32, and 37, refined into five subclusters each through spectral clustering. Focusing first on Cluster 17, closely aligned with the ribosome 50S ground truth in the UMAP space, the subclusters reveal intriguing insights. Subclusters 1 and 3 show particles with size and shape that correspond well to ribosome 50S particles. In particular, the particles in subcluster 1 demonstrate high structural similarity

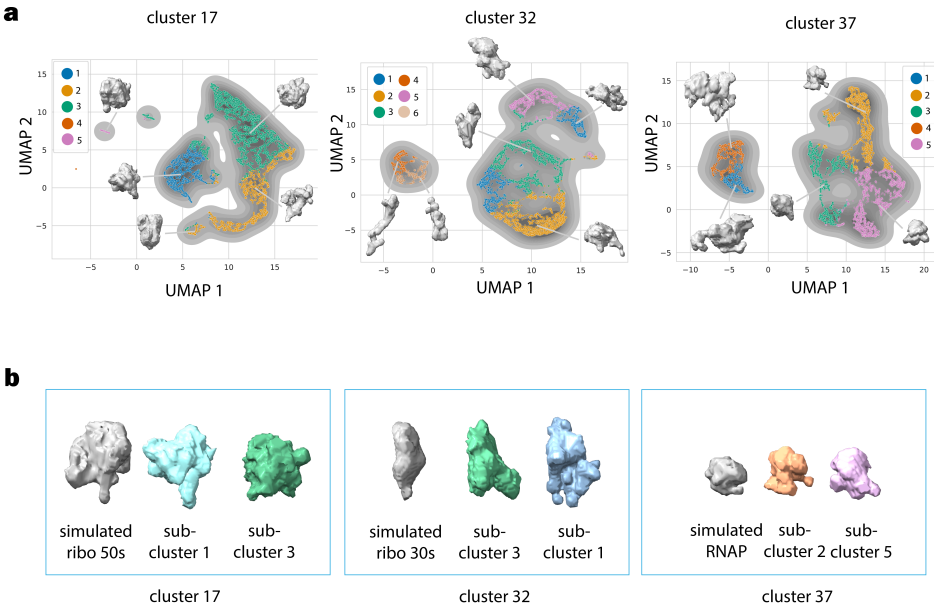

Figure 28: **Detailed analysis of selected clusters through UMAP visualization and spectral clustering.**
(a) UMAP projections of clusters 17, 32, and 37, obtained from the first stage of K-means clustering, are shown with further refinement into five (or six) sub-clusters using spectral clustering. The sub-clusters are color-coded, and representative surface visualizations of denoised subtomograms from each sub-cluster are displayed around the UMAP plots. (b) Comparison between simulated reference particles and sub-clustered densities. For cluster 17, sub-clusters 1 and 3 are shown alongside the simulated 50S ribosome. For cluster 32, sub-clusters 1 and 3 are compared with the simulated 30S ribosome. For cluster 37, sub-clusters 1 and 3 are compared to the simulated RNA polymerase (RNAP).

to the simulated ribosome 50S reference, suggesting accurate identification by the CryoSiam framework. The comparison in panel (b) supports this observation, where the simulated ribosome 50S surface is visually compared with examples from subclusters 1 and 3, highlighting their resemblance in both size and shape.

However, subcluster 2 presents particles that deviate from the expected ribosome morphology. These particles have a similar size to ribosomes but do not exhibit typical ribosomal structures. This observation suggests that these embeddings likely represent other macromolecular complexes or potential noise from the segmentation process. Additionally, subcluster 5 includes ribosome 50S particles with additional densities, which could indicate ribosomes interacting with other molecular components or segmentation artifacts caused by over-segmentation during the instance segmentation stage.

Moving on to cluster 32, this cluster was mapped close to the simulated ground truth embeddings for isolated ribosome 30S particles. Upon further investigation through spectral clustering, distinct structural variations were observed within this cluster. Subclusters 1 and 3 notably stood out as they exhibited size and shape characteristics aligned with the isolated ribosome 30S structure. These two subclusters are shown in panel (b) of Figure 28, highlight-

ing their resemblance to the simulated ribosome 30S reference. However, an important observation is that the ribosome 30S particles identified in the real tomograms were not entirely isolated. Instead, they appeared to interact with other molecular components, particularly in subcluster 3. Interestingly, subcluster 3 displayed a structural arrangement that closely matched the size and shape of RNA polymerase, suggesting a potential interaction between ribosome 30S and RNA polymerase in the tomogram. This observation aligns with the biological context of Mycoplasma pneumoniae, where active transcription-translation is standard, with ribosomes often engaged in complex interactions during the protein synthesis process.

Lastly, cluster 37 was mapped in the UMAP space close to the simulated RNA polymerase references. This cluster was exciting due to its association with a key transcription factor in bacterial cells. Initial K-means clustering did not separate the RNA polymerase particles, resulting in a mixture of different structures within cluster 37. However, the second-stage spectral clustering provided a more refined separation, revealing five subclusters. Among these, subclusters 1 and 4 emerged as the largest and were distinct from the other subclusters. These two subclusters captured particles with sizes and shapes more consistent with RNA polymerase, suggesting that the spectral clustering step was essential for isolating RNA polymerase particles more accurately.

For visualization, examples from subclusters 2 and 5 are presented in panel (b) of Figure 28. These examples show that the overall shape and size of the particles in these subclusters match the RNA polymerase reference from the simulated dataset. Interestingly, both subclusters include additional densities around the core RNA polymerase structure, which were not part of the simulated reference. These additional densities indicate active RNA polymerase complexes engaged in transcription processes, where the protein interacts with DNA, RNA, and other associated factors. This observation underscores the importance of identifying isolated particles and functional complexes, which are more representative of the native biological context.

Overall, the analysis of clusters 17, 32, and 37 highlights the importance of the two-stage clustering approach. While K-means clustering provides an initial separation of subtomograms based on general features, the more nuanced spectral clustering identifies distinct subpopulations within each cluster, capturing isolated particles and larger complexes. This multi-step approach allows CryoSiam to distinguish between functionally relevant complexes in real tomograms, demonstrating its utility in particle identification and classification in complex cryo-ET datasets.

5.4 DISCUSSION

The application of CryoSiam to real cryo-ET datasets underscores the framework's ability to generalize across a range of biological and imaging conditions. CryoSiam has demonstrated robust performance in denoising, semantic segmentation, and particle identification tasks by training exclusively on simulated data. However, the transition from simulated to real data highlighted several strengths and limitations warranting discussion.

One of the framework's key strengths lies in its SSL approach, which facilitates the extraction of meaningful representations without relying on annotated real data. This design choice enabled CryoSiam to achieve competitive results compared to supervised methods like MemBrainSeg and CryoCARE, even under challenging imaging conditions. In particular, the denoising module successfully suppressed noise while preserving structural fidelity, and the segmentation module accurately captured complex membrane geometries in datasets like EMPIAR-10988.

The particle identification task further demonstrated CryoSiam's versatility. The embeddings generated from subtomograms enabled initial K-means clustering, visualized using UMAP. The visualization showed clear separations between clusters, indicating that meaningful embeddings were created for the particles within the tomograms. Upon projecting known simulated particle embeddings onto the same UMAP space, distinctions between ribosomes, RNA polymerase, and their subunits became apparent, underscoring CryoSiam's potential to distinguish between particles based on their size and shape. However, some clusters remained mixed, which led to a second stage of spectral clustering to refine the separation.

The subsequent analysis of clusters revealed interesting findings. For instance, in cluster 17, which mapped closely to the ribosome 50S, several subclusters emerged that corresponded to ribosomes in complex with additional densities. This suggests that CryoSiam can detect particles in interaction with other structures, providing insights into the functional state of the particles in situ. Similarly, cluster 32 aligned closely with ribosome 30S subunits, but it also revealed interactions with RNA polymerase, potentially representing active transcription-translation complexes. Cluster 37, associated with RNA polymerase, was initially mixed but was refined into distinct subclusters through spectral clustering. The separation highlighted CryoSiam's ability to differentiate particles of similar size and shape, even when mixed in the initial k-means stage.

These findings demonstrate that CryoSiam is a robust tool for analyzing complex cells with dense molecular environments and interacting particles. However, some critical challenges remain. Notably, the missing wedge artifact, inherent to cryo-ET data, results in an anisotropic resolution that can hinder accurate particle identification. This artifact was not addressed in the simulated training data, and its effects were visible in the real data evaluations. Addressing this limitation would be a key future direction, possibly by incorporating missing-wedge-aware methods into CryoSiam's workflow.

Another limitation is the lack of comprehensive ground truth data for real tomograms, making it challenging to perform quantitative evaluations. For datasets like EMPIAR-10988, ground truths were available for membrane segmentation and ribosome identification, but these annotations remain incomplete and do not encompass all particles or structures within the tomograms. Qualitative comparisons were the primary evaluation method for other datasets, such as EMPIAR-10499 and EMPIAR-11756. These challenges emphasize the importance of ongoing collaboration with structural biologists to verify CryoSiam's predictions and identify particles of smaller sizes or less well-defined structures.

Despite these challenges, CryoSiam offers a significant advantage in scalability and adaptability to new experimental setups. Its reliance on simulated data removes the dependence on extensive annotated real datasets, which are often difficult to obtain. CryoSiam complements existing supervised frameworks by providing robust generalization and the potential to explore structural interactions in regions with little prior knowledge. However, additional steps should be taken to refine the particle identification pipeline to maximize its utility. For example, applying sub-tomogram averaging to identified clusters could help determine the high-resolution structures of particles of interest, providing more precise insights into their molecular architecture.

Future work should also focus on investigating CryoSiam's capability to detect smaller particles and explore the limits of its particle identification accuracy. This would require expertise from structural biologists to interpret the clusters and validate the biological relevance of the predicted structures. The potential to uncover previously uncharacterized particles or complexes within tomograms is a promising direction for future research, leveraging CryoSiam's self-supervised embeddings to unlock novel insights into the molecular organization of cells.

