GraphConv: Geometric Deep Learning for Multiple Conformation Generation from Electron Density Images

Saurav Jayakumar

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Dong Si

Renzhi Cao

Jie Hou

Erika Parsons

Program Authorized to Offer Degree:
Computer Science & Software Engineering

University of Washington

Abstract

GraphConv: Geometric Deep Learning
for Multiple Conformation Generation
from Electron Density Images

Saurav Jayakumar

Chair of the Supervisory Committee:
Dong Si
Computing & Software Systems

In the field of cryo-electron microscopy (cryo-EM) structural analysis, the precise prediction of molecular conformations within datasets is essential. Despite strides made in deep learning methodologies, existing solutions often yield volumes of suboptimal quality. Addressing this critical limitation, our research introduces GraphConv, an innovative encoder model designed to embed particle images into a latent space, thereby substituting the conventional encoder utilized by CryoDRGN. This novel approach employs a Graph Neural Network (GNN) architecture featuring multiple GraphConv and Convolutional layers, aimed at capturing richer information from particle images and precisely reconstructing corresponding 3D volumes. Rigorous