CryoSiam demonstrates its robustness across various real cryo-ET datasets, effectively performing denoising, segmentation, and particle identification. The framework shows particular strength in segmenting large, complex particles and identifying molecule interactions. However, challenges such as the missing wedge artifact and the need for more thorough validation with real ground truth data remain. Addressing these challenges through methodological improvements and collaborative efforts will be essential further to enhance CryoSiam's role in cryo-ET analysis workflows.

5.5 CONCLUSION AND FUTURE DIRECTIONS

The evaluation of CryoSiam on real cryo-ET datasets highlights its potential as a robust and versatile framework for denoising, segmentation, and particle identification. By relying solely on models trained on simulated data, CryoSiam demonstrates that meaningful insights can be derived from real tomograms without the need for extensively annotated datasets. The results showcase CryoSiam's capacity to achieve competitive performance in key tasks such as membrane segmentation and denoising under various imaging conditions, emphasizing its generalizability and scalability across diverse biological samples.

CryoSiam's performance in membrane segmentation and particle identification particularly underscores its capability to handle complex cellular environments with high densities and interacting particles. The framework successfully identified major macromolecular complexes within Mycoplasma pneumoniae cells, such as ribosomes and RNA polymerase, while distinguishing between their subunits and interaction states. Notably, CryoSiam achieved this level of particle identification without prior exposure to real data, validating the effectiveness of its self-supervised embedding space for clustering and classification tasks. Furthermore, combining K-means and spectral clustering, the two-stage clustering approach allowed for the refinement of clusters and

the detection of distinct molecular complexes, providing a pathway for more detailed structural analysis.

However, the transition from simulated to real data also revealed key limitations that must be addressed in future work. One significant challenge is the missing wedge artifact, which introduces anisotropic resolution in cryo-ET reconstructions. This artifact affects the appearance of particles in tomograms, complicating their identification and classification. CryoSiam, trained exclusively on isotropic simulated data, does not currently account for this artifact, which may impact the accuracy of its predictions. Addressing this limitation by incorporating missing-wedge-aware training or domain adaptation techniques could substantially improve the framework's performance on real tomograms.

Another area for improvement is the diversity of the simulated training data. The current simulated datasets used for training do not fully capture the structural variability observed in real biological samples, particularly for smaller and more flexible particles such as actin filaments or diverse microtubules. To improve generalization, future iterations of CryoSiam should incorporate more diverse simulated datasets that represent a wider range of structural complexities.

The results presented in this study suggest that CryoSiam can be further refined to incorporate additional particle characterization capabilities. One promising direction is to complement the particle identification workflow with STA techniques to resolve higher-resolution structures of the detected particles. This would enable more precise identification and characterization of macromolecular complexes within tomograms, facilitating deeper biological insights. Additionally, integrating CryoSiam with supervised fine-tuning on real data, where available, could enhance its accuracy and reliability, especially for challenging datasets with unique imaging artifacts or rare structural features.

CryoSiam has demonstrated significant potential to streamline cryo-ET workflows by reducing reliance on annotated real datasets and providing robust generalization across various imaging conditions. The ability to predict both large-scale membrane structures and smaller, membrane-associated particles, such as ribosomes and RNA polymerase, highlights its versatility in addressing the unique challenges of cryo-ET analysis. By identifying macromolecular complexes in dense cellular environments, CryoSiam opens the door to automated analysis pipelines that can handle the complexities of in situ tomograms, accelerating discoveries in structural biology.

In conclusion, this thesis has demonstrated the feasibility and effectiveness of CryoSiam in real cryo-ET data analysis. The results validate the core hypothesis that models trained on simulated data can generalize well to real-world scenarios, even without direct supervision when SSL pretasks are incorporated into the framework. CryoSiam's success in denoising, segmentation, and particle identification lays a strong foundation for future advancements in cryo-ET workflows, where the goal is to make structural biology more accessible, efficient, and automated. Further refinements in handling imaging artifacts, expanding the diversity of training data, and exploring hybrid training approaches will be key to maximizing the impact of CryoSiam in cryo-ET and beyond.

Part IV

RESEARCH SYNTHESIS AND CONCLUSIONS

This final part reflects on the research outcomes, synthesizing findings from both simulated and real cryo-ET data experiments. It discusses the implications of the proposed methodology for segmentation in cryo-ET and the broader fields of structural biology and machine learning. The discussion critically evaluates the strengths, limitations, and challenges encountered during the research process. The part concludes by summarizing the contributions of this work and exploring potential directions for future research, highlighting opportunities for further advancing cryo-ET data analysis and self-supervised learning techniques.

6 DISCUSSIONS

This chapter critically evaluates the contributions of the CryoSiam framework, emphasizing its design, implementation, and performance across simulated and real cryo-ET datasets. The discussion highlights the framework's strengths in addressing key challenges in cryo-ET data analysis, including high noise levels, imaging conditions variability, and annotated ground truth data scarcity. Using SSL and leveraging simulated datasets, CryoSiam demonstrated its capacity to handle diverse downstream tasks such as denoising, semantic segmentation, instance segmentation, and particle identification.

The chapter contextualizes these achievements by analyzing results from controlled simulated and real cryo-ET datasets and comparing them with existing state-of-the-art methods. It explores insights from ablation studies, qualitative evaluations, and task-specific benchmarks to assess CryoSiam's ability to generalize effectively across datasets with different imaging conditions. This evaluation underscores the framework's practical contributions to cryo-ET workflows and its broader implications for data-driven approaches in structural biology.

Finally, this chapter identifies the limitations of the current framework and discusses the lessons learned from its application to real data. These reflections provide the foundation for proposing future research directions, which will be elaborated in the subsequent chapter.

6.1 OVERVIEW OF CONTRIBUTIONS

The CryoSiam framework represents a significant advancement in the analysis of cryo-ET data, addressing key challenges in denoising, segmentation, and particle identification. By leveraging SSL, this thesis introduced a novel framework capable of generalizing from simulated data to real tomograms, reducing the reliance on annotated real-world datasets. This contribution is particularly relevant in cryo-ET, where obtaining high-quality annotations is time-consuming and prone to human error.

One of the core contributions of this thesis is the development of a two-stage embedding framework that operates at both voxel and subtomogram levels. DenseSimSiam, the voxel-level embedding generator, captures fine structural details within tomograms, enabling dense semantic segmentation for complex cellular environments. SimSiam, on the other hand, provides subtomogram-level embeddings tailored for particle identification tasks. Integrating these components within a unified pipeline enables seamless processing of cryo-ET data, covering the entire workflow from denoising and membrane segmentation to the identification and classification of macromolecular complexes. This holistic approach addresses a critical gap in existing cryo-ET analysis methods by providing a fully automated and transferable solution.

Another key achievement is the framework's ability to perform instance segmentation and particle identification on real tomograms without requiring any fine-tuning on real data. This work demonstrated that CryoSiam could effectively process tomograms from different organisms and imaging conditions, as shown by its application to datasets such as EMPIAR-10499, EMPIAR-10988, and EMPIAR-11756. In particular, CryoSiam successfully segmented membrane structures and identified particle classes like ribosomes, RNA polymerase, and other macromolecular complexes, highlighting its capacity to handle dense, complex cellular environments. Notably, the framework detected isolated particles and particles in active interactions, such as transcription-translation complexes, which are notoriously challenging to identify.

The framework's reliance on SSL allowed it to overcome a significant limitation faced by traditional supervised methods: the need for extensive labeled data. CryoSiam leveraged simulated datasets to create transferable embeddings that generalized well to real-world data. This contribution demonstrates the practical feasibility of using simulated training data to bridge the gap between synthetic and real cryo-ET data, a key challenge in structural biology. Furthermore, CryoSiam's ability to achieve competitive performance compared to state-of-the-art supervised methods, such as MemBrainSeg and CryoCARE, underscores the robustness of its self-supervised approach. For instance, the framework delivered comparable or superior membrane segmentation results without supervised fine-tuning on real data.

The thesis also presented a comprehensive set of ablation studies, providing valuable insights into the impact of various design choices on the model's performance. These studies investigated critical factors such as embedding sizes, loss functions, and data augmentations, revealing how each component contributes to the overall effectiveness of the framework. This empirical analysis was essential for optimizing the CryoSiam framework, ensuring it could generalize well across different datasets and imaging conditions.

Beyond segmentation tasks, CryoSiam demonstrated instance segmentation and particle identification capabilities by clustering embeddings and visualizing their separability using UMAP. The framework achieved two-stage clustering through k-means followed by spectral clustering, enabling the detection of distinct molecular complexes. This approach revealed the framework's potential to identify large macromolecular complexes, such as ribosomes and RNA polymerase, and finer substructures, such as ribosomal subunits. Importantly, CryoSiam detected membrane-associated particles, a challenging task due to the variability in particle interactions and density. These findings highlight the framework's utility in identifying molecular interactions and structural arrangements within cells, paving the way for deeper biological insights.

Despite its achievements, the thesis also identified key limitations that inform future directions. One notable limitation is CryoSiam's sensitivity to the missing wedge artifact, which causes anisotropic resolution in cryo-ET reconstructions. The missing wedge remains a critical challenge in cryo-ET analysis, particularly for particle identification, and addressing this artifact would enhance the framework's ability to resolve smaller particles and finer structural details. Another limitation is the need for more diverse simulated training data

to better capture the variability in real tomograms, particularly for smaller and more flexible particles, such as DNA/RNA filaments or protein complexes in different conformational states.

Finally, this thesis contributes to advancing the practical implementation of cryo-ET workflows by presenting a comprehensive evaluation across simulated and real datasets, integrating innovative SSL techniques, rigorous evaluation, and meaningful comparisons with established methods, which positions CryoSiam as a valuable tool for cryo-ET researchers. By demonstrating that models trained solely on simulated data can achieve competitive performance on real tomograms, this work opens new possibilities for automated analysis pipelines in structural biology, particularly in cases where annotated data are scarce or unavailable.

In summary, this thesis introduced a novel self-supervised cryo-ET analysis framework capable of performing denoising, semantic segmentation, instance segmentation, and particle identification tasks with high accuracy and adaptability. CryoSiam addresses critical challenges in the field by reducing the reliance on annotated data and demonstrating robust performance across various biological samples and imaging conditions. The insights gained from this work lay the foundation for further advancements in cryo-ET workflows, pushing the boundaries of automated tomographic analysis and structural discovery.

6.2 EVALUATION OF PERFORMANCE

The evaluation of CryoSiam on both simulated and real cryo-ET datasets underscores the framework's versatility across a wide range of tasks, including denoising, semantic segmentation, instance segmentation, and particle identification. This performance assessment highlights CryoSiam's ability to generalize from simulated data to real-world conditions, demonstrating its potential as a comprehensive tool for structural biology research. The evaluation process, however, also brought to light key strengths, limitations, and areas for future refinement.

One of the most significant outcomes of this evaluation is CryoSiam's effective handling of noisy tomograms, particularly under challenging imaging conditions. The denoising module produced tomograms with reduced noise and restored contrast, improving the visibility of structural features critical for downstream analysis. Comparisons with supervised methods such as CryoCARE revealed that CryoSiam could achieve comparable noise suppression without requiring annotated real data. CryoSiam excelled in handling defocus imaging conditions, where contrast distortions introduced by the contrast transfer function (CTF) are prevalent. CryoSiam demonstrated its robustness and adaptability to diverse imaging setups by restoring structural clarity in tomograms affected by CTF modulation.

In the semantic segmentation task, CryoSiam achieved accurate membrane segmentation, identifying continuous membrane structures and distinguishing them from background noise. While supervised methods like MemBrainSeg were trained on real tomograms, CryoSiam relied solely on simulated data yet achieved comparable segmentation results. This success highlights

the generalizability of the framework's voxel embeddings, even without fine-tuning on real data. Importantly, CryoSiam was also able to detect membrane-associated particles, a particularly challenging task due to the variability in particle interactions with cellular membranes. Identifying these particles within membrane boundaries is crucial for understanding cellular processes, such as transcription-translation coupling, and demonstrates CryoSiam's practical utility in complex biological analyses.

CryoSiam exhibited robust performance in detecting and classifying macromolecular complexes in the instance segmentation and particle identification tasks. The framework successfully performed instance segmentation on real tomograms, identifying distinct particles across different datasets. The clustering of subtomogram embeddings revealed meaningful separations between particle classes, as shown through UMAP visualizations. While the initial K-means clustering step provided a functional first layer of separation, the second-stage spectral clustering further refined particle identification, enabling the detection of biologically relevant complexes such as ribosomes, RNA polymerase, and ribosomal subunits.

One key finding from the particle identification evaluation is CryoSiam's ability to capture complex particle interactions in real tomograms. For example, clusters corresponding to ribosomes contained transcription-translation complexes, where ribosomes interact with RNA polymerase. These interactions, critical to understanding bacterial gene expression, were identified directly from the tomograms without any prior annotations or labels. The ability to detect such complexes highlights CryoSiam's practical applicability in biological discovery and reinforces the value of unsupervised learning approaches for cryo-ET analysis.

Despite these successes, several limitations emerged during the evaluation. The missing wedge artifact, a common issue in cryo-ET data due to limited angular coverage during tilt-series acquisition, poses a significant challenge for particle identification. The anisotropic resolution caused by the missing wedge leads to incomplete representations of particles in tomograms, making it difficult for the embedding models to capture particle shapes and interactions fully. While CryoSiam demonstrated resilience to noise and variability, it remains sensitive to the missing wedge, particularly when identifying smaller particles or particles with elongated geometries.

Another limitation identified during the evaluation is the framework's sensitivity to sample variability. While CryoSiam successfully identified ribosomes, RNA polymerase, and other macromolecular complexes in Mycoplasma pneumoniae tomograms, its performance on more structurally diverse datasets revealed areas for improvement. For example, the semantic segmentation model trained on simulated data did not fully capture actin filaments and microtubules, which exhibit a range of conformations and organizational patterns. This limitation emphasizes the need for further refinement of the training datasets to include more diverse particle types and conformational states.

The evaluation also highlighted the importance of second-stage clustering for particle identification. Spectral clustering revealed subclusters that provided more refined separations between particle types. In particular, identifying subclusters corresponding to ribosomal subunits and RNA polymerase

complexes demonstrates the utility of spectral clustering in distinguishing between particles that are otherwise difficult to separate using more straightforward methods like k-means. This two-stage clustering approach improved particle classification and provided insights into the structural variability within each cluster, suggesting potential biological relevance.

In addition to structural variability, the appearance of particles under different imaging conditions remains a challenge. For instance, the CryoSiam framework performed well on tomograms acquired with defocus imaging conditions but showed reduced sensitivity to specific structures in VP) tomograms. This discrepancy likely stems from the simulated training data not accounting for the unique contrast characteristics introduced by VPP imaging. Addressing this limitation would require the generation of additional simulated datasets that replicate VPP imaging conditions, improving CryoSiam's generalization to all types of real tomograms.

Overall, the evaluation of CryoSiam across multiple cryo-ET datasets demonstrates the framework's potential as a robust tool for automated cryo-ET analysis. Its ability to perform denoising, segmentation, and particle identification without prior training on real data presents a significant advancement in the field. However, future iterations of CryoSiam must address the limitations identified during this study, particularly the handling of the missing wedge artifact, including more diverse training datasets and adaptations to different imaging conditions. These improvements will further enhance CryoSiam's applicability to real-world cryo-ET workflows, paving the way for more accurate and efficient structural biology research.

In summary, the evaluation of CryoSiam underscores its strengths in denoising and segmentation and its capacity to detect complex particle interactions. While some challenges remain, particularly related to the missing wedge and sample variability, the framework provides a promising foundation for future advancements in cryo-ET analysis. By building on these findings, CryoSiam has the potential to become a key component of automated cryo-ET workflows, driving discoveries in structural biology.

6.3 LIMITATIONS OF THE PROPOSED FRAMEWORK

While CryoSiam demonstrated strong performance across various tasks in cryo-ET analysis, several limitations became evident while evaluating real tomograms. These limitations highlight key areas where improvements are needed to enhance the framework's robustness, generalizability, and practical applicability.

One of the most significant limitations stems from the missing wedge artifact, an inherent issue in cryo-ET data caused by incomplete angular sampling during tilt-series acquisition. The missing wedge introduces anisotropic resolution, resulting in particles that are incompletely represented in tomograms. This artifact poses a challenge for segmentation and particle identification, as the CryoSiam framework was trained on simulated datasets that do not account for these anisotropies. While CryoSiam's embeddings proved resilient to noise and other imaging inconsistencies, they remain sensitive to the missing wedge, particularly when distinguishing particles with elongated geometries

or asymmetrical structures. Addressing this limitation would require the incorporation of simulated datasets without missing wedges and training a task for correcting the missing wedge to compensate for anisotropic distortions in real tomograms.

Another limitation relates to the variability of biological structures present in real tomograms. While CryoSiam performed well in detecting membrane structures and larger particles like ribosomes and RNA polymerase, it struggled to generalize to more intricate structural features, such as actin filaments. Consequently, CryoSiam's semantic segmentation model was less effective in capturing the complexity of these structures. This limitation underscores the importance of expanding the diversity of training datasets to include a broader range of structural complexity. In particular, simulating different types of actin filaments, microtubule geometries, and other filamentous structures would improve the model's generalization to real biological samples.

CryoSiam's reliance on simulated data presents both a strength and a limitation. Using SSL on simulated datasets enabled the framework to perform strongly without requiring annotated real tomograms. However, simulated data inherently differ from real data in several important aspects, including noise characteristics, imaging artifacts, and structural variability. While CryoSiam successfully transferred its embeddings to real data, certain discrepancies between simulated and real tomograms remain apparent. For example, the appearance of particles under different imaging conditions, such as defocus and VPP imaging, varied significantly from the simulated training data. CryoSiam performed better on defocus tomograms than VPP tomograms, likely because the simulated data did not account for the unique contrast characteristics introduced by VPP imaging. Incorporating simulated datasets replicating different imaging conditions would improve the framework's adaptability to various experimental setups.

Another limitation identified during particle identification was distinguishing smaller particles or particles with low contrast. While CryoSiam effectively identified large particles like ribosomes and RNA polymerase, detecting smaller particles was more challenging. This limitation is likely due to both the resolution of the embeddings and the clustering methods used. The initial K-means clustering provided a functional first layer of separation but struggled to distinguish smaller, less prominent particles. The second-stage spectral clustering improved particle identification by refining clusters into subclusters, but further improvements are needed to handle small particles more effectively. This task would benefit from more refined embeddings, advanced clustering techniques, and input from domain experts to verify and validate the biological relevance of the identified clusters.

A related limitation concerns the need for manual inspection and validation. While CryoSiam automates many steps in the cryo-ET analysis workflow, interpreting clustering results and identifying specific particles still require expert knowledge. For example, ribosome subunits and RNA polymerase interactions were identified based on comparisons with known particle structures from simulated data. However, the framework does not currently provide an automated mechanism for verifying the identity of unknown particles or interactions. Future work could focus on integrating additional tools, such as

automated template matching or subtomogram averaging, to provide more definitive particle identifications.

Despite these limitations, CryoSiam's performance on real cryo-ET datasets demonstrates the framework's potential to address key challenges in structural biology. By identifying areas for improvement, this thesis lays the groundwork for future enhancements that will further refine CryoSiam's capabilities. Addressing these limitations will require a combination of improved training datasets, advanced clustering methods, and collaborative efforts with structural biologists to validate the framework's outputs.

7 CONCLUSIONS AND FUTURE WORK

In this chapter, I reflect on this thesis's key findings and contributions, summarizing the outcomes achieved through the development and application of the CryoSiam framework. I discuss how this work addresses critical challenges in cryo-ET analysis, particularly the ability to transfer knowledge from simulated datasets to real-world tomograms without fine-tuning. The insights gained throughout this project are presented, along with a discussion of the framework's limitations and avenues for future improvements.

By outlining potential directions for further research, I aim to provide a clear vision of how CryoSiam can be refined and expanded to tackle more complex cryo-ET datasets and diverse biological systems. The chapter concludes by emphasizing the broader implications of this work for advancing automated cryo-ET analysis workflows and supporting structural biology research in addressing fundamental questions about cellular architecture and molecular interactions.

7.1 SUMMARY OF CONTRIBUTIONS

In this thesis, I introduced and validated the CryoSiam framework, an SSL approach for cryo-ET data analysis. This work addresses a significant gap in cryo-ET research by providing a robust method to transfer knowledge from simulated datasets to real tomograms without requiring annotated real-world data. The framework demonstrates strong performance across key cryo-ET tasks, including tomogram denoising, semantic segmentation, instance segmentation, and particle identification.

One of the core contributions is the development of DenseSimSiam, a voxel-level embedding model that captures fine structural details in tomograms, and SimSiam, a subtomogram-level embedding model used for particle identification. By integrating these components, I developed a versatile pipeline to process cryo-ET data from raw tomograms to downstream structural analysis, covering segmentation and identification tasks. Using SSL ensured that CryoSiam could generate meaningful embeddings even when trained solely on simulated data, proving its generalizability to real tomograms under varying imaging conditions.

The evaluation of real cryo-ET datasets, including EMPIAR-10499, EMPIAR-10988, and EMPIAR-11756, demonstrated CryoSiam's robustness in denoising and membrane segmentation tasks. Despite the lack of real data in its training set, CryoSiam's performance in membrane segmentation was comparable to that of state-of-the-art supervised methods. In particle identification, CryoSiam successfully identified biologically relevant structures in Mycoplasma pneumoniae cells, proving evidence of how SSL frameworks can contribute to discovering complex molecular interactions in cryo-ET data.

In addition to the practical contributions, I conducted extensive ablation studies to identify the key design choices that influence CryoSiam's performance. These studies provided valuable insights into embedding sizes, loss functions, and transformations that maximize the framework's effectiveness. The results of these experiments informed the final design of CryoSiam, ensuring that it can handle real cryo-ET datasets with diverse biological and imaging conditions.

This work contributes a novel technical solution and demonstrates the feasibility of using simulated data to bridge the gap between experimental cryo-ET data and automated analysis workflows. By reducing the reliance on annotated real data, CryoSiam presents an opportunity to accelerate structural biology research and enhance the accessibility of cryo-ET analysis for researchers across the field.

7.2 INSIGHTS GAINED FROM REAL DATA EVALUATION

Applying CryoSiam to real cryo-ET datasets provided valuable insights into the framework's practical capabilities and limitations when transitioning from simulated to experimental data. The evaluation confirmed that models trained exclusively on simulated tomograms can generate meaningful predictions on real datasets without fine-tuning. This result underscores the robustness of SSL for cryo-ET analysis. It highlights the potential of using simulated data to bridge the gap between experimental and computational workflows.

One of the most significant insights gained is the framework's ability to generalize membrane segmentation to diverse real tomograms. For instance, the segmentation results on EMPIAR-10988 revealed that CryoSiam can accurately identify continuous membrane structures across different imaging conditions, even in the presence of noise and anisotropic artifacts like the missing wedge. The framework's membrane predictions showed consistency across datasets with varied biological compositions and imaging setups, demonstrating that the embedding space generated by DenseSimSiam captures relevant features beyond the training environment.

The particle identification results on the EMPIAR-10499 dataset highlighted the versatility of the subtomogram embedding approach. By clustering subtomogram embeddings, CryoSiam successfully identified biologically meaningful structures such as ribosomes and their subunits. These results suggest that the embeddings contain sufficient information to differentiate between particles of varying sizes and functions, even in noisy tomograms. However, the analysis also revealed the complexity of real cellular environments, where particles interact and form composite structures. The framework's ability to detect these interactions highlights its practical utility for studying molecular dynamics in situ.

Despite these successes, the evaluation also brought attention to several challenges that remain to be addressed. The missing wedge artifact in cryo-ET data presents a significant obstacle to achieving uniform segmentation and particle identification across all orientations. This anisotropy can hinder the recognition of smaller particles or complex structures, particularly those with weak contrast or ambiguous boundaries. Additionally, the diversity of structures

in real cells, especially for particles like actin filaments and microtubules, is far greater than what was captured in the simulated training data, which impacted the framework's ability to generalize to these classes.

Another important takeaway is the influence of imaging conditions on the results. For example, while CryoSiam performed well on defocus datasets, its performance on VPP tomograms indicated potential limitations in handling data collected under unique imaging conditions not represented in the training set. This suggests incorporating a broader range of imaging parameters into the simulated datasets could improve the framework's adaptability.

Overall, the real data evaluation demonstrated that CryoSiam provides a solid foundation for cryo-ET analysis, offering robust performance across key tasks without relying on annotated real data. However, it also highlighted areas where further improvements are necessary to enhance the framework's performance, particularly in addressing artifacts, handling diverse imaging conditions, and capturing the full range of structural variability in real biological samples.

7.3 LIMITATIONS AND CHALLENGES

While CryoSiam has demonstrated promising capabilities across many cryo-ET tasks, several limitations became evident while evaluating both simulated and real datasets. These challenges underscore areas where further refinement is necessary to improve the framework's robustness and practical utility in real-world scenarios.

One critical limitation is the reliance on simulated data for training. Simulated tomograms provide an invaluable resource for SSL, allowing models to be trained without the need for annotated real tomograms. However, this reliance introduces potential domain gaps, as simulated data can only approximate the biological complexities in real tomograms. As a result, CryoSiam struggled to generalize to certain particle types that were either underrepresented or oversimplified in the simulated datasets, such as actin filaments and microtubules. These structures exhibit significant variability in real cells, posing additional challenges for accurate segmentation and particle identification.

The missing wedge artifact also emerged as a challenge during the evaluation. This artifact, inherent to cryo-ET due to limited angular coverage during tilt-series acquisition, causes anisotropic resolution in reconstructed tomograms. Although CryoSiam demonstrated robustness in handling noise and structural variability, the missing wedge remains a critical obstacle, particularly for identifying particles that are oriented along missing wedge directions. Addressing this artifact requires incorporating compensation strategies during training or embedding generation to improve the model's generalization ability across anisotropic reconstructions.

Another notable limitation was the diversity of training data used to develop the framework. The simulated datasets comprised only a handful of particle types and structural variations. Expanding the variety of training data to include more realistic biological structures, such as filamentous proteins, diverse membrane morphologies, and complex macromolecular assemblies, could sig-

nificantly improve the generalization of CryoSiam across a broader range of cellular environments.

Additionally, CryoSiam's performance varied depending on the imaging conditions of the tomograms. The framework achieved better segmentation results on defocus tomograms than VPP tomograms, suggesting that certain imaging conditions present challenges that were not fully addressed during training. Incorporating simulated tomograms with varied imaging parameters, such as different defocus levels, contrast transfer functions, and noise profiles, could further enhance CryoSiam's ability to adapt to diverse real-world datasets.

The particle identification pipeline revealed another challenge: interpretability. While CryoSiam effectively clustered subtomograms and separated distinct particle types, the biological relevance of these clusters required further validation through subtomogram averaging or manual curation by domain experts. This reliance on external verification steps indicates that CryoSiam's clustering methods, though effective, could be made more autonomous by integrating additional refinement processes or downstream analysis tools.

A final challenge was the limited availability of comprehensive ground truth annotations for real datasets. This lack of ground truth made it challenging to evaluate the framework's performance for specific tasks quantitatively, particularly instance segmentation and particle identification. Most evaluations relied on qualitative assessments, which, while insightful, do not offer the same level of rigor as quantitative benchmarks. Developing strategies to generate more complete annotations, such as leveraging semi-automated annotation tools or crowdsourced efforts, could address this limitation and enhance the evaluation process.

Despite these challenges, CryoSiam demonstrates remarkable potential as a robust framework for cryo-ET analysis. The ability to apply models trained solely on simulated data to real tomograms is a significant advancement in the field, reducing the dependency on annotated datasets and offering a scalable solution for structural biology research. Addressing the identified limitations will be crucial for future development, ensuring that CryoSiam can continue contributing to the advancement of cryo-ET workflows and the broader understanding of cellular structures.

7.4 BROADER IMPLICATIONS AND FUTURE DIRECTIONS

The CryoSiam framework presents a significant step forward in cryo-ET data analysis by introducing self-supervised learning techniques capable of generalizing from simulated to real tomograms. Its development aligns with the growing need for automated tools that reduce the reliance on annotated datasets while maintaining high accuracy across tasks such as denoising, segmentation, and particle identification. However, the impact of CryoSiam extends beyond its immediate applications in this thesis, pointing toward broader implications and future directions that could shape the evolution of cryo-ET workflows.

One of the most profound implications of this work is the potential to democratize cryo-ET analysis by enabling researchers to process their data without requiring large annotated datasets. The traditional reliance on supervised

learning methods has placed a considerable burden on researchers to annotate tomograms, a time-intensive and often subjective process manually. By demonstrating that models trained on simulated data can achieve comparable results to those trained on real annotations, CryoSiam provides a pathway to more accessible and scalable analysis tools. This democratization could accelerate discoveries in structural biology, particularly for less-resourced laboratories or research groups focusing on underrepresented biological systems.

Another important aspect of CryoSiam is its ability to handle complex real tomograms, including those with varying noise levels, imaging artifacts, and structural complexities. This capability highlights the growing importance of developing robust, adaptable tools to account for the variability inherent in cryo-ET data. However, future work must address unaddressed challenges, such as the missing wedge artifact, variations in imaging conditions, and the need for more diverse training datasets. CryoSiam could further enhance its generalizability and reliability across a broader range of cryo-ET datasets by incorporating these improvements.

An inspiring future direction is integrating CryoSiam's particle identification module into workflows to construct whole-cell models from cryo-ET data. The ability to retrieve, classify, and spatially segregate particles from entire tomograms lays the foundation for building comprehensive cellular reconstructions. Unlike traditional TM approaches, which constrain particles using predefined particle libraries, CryoSiam's clustering-based identification method can discover previously uncharacterized particles or interactions. This potential to achieve unbiased, large-scale particle identification could significantly contribute to developing whole-cell models, offering a more complete picture of cellular organization and interactions at molecular resolution.

Whole-cell modeling from cryo-ET data represents a transformative goal in structural biology, where the spatial arrangement of cellular components is reconstructed to reveal the dynamic interplay between macromolecules within their native environment. CryoSiam's capacity to provide detailed, accurate particle maps could be a foundational tool. However, achieving this goal will require further advancements, particularly in resolving smaller particles, handling structural variability, and addressing artifacts introduced by imaging conditions. Additionally, integrating CryoSiam's outputs with downstream structural refinement methods, such as subtomogram averaging, will be essential to achieve high-resolution insights from the identified particles.

A promising avenue for future research is integrating domain adaptation techniques to bridge the gap between simulated and real data more effectively. While CryoSiam has successfully transferred knowledge from simulated to real tomograms, domain adaptation could help further reduce discrepancies between these two domains by fine-tuning models with limited real data or using adversarial training to align simulated and real data distributions. Exploring hybrid approaches that combine self-supervised pretraining with supervised fine-tuning on a small set of real annotations could balance scalability and accuracy.

Expanding the diversity of simulated training data remains a priority for future improvements. The CryoSiam framework primarily focuses on common structural elements such as membranes, ribosomes, and generic parti-

cles. Incorporating more varied biological structures, such as different filamentous proteins, organelles, and macromolecular complexes, into the simulated datasets could significantly enhance the model's capacity to generalize to a broader range of real biological systems. This task will require collaboration with structural biologists to ensure that simulated data captures the heterogeneity observed in actual cells.

Furthermore, CryoSiam's clustering-based particle identification introduces new challenges related to interpretability and validation. While the framework excels at identifying particle clusters, future research must focus on developing methods to validate these clusters, potentially through subtomogram averaging or integration with other structural analysis tools. Collaborations with domain experts in structural biology will be essential to interpret these clusters correctly and to ensure that CryoSiam's outputs lead to meaningful biological insights.

In conclusion, the development of CryoSiam marks a step toward more automated, scalable, and accessible cryo-ET analysis workflows. The framework's ability to perform robust particle identification and membrane segmentation directly on real tomograms highlights its potential to become a core tool in cryo-ET analysis. The insights gained from this thesis pave the way for future advancements, particularly in developing whole-cell models and comprehensive structural reconstructions. By addressing current limitations and exploring new methodologies, CryoSiam can contribute to a deeper understanding of cellular architecture and dynamics, ultimately transforming how we study biological systems at molecular resolution.

BIBLIOGRAPHY

[1] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, et al. *Molecular biology of the cell*. Vol. 3. Garland New York, 1994.

[2] Hong-Wei Wang and Jia-Wei Wang. "How cryo-electron microscopy and X-ray crystallography complement each other". In: *Protein Science* 26.1 (2017), pp. 32–39.

[3] Georgios Skiniotis and Daniel R Southworth. "Single-particle cryo-electron microscopy of macromolecular complexes". In: *Journal of Electron Microscopy* 65.1 (2015), pp. 9–22.

[4] Werner Kühlbrandt. "The resolution revolution". In: *Science* 343.6178 (2014), pp. 1443–1444.

[5] Joachim Frank. "Advances in the field of single-particle cryo-electron microscopy over the last decade". In: *Nature protocols* 12.2 (2017), pp. 209–212.

[6] Wolfgang Baumeister. "Cryo-electron tomography: A long journey to the inner space of cells". In: *Cell* 185.15 (2022), pp. 2649–2652.

[7] Vladan Lučić, Friedrich Förster, and Wolfgang Baumeister. "Structural studies by electron tomography: from cells to molecules". In: *Annu. Rev. Biochem.* 74.1 (2005), pp. 833–865.

[8] Vladan Lučić, Andrew Leis, and Wolfgang Baumeister. "Cryo-electron tomography of cells: connecting structure and function". In: *Histochemistry and Cell Biology* 130 (2008), pp. 185–196.

[9] Julia Mahamid, Stefan Pfeffer, Miroslava Schaffer, Elizabeth Villa, Radostin Danev, Luis Kuhn Cuellar, Friedrich Förster, Anthony A Hyman, Jürgen M Plitzko, and Wolfgang Baumeister. "Visualizing the molecular sociology at the HeLa cell nuclear periphery". In: *Science* 351.6276 (2016), pp. 969–972.

[10] Muyuan Chen, Wei Dai, Stella Y Sun, Darius Jonasch, Cynthia Y He, Michael F Schmid, Wah Chiu, and Steven J Ludtke. "Convolutional neural networks for automated annotation of cellular cryo-electron tomograms". In: *Nature methods* 14.10 (2017), pp. 983–985.

[11] Cuicui Zhao, Da Lu, Qian Zhao, Chongjiao Ren, Huangtao Zhang, Jiaqi Zhai, Jiaxin Gou, Shilin Zhu, Yaqi Zhang, and Xinqi Gong. "Computational methods for in situ structural studies with cryogenic electron tomography". In: *Frontiers in Cellular and Infection Microbiology* 13 (2023), p. 1135013.

[12] Mohammed Majid Abdulrazzaq, Nehad TA Ramaha, Alaa Ali Hameed, Mohammad Salman, Dong Keon Yon, Norma Latif Fitriyani, Muhammad Syafrudin, and Seung Won Lee. "Consequential Advancements of Self-Supervised Learning (SSL) in Deep Learning Contexts". In: *Mathematics* 12.5 (2024), p. 758.

[13] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al. "Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks". In: *Advances in Neural Information Processing Systems* 36 (2024).

[14] Tarun Gupta, Xuehai He, Mostofa Rafid Uddin, Xiangrui Zeng, Andrew Zhou, Jing Zhang, Zachary Freyberg, and Min Xu. "Self-supervised learning for macromolecular structure classification based on cryo-electron tomograms". In: *Frontiers in Physiology* 13 (2022), p. 957484.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020. arXiv: 1911.05722 [cs.CV].

[17] Lindsey N Young and Elizabeth Villa. "Bringing structure to cell biology with cryo-electron tomography". In: *Annual review of biophysics* 52.1 (2023), pp. 573–595.

[18] Sharon Grayer Wolf, Lothar Houben, and Michael Elbaum. "Cryo-scanning transmission electron tomography of vitrified cells". In: *Nature methods* 11.4 (2014), pp. 423–428.

[19] Wim JH Hagen, William Wan, and John AG Briggs. "Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging". In: *Journal of structural biology* 197.2 (2017), pp. 191–198.

[20] Martin Turk and Wolfgang Baumeister. "The promise and the challenges of cryo-electron tomography". In: *FEBS letters* 594.20 (2020), pp. 3243–3261.

[21] Benjamin A Barad, Michaela Medina, Daniel Fuentes, R Luke Wiseman, and Danielle A Grotjahn. "Quantifying organellar ultrastructure in cryo-electron tomography using a surface morphometrics pipeline". In: *Journal of Cell Biology* 222.4 (2023), e202204093.

[22] Barrett M Powell, Tyler S Brant, Joseph H Davis, and Shyamal Mosalaganti. "Rapid structural analysis of bacterial ribosomes in situ". In: *bioRxiv* (2024).

[23] Vladan Lučić, Alexander Rigort, and Wolfgang Baumeister. "Cryo-electron tomography: the challenge of doing structural biology in situ". In: *Journal of Cell Biology* 202.3 (2013), pp. 407–419.

[24] Saikat Chakraborty, Marion Jasnin, and Wolfgang Baumeister. "Three-dimensional organization of the cytoskeleton: a cryo-electron tomography perspective". In: *Protein Science* 29.6 (2020), pp. 1302–1320.

[25] Georges Chreifi, Songye Chen, and Grant J Jensen. "Rapid tilt-series method for cryo-electron tomography: Characterizing stage behavior during FISE acquisition". In: *Journal of structural biology* 213.2 (2021), p. 107716.

[26] Stefan Pfeffer and Julia Mahamid. "Unravelling molecular complexity in structural cell biology". In: *Current opinion in structural biology* 52 (2018), pp. 111–118.

[27] Vinson Lam and Elizabeth Villa. "Practical approaches for cryo-FIB milling and applications for cellular cryo-electron tomography". In: *cryoEM: Methods and Protocols* (2021), pp. 49–82.

[28] Robert Englmeier and Friedrich Förster. "Cryo-electron tomography for the structural study of mitochondrial translation". In: *Tissue and Cell* 57 (2019), pp. 129–138.

[29] Emmanuelle RJ Quemin, Emily A Machala, Benjamin Vollmer, Vojtěch Pražák, Daven Vasishtan, Rene Rosch, Michael Grange, Linda E Franken, Lindsay A Baker, and Kay Grünewald. "Cellular electron cryo-tomography to study virus-host interactions". In: *Annual review of virology* 7.1 (2020), pp. 239–262.

[30] 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". In: *Elife* 6 (2017), e32493.

[31] Florian KM Schur, Martin Obr, Wim JH Hagen, William Wan, Arjen J Jakobi, Joanna M Kirkpatrick, Carsten Sachse, Hans-Georg Kräusslich, and John AG Briggs. "An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation". In: *Science* 353.6298 (2016), pp. 506–508.

[32] Benjamin E Bammes, Ryan H Rochat, Joanita Jakana, Dong-Hua Chen, and Wah Chiu. "Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency". In: *Journal of structural biology* 177.3 (2012), pp. 589–601.

[33] Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J Noble, and Bonnie Berger. "TOPAZ: A positive-unlabeled convolutional neural network CryoEM particle picker that can pick any size and shape particle". In: *Microscopy and Microanalysis* 25.S2 (2019), pp. 986–987.

[34] Daniel Asarnow, Vada A Becker, Daija Bobe, Charlie Dubbledam, Jake D Johnston, Mykhailo Kopylov, Nathalie R Lavoie, Qiuye Li, Jacob M Mattingly, Joshua H Mendez, et al. "Recent advances in infectious disease research using cryo-electron tomography". In: *Frontiers in Molecular Biosciences* 10 (2024), p. 1296941.

[35] Peijun Zhang. "Correlative cryo-electron tomography and optical microscopy of cells". In: *Current opinion in structural biology* 23.5 (2013), pp. 763–770.

[36] Hana Nedozralova, Nirakar Basnet, Iosune Ibiricu, Satish Bodakuntla, Christian Biertümpfel, and Naoko Mizuno. "In situ cryo-electron tomography reveals local cellular machineries for axon branch development". In: *Journal of Cell Biology* 221.4 (2022), e202106086.

[37] Emmanuel Moebel and Charles Kervrann. "Towards unsupervised classification of macromolecular complexes in cryo electron tomography: Challenges and opportunities". In: *Computer Methods and Programs in Biomedicine* 225 (2022), p. 107017.

[38] Ryan K Hylton and Matthew T Swulius. "Challenges and triumphs in cryo-electron tomography". In: *Iscience* 24.9 (2021).

[39] Euan Pyle and Giulia Zanetti. "Current data processing strategies for cryo-electron tomography and subtomogram averaging". In: *Biochemical journal* 478.10 (2021), pp. 1827–1845.

[40] Barrett M Powell and Joseph H Davis. "Learning structural heterogeneity from cryo-electron sub-tomograms with tomoDRGN". In: *Nature Methods* (2024), pp. 1–12.

[41] Xiangrui Zeng, Yizhe Ding, Yueqian Zhang, Mostofa Rafid Uddin, Ali Dabouei, and Min Xu. "DUAL: deep unsupervised simultaneous simulation and denoising for cryo-electron tomography". In: *bioRxiv* (2024).

[42] Hongjia Li, Hui Zhang, Xiaohua Wan, Zhidong Yang, Chengmin Li, Jintao Li, Renmin Han, Ping Zhu, and Fa Zhang. "Noise-Transfer2Clean: denoising cryo-EM images based on noise modeling and transfer". In: *Bioinformatics* 38.7 (2022), pp. 2022–2029.

[43] Han Xue, Meng Zhang, Jianfang Liu, Jianjun Wang, and Gang Ren. "Cryo-electron tomography related radiation-damage parameters for individual-molecule 3D structure determination". In: *Frontiers in Chemistry* 10 (2022), p. 889203.

[44] Achilleas S Frangakis. "It's noisy out there! A review of denoising techniques in cryo-electron tomography". In: *Journal of Structural Biology* 213.4 (2021), p. 107804.

[45] Tim-Oliver Buchholz, Mareike Jordan, Gaia Pigino, and Florian Jug. "Cryo-CARE: Content-aware image restoration for cryo-transmission electron microscopy data". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 502–506.

[46] Radostin Danev, Hirofumi Iijima, Mizuki Matsuzaki, and Sohei Motoki. "Fast and accurate defocus modulation for improved tunability of cryo-EM experiments". In: *IUCrJ* 7.3 (2020), pp. 566–574.

[47] Markus Matthias Hohle, Katja Lammens, Fabian Gut, Bingzhi Wang, Sophia Kahler, Kathrin Kugler, Michael Till, Roland Beckmann, Karl-Peter Hopfner, and Christophe Jung. "Ice thickness monitoring for cryo-EM grids by interferometry imaging". In: *Scientific Reports* 12.1 (2022), p. 15330.

[48] Felix JB Bäuerlein, Max Renner, Dana El Chami, Stephan E Lehnart, José C Pastor-Pareja, and Rubén Fernández-Busnadiego. "Cryo-electron tomography of large biological specimens vitrified by plunge freezing". In: *BioRxiv* (2021), pp. 2021–04.

[49] Kasahun Neselu, Bing Wang, William J Rice, Clinton S Potter, Bridget Carragher, and Eugene YD Chua. "Measuring the effects of ice thickness on resolution in single particle cryo-EM". In: *Journal of Structural Biology: X* 7 (2023), p. 100085.

[50] Jessica E Heebner, Carson Purnell, Ryan K Hylton, Mike Marsh, Michael A Grillo, and Matthew T Swulius. "Deep learning-based segmentation of cryo-electron tomograms". In: *J. Vis. Exp* 64435.10.3791 (2022), p. 64435.

[51] Felix Weis, Wim JH Hagen, Martin Schorb, and Simone Mattei. "Strategies for optimization of cryogenic electron tomography data acquisition". In: *JoVE (Journal of Visualized Experiments)* 169 (2021), e62383.

[52] Julio Kovacs, Junha Song, Manfred Auer, Jing He, Wade Hunter, and Willy Wriggers. "Correction of missing-wedge artifacts in filamentous tomograms by template-based constrained deconvolution". In: *Journal of chemical information and modeling* 60.5 (2020), pp. 2626–2633.

[53] Haonan Zhang, Yan Li, Yanan Liu, Dongyu Li, Lin Wang, Kai Song, Keyan Bao, and Ping Zhu. "A method for restoring signals and revealing individual macromolecule states in cryo-ET, REST". In: *Nature Communications* 14.1 (2023), p. 2937.

[54] Yun-Tao Liu, Heng Zhang, Hui Wang, Chang-Lu Tao, Guo-Qiang Bi, and Z Hong Zhou. "Isotropic reconstruction for electron tomography with deep learning". In: *Nature communications* 13.1 (2022), p. 6482.

[55] Dave Van Veen, Jesús G Galaz-Montoya, Liyue Shen, Philip Baldwin, Akshay S Chaudhari, Dmitry Lyumkis, Michael F Schmid, Wah Chiu, and John Pauly. "Missing wedge completion via unsupervised learning with coordinate networks". In: *International Journal of Molecular Sciences* 25.10 (2024), p. 5473.

[56] Guanglei Ding, Yitong Liu, Rui Zhang, and Huolin L Xin. "A joint deep learning model to recover information and reduce artifacts in missing-wedge sinograms for electron tomography and beyond". In: *Scientific reports* 9.1 (2019), p. 12803.

[57] Rebecca F Thompson, Matt Walker, C Alistair Siebert, Stephen P Muench, and Neil A Ranson. "An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology". In: *Methods* 100 (2016), pp. 3–15.

[58] Irene de Teresa-Trueba, Sara K Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E Zimmerli, Mauricio Toro-Nahuelpan, Dorothy WC Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, et al. "Convolutional networks for supervised mining of molecular patterns within cellular context". In: *Nature Methods* 20.2 (2023), pp. 284–294.

[59] Lorenz Lamm, Simon Zufferey, Ricardo D Righetto, Wojciech Wietrzynski, Kevin A Yamauchi, Alister Burt, Ye Liu, Hanyi Zhang, Antonio Martinez-Sanchez, Sebastian Ziegler, et al. "MemBrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography". In: *bioRxiv* (2024), pp. 2024–01.

[60] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al. "Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms". In: *Nature methods* 18.11 (2021), pp. 1386–1394.

[61] Corey W Hecksel, Michele C Darrow, Wei Dai, Jesús G Galaz-Montoya, Jessica A Chin, Patrick G Mitchell, Shurui Chen, Jemba Jakana, Michael F Schmid, and Wah Chiu. "Quantifying variability of manual annotation in cryo-electron tomograms". In: *Microscopy and Microanalysis* 22.3 (2016), pp. 487–496.

[62] Paula P Navarro. "Quantitative cryo-electron tomography". In: *Frontiers in Molecular Biosciences* 9 (2022), p. 934465.

[63] Xiangrui Zeng, Anson Kahng, Liang Xue, Julia Mahamid, Yi-Wei Chang, and Min Xu. "High-throughput cryo-ET structural pattern mining by unsupervised deep iterative subtomogram clustering". In: *Proceedings of the National Academy of Sciences* 120.15 (2023), e2213149120.

[64] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. "DeepETPicker: Fast and accurate 3D particle picking for cryo-electron tomography using weakly supervised deep learning". In: *Nature Communications* 15.1 (2024), p. 2090.

[65] Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. "TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining". In: *Nature methods* 20.6 (2023), pp. 871–880.

[66] Yizhou Zhao, Hengwei Bian, Michael Mu, Mostofa R Uddin, Zhenyang Li, Xiang Li, Tianyang Wang, and Min Xu. "Cryosam: Training-free cryoet tomogram segmentation with foundation models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 124–134.

[67] Qinwen Huang, Ye Zhou, and Alberto Bartesaghi. "MiLoPYP: self-supervised molecular pattern mining and particle localization in situ". In: *Nature Methods* 21.10 (2024), pp. 1863–1872.

[68] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. "Self-supervised representation learning: Introduction, advances, and challenges". In: *IEEE Signal Processing Magazine* 39.3 (2022), pp. 42–62.

[69] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6707–6717.

[70] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. "Masked siamese networks for label-efficient learning". In: *European Conference on Computer Vision*. Springer. 2022, pp. 456–473.

[71] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, Xuelong Li, and Yue Lu. "Distance-based Weighted Transformer Network for image completion". In: *Pattern Recognition* 147 (2024), p. 110120.

[72] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 16684–16693.

[73] Dongsu Zhang, Junha Chun, Sang Kyun Cha, and Young Min Kim. "Spatial semantic embedding network: Fast 3d instance segmentation with deep metric learning". In: *arXiv preprint arXiv:2007.03169* (2020).

[74] Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. "Self-Supervised Anomaly Detection with Neural Transformations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[75] Jiaye Teng, Weiran Huang, and Haowei He. "Can pretext-based self-supervised learning be boosted by downstream data? a theoretical analysis". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 4198–4216.

[76] Wataru Shimoda and Keiji Yanai. "Self-supervised difference detection for weakly-supervised semantic segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5208–5217.

[77] Rob Chew, Michael Wenger, Caroline Kery, Jason Nance, Keith Richards, Emily Hadley, and Peter Baumgartner. "SMART: an open source data labeling platform for supervised learning". In: *Journal of Machine Learning Research* 20.82 (2019), pp. 1–5.

[78] Ariana Peck, Yue Yu, Jonathan Schwartz, Anchi Cheng, Utz Heinrich Ermel, Saugat Kandel, Dari Kimanius, Elizabeth Montabana, Daniel Serwas, Hannah Siems, et al. "Annotating CryoET Volumes: A Machine Learning Challenge". In: *bioRxiv* (2024), pp. 2024–11.

[79] Thalles Silva, Helio Pedrini, and Adín Ramírez. "Self-supervised learning of contextualized local visual embeddings". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 177–186.

[80] Donal M McSweeney, Sean M McSweeney, and Qun Liu. "A self-supervised workflow for particle picking in cryo-EM". In: *IUCrJ* 7.4 (2020), pp. 719–727.

[81] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. "Benchmarking self-supervised learning on diverse pathology datasets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3344–3354.

[82] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058.

[83] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. "Selfreg: Self-supervised contrastive regularization for domain generalization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9619–9628.

[84] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. "Towards the generalization of contrastive self-supervised learning". In: *arXiv preprint arXiv:2111.00743* (2021).

[85] Friedrich Förster, Sabine Pruggnaller, Anja Seybert, and Achilleas S Frangakis. "Classification of cryo-electron sub-tomograms using constrained correlation". In: *Journal of structural biology* 161.3 (2008), pp. 276–286.

[86] Valentin J Maurer, Marc Siggel, and Jan Kosinski. "What shapes template-matching performance in cryogenic electron tomography in situ?" In: *Acta Crystallographica Section D: Structural Biology* (2024).

[87] Ilja Gubins, Marten L Chaillet, Gijs van Der Schot, Remco C Veltkamp, Friedrich Förster, Yu Hao, Xiaohua Wan, Xuefeng Cui, Fa Zhang, Emmanuel Moebel, et al. "SHREC 2020: Classification in cryo-electron tomograms". In: *Computers & Graphics* 91 (2020), pp. 279–289.

[88] Andrea Raffo, Ulderico Fugacci, Silvia Biasotti, Walter Rocchia, Yonghuai Liu, Ekpo Otu, Reyer Zwiggelaar, David Hunter, Evangelia I Zacharaki, Eleftheria Psatha, et al. "SHREC 2021: Retrieval and classification of protein surfaces equipped with physical and chemical properties". In: *Computers & Graphics* 99 (2021), pp. 1–21.

[89] Carson Purnell, Jessica Heebner, Michael T Swulius, Ryan Hylton, Seth Kabonick, Michael Grillo, Sergei Grigoryev, Fred Heberle, M Neal Waxham, and Matthew T Swulius. "Rapid synthesis of cryo-et data for training deep learning models". In: *bioRxiv* (2023).

[90] Antonio Martinez-Sanchez, Lorenz Lamm, Marion Jasnin, and Harold Phelippeau. "Simulating the cellular context in synthetic datasets for cryo-electron tomography". In: *IEEE Transactions on Medical Imaging* (2024).

[91] Ye Hong, Yutong Song, Zheyuan Zhang, and Sai Li. "Cryo-electron tomography: the resolution revolution and a surge of in situ virological discoveries". In: *Annual Review of Biophysics* 52.1 (2023), pp. 339–360.

[92] Casper Berger, Navya Premaraj, Raimond BG Ravelli, Kèvin Knoops, Carmen López-Iglesias, and Peter J Peters. "Cryo-electron tomography on focused ion beam lamellae transforms structural cell biology". In: *Nature Methods* 20.4 (2023), pp. 499–511.

[93] Rui Yan, Singanallur V Venkatakrishnan, Jun Liu, Charles A Bouman, and Wen Jiang. "MBIR: A cryo-ET 3D reconstruction method that effectively minimizes missing wedge artifacts and restores missing information". In: *Journal of structural biology* 206.2 (2019), pp. 183–192.

[94] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. "Using self-supervised learning can improve model robustness and uncertainty". In: *Advances in neural information processing systems* 32 (2019).

[95] Tin Ki Tsang, Eric A Bushong, Daniela Boassa, Junru Hu, Benedetto Romoli, Sebastien Phan, Davide Dulcis, Chih-Ying Su, and Mark H Ellisman. "High-quality ultrastructural preservation using cryofixation for 3D electron microscopy of genetically labeled tissues". In: *Elife* 7 (2018), e35524.

[96] Charles J Hunt. "Cryopreservation: vitrification and controlled rate cooling". In: *Stem cell banking: concepts and protocols* (2017), pp. 41–77.

[97] Lucille A Giannuzzi and Frederick A Stevie. "A review of focused ion beam milling techniques for TEM specimen preparation". In: *Micron* 30.3 (1999), pp. 197–204.

[98] Christopher J Russo and Lori A Passmore. "Ultrastable gold substrates: properties of a support for high-resolution electron cryomicroscopy of biological specimens". In: *Journal of structural biology* 193.1 (2016), pp. 33–44.

[99] Miroslava Schaffer, Julia Mahamid, Benjamin D Engel, Tim Laugks, Wolfgang Baumeister, and Jürgen M Plitzko. "Optimized cryo-focused ion beam sample preparation aimed at in situ structural studies of membrane proteins". In: *Journal of structural biology* 197.2 (2017), pp. 73–82.

[100] Jose-Jesus Fernandez, Sam Li, Tanmay AM Bharat, and David A Agard. "Cryo-tomography tilt-series alignment with consideration of the beam-induced sample motion". In: *Journal of structural biology* 202.3 (2018), pp. 200–209.

[101] Shawn Zheng, Axel Brilot, Yifan Cheng, and David A Agard. "Beam-Induced Motion Mechanism and Correction for Improved Cryo-Electron Microscopy and Cryo-Electron Tomography". In: *Cryo-Electron Tomography: Structural Biology in situ*. Springer, 2024, pp. 293–314.

[102] Shawn Q Zheng, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng, and David A Agard. "MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy". In: *Nature methods* 14.4 (2017), pp. 331–332.

[103] Timothy Grant and Nikolaus Grigorieff. "Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6". In: *elife* 4 (2015), e06980.

[104] David N Mastronarde and Susannah R Held. "Automated tilt series alignment and tomographic reconstruction in IMOD". In: *Journal of structural biology* 197.2 (2017), pp. 102–113.

[105] Alex J Noble and Scott M Stagg. "Automated batch fiducial-less tilt-series alignment in Appion using Protomo". In: *Journal of structural biology* 192.2 (2015), pp. 270–278.

[106] Dimitry Tegunov and Patrick Cramer. "Real-time cryo-electron microscopy data preprocessing with Warp". In: *Nature methods* 16.11 (2019), pp. 1146–1152.

[107] Hanspeter Winkler and Kenneth A Taylor. "Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography". In: *Ultramicroscopy* 106.3 (2006), pp. 240–254.

[108] Shawn Zheng, Georg Wolff, Garrett Greenan, Zhen Chen, Frank GA Faas, Montserrat Bárcena, Abraham J Koster, Yifan Cheng, and David A Agard. "AreTomo: An integrated software package for automated marker-free, motion-corrected cryo-electron tomographic alignment and reconstruction". In: *Journal of Structural Biology: X* 6 (2022), p. 100068.

[109] Min Xu and Frank Alber. "High precision alignment of cryo-electron subtomograms through gradient-based parallel optimization". In: *BMC systems biology* 6 (2012), pp. 1–13.

[110] Giulia Zanetti, James D Riches, Stephen D Fuller, and John AG Briggs. "Contrast transfer function correction applied to cryo-electron tomography and sub-tomogram averaging". In: *Journal of structural biology* 168.2 (2009), pp. 305–312.

[111] Alexis Rohou and Nikolaus Grigorieff. "CTFFIND4: Fast and accurate defocus estimation from electron micrographs". In: *Journal of structural biology* 192.2 (2015), pp. 216–221.

[112] Kai Zhang. "Gctf: Real-time CTF determination and correction". In: *Journal of structural biology* 193.1 (2016), pp. 1–12.

[113] Beata Turoňová, Florian KM Schur, William Wan, and John AG Briggs. "Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4 Å". In: *Journal of structural biology* 199.3 (2017), pp. 187–195.

[114] JI Agulleiro and José-Jesús Fernandez. "Fast tomographic reconstruction on multicore computers". In: *Bioinformatics* 27.4 (2011), pp. 582–583.

[115] Euan Pyle, Joshua Hutchings, and Giulia Zanetti. "Strategies for picking membrane-associated particles within subtomogram averaging workflows". In: *Faraday Discussions* 240 (2022), pp. 101–113.

[116] Jasenko Zivanov, Joaquin Oton, Zunlong Ke, Andriko von Kugelgen, Euan Pyle, Kun Qu, Dustin Morado, Daniel Castanjo-Diez, Giulia Zanetti, Tanmay AM Bharat, et al. "A Bayesian approach to single-particle electron cryo-tomography in RELION-4.0". In: *Elife* 11 (2022), e83724.

[117] Michael Grange, Daven Vasishtan, and Kay Grünewald. "Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes". In: *Journal of Structural Biology* 197.2 (2017), pp. 181–190.

[118] Tanmay AM Bharat and Sjors HW Scheres. "Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION". In: *Nature protocols* 11.11 (2016), pp. 2054–2065.

[119] Daniel Castaño-Díez, Mikhail Kudryashev, Marcel Arheit, and Henning Stahlberg. "Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments". In: *Journal of structural biology* 178.2 (2012), pp. 139–151.

[120] Jesús G Galaz-Montoya, John Flanagan, Michael F Schmid, and Steven J Ludtke. "Single particle tomography in EMAN2". In: *Journal of structural biology* 190.3 (2015), pp. 279–290.

[121] Renmin Han, Lun Li, Peng Yang, Fa Zhang, and Xin Gao. "A novel constrained reconstruction model towards high-resolution subtomogram averaging". In: *Bioinformatics* 37.11 (2021), pp. 1616–1626.

[122] Kendra E Leigh, Paula P Navarro, Stefano Scaramuzza, Wenbo Chen, Yingyi Zhang, Daniel Castaño-Díez, and Misha Kudryashev. "Subtomogram averaging from cryo-electron tomograms". In: *Methods in cell biology* 152 (2019), pp. 217–259.

[123] J Bernard Heymann. "The progressive spectral signal-to-noise ratio of cryo-electron micrograph movies as a tool to assess quality and radiation damage". In: *Computer methods and programs in biomedicine* 220 (2022), p. 106799.

[124] Mikhail Kudryashev, Daniel Castaño-Díez, and Henning Stahlberg. "Limiting factors in single particle cryo electron tomography". In: *Computational and structural biotechnology journal* 1.2 (2012), e201207002.

[125] Tristan Bepler, Kotaro Kelley, Alex J Noble, and Bonnie Berger. "Topaz-Denoise: general deep denoising models for cryoEM and cryoET". In: *Nature communications* 11.1 (2020), p. 5208.

[126] Simon Wiedemann and Reinhard Heckel. "A deep learning method for simultaneous denoising and missing wedge reconstruction in cryogenic electron tomography". In: *Nature Communications* 15.1 (2024), p. 8255.

[127] Mart GF Last, Leoni Abendstein, Lenard M Voortman, and Thomas H Sharp. "Streamlining segmentation of cryo-electron tomography datasets with Ais". In: *eLife* 13 (2024), RP98552.

[128] Jochen Böhm, Achilleas S Frangakis, Reiner Hegerl, Stephan Nickell, Dieter Typke, and Wolfgang Baumeister. "Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms". In: *Proceedings of the National Academy of Sciences* 97.26 (2000), pp. 14245–14250.

[129] Martin Beck and Wolfgang Baumeister. "Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail?" In: *Trends in cell biology* 26.11 (2016), pp. 825–837.

[130] Detlev Stalling, Malte Westerhoff, Hans-Christian Hege, et al. "Amira: A highly interactive system for visual data analysis." In: *The visualization handbook* 38 (2005), pp. 749–767.

[131] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. "UCSF Chimera - a visualization system for exploratory research and analysis". In: *Journal of computational chemistry* 25.13 (2004), pp. 1605–1612.

[132] James R Kremer, David N Mastronarde, and J Richard McIntosh. "Computer visualization of three-dimensional image data using IMOD". In: *Journal of structural biology* 116.1 (1996), pp. 71–76.

[133] Sergio Cruz-León, Tomáš Majtner, Patrick Hoffmann, Jan P Kreysing, Maarten W Tuijtel, Stefan L Schaefer, Katharina Geißler, Martin Beck, Beata Turoňová, and Gerhard Hummer. "High-confidence 3D template matching for cryo-electron tomography". In: *Biophysical Journal* 123.3 (2024), 183a.

[134] Sheng Xu, Amnon Balanov, and Tamir Bendory. "Bayesian Perspective for Orientation Estimation in Cryo-EM and Cryo-ET". In: *bioRxiv* (2024), pp. 2024–12.

[135] Martin Beck, Johan A Malmström, Vinzenz Lange, Alexander Schmidt, Eric W Deutsch, and Ruedi Aebersold. "Visual proteomics of the human pathogen Leptospira interrogans". In: *Nature methods* 6.11 (2009), pp. 817–823.

[136] Noushin Hajarolasvadi, Harold Phelippeau, Robert Brandt, Pierre Nicolas Suau, Antonio Martinez-Sanchez, and Daniel Baum. "Deep orientation estimation of macromolecules in cryo-electron tomography". In: *BIO Web of Conferences*. Vol. 129. EDP Sciences. 2024, p. 10016.

[137] Qinwen Huang, Ye Zhou, Hsuan-Fu Liu, and Alberto Bartesaghi. "Accurate detection of proteins in cryo-electron tomograms from sparse labels". In: *European Conference on Computer Vision*. Springer. 2022, pp. 644–660.

[138] Thomas Hrabe, Yuxiang Chen, Stefan Pfeffer, Luis Kuhn Cuellar, Ann-Victoria Mangold, and Friedrich Förster. "PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis". In: *Journal of structural biology* 178.2 (2012), pp. 177–188.

[139] Marten L Chaillet, Gijs van der Schot, Ilja Gubins, Sander Roet, Remco C Veltkamp, and Friedrich Förster. "Extensive angular sampling enables the sensitive localization of macromolecules in electron tomograms". In: *International Journal of Molecular Sciences* 24.17 (2023), p. 13375.

[140] Valentin J Maurer, Marc Siggel, and Jan Kosinski. "PyTME (Python Template Matching Engine): A fast, flexible, and multi-purpose template matching library for cryogenic electron microscopy data". In: *SoftwareX* 25 (2024), p. 101636.

[141] William Wan, Sagar Khavnekar, and Jonathan Wagner. "STOPGAP: an open-source package for template matching, subtomogram alignment and classification". In: *Acta Crystallographica Section D: Structural Biology* 80.5 (2024).

[142] Chang Liu, Xiangrui Zeng, Ruogu Lin, Xiaodan Liang, Zachary Freyberg, Eric Xing, and Min Xu. "Deep learning based supervised semantic segmentation of electron cryo-subtomograms". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 1578–1582.

[143] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[144] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[145] Hsuan-Fu Liu, Ye Zhou, Qinwen Huang, Jonathan Piland, Weisheng Jin, Justin Mandel, Xiaochen Du, Jeffrey Martin, and Alberto Bartesaghi. "nextPYP: a comprehensive and scalable platform for characterizing protein variability in situ using single-particle cryo-electron tomography". In: *Nature Methods* 20.12 (2023), pp. 1909–1919.

[146] Dari Kimanius and Johannes Schwab. "Confronting heterogeneity in cryogenic electron microscopy data: Innovative strategies and future perspectives with data-driven methods". In: *Current Opinion in Structural Biology* 86 (2024), p. 102815.

[147] Haoran Li, Xingjian Li, Jiahua Shi, Huaming Chen, Bo Du, Daisuke Kihara, Johan Barthelemy, Jun Shen, and Min Xu. "Vox-UDA: Voxel-wise Unsupervised Domain Adaptation for Cryo-Electron Subtomogram Segmentation with Denoised Pseudo Labeling". In: *arXiv preprint arXiv:2406.18610* (2024).

[148] Sanket Rajan Gupte, Cathy Hou, Gong-Her Wu, Jesus G Galaz-Montoya, Wah Chiu, and Serena Yeung-Levy. "CryoViT: Efficient segmentation of cryogenic electron tomograms with vision foundation models". In: *bioRxiv* (2024), pp. 2024–06.

[149] J Lehtinen. "Noise2Noise: Learning Image Restoration without Clean Data". In: *arXiv preprint arXiv:1803.04189* (2018).

[150] Jose Inacio Costa-Filho, Liam Theveny, Marilina de Sautu, and Tom Kirchhausen. "CryoSamba: self-supervised deep volumetric denoising for cryo-electron tomography data". In: *Journal of Structural Biology* (2024), p. 108163.

[151] Eugene Palovcak, Daniel Asarnow, Melody G Campbell, Zanlin Yu, and Yifan Cheng. "Enhancing SNR and generating contrast for cryo-EM images with convolutional neural networks". In: *bioRxiv* (2020), pp. 2020–08.

[152] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. "Self-supervised learning: A succinct review". In: *Archives of Computational Methods in Engineering* 30.4 (2023), pp. 2761–2775.

[153] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. "Revisiting self-supervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1920–1929.

[154] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[155] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[156] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020).

[157] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 15750–15758.

[158] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[159] Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. "The ssl interplay: Augmentations, inductive bias, and generalization". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 3252–3298.

[160] Vitaliy Kinakh, Olga Taran, and Svyatoslav Voloshynovskiy. "ScatSimCLR: self-supervised contrastive learning with pretext task regularization for small-scale datasets". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1098–1106.

[161] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. "Dense siamese network for dense unsupervised learning". In: *European Conference on Computer Vision*. Springer. 2022, pp. 464–480.

[162] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[163] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[164] Serge Beucher. "The watershed transformation applied to image segmentation". In: *Scanning microscopy* 1992.6 (1992), p. 28.

[165] Steffen Wolf, Constantin Pape, Alberto Bailoni, Nasim Rahaman, Anna Kreshuk, Ullrich Kothe, and FredA Hamprecht. "The mutex watershed: efficient, parameter-free image partitioning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 546–562.

[166] Pavol Harar, Lukas Herrmann, Philipp Grohs, and David Haselbach. "Faket: Simulating cryo-electron tomograms with neural style transfer". In: *arXiv preprint arXiv:2304.02011* (2023).

[167] Zhidong Yang, Fa Zhang, and Renmin Han. "Self-supervised cryo-electron tomography volumetric image restoration from single noisy volume with sparsity constraint". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4056–4065.

[168] John A Schellman. "Flexibility of DNA". In: *Biopolymers: Original Research on Biomolecules* 13.1 (1974), pp. 217–226.

[169] Ilya Loshchilov and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts". In: *arXiv preprint arXiv:1608.03983* (2016).

[170] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[171] Leslie N Smith and Nicholay Topin. "Super-convergence: Very fast training of neural networks using large learning rates. arXiv". In: *arXiv preprint arXiv:1708.07120* 6 (2017).

[172] I Loshchilov. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[173] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[174] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[175] Andrii Iudin, Paul K Korir, José Salavert-Torres, Gerard J Kleywegt, and Ardan Patwardhan. "EMPIAR: a public archive for raw electron microscopy image data". In: *Nature methods* 13.5 (2016), pp. 387–388.

[176] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. "EMPIAR: the electron microscopy public image archive". In: *Nucleic Acids Research* 51.D1 (2023), pp. D1503–D1511.

[177] Utz Ermel, Anchi Cheng, Jun Xi Ni, Jessica Gadling, Manasa Venkatakrishnan, Kira Evans, Jeremy Asuncion, Andrew Sweet, Janeece Pourroy, Zun Shi Wang, et al. "A data portal for providing standardized annotations for cryo-electron tomography". In: *Nature Methods* (2024), pp. 1–3.

[178] Dimitry Tegunov, Liang Xue, Christian Dienemann, Patrick Cramer, and Julia Mahamid. "Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells". In: *Nature methods* 18.2 (2021), pp. 186–193.

[179] Sagar Khavnekar, Ron Kelley, Florent Waltz, Wojciech Wietrzynski, Xianjun Zhang, Martin Obr, Grigory Tagiltsev, Florian Beck, William Wan, John Briggs, et al. *Towards the Visual Proteomics of C. reinhardtii using High-throughput Collaborative in situ Cryo-ET*. 2023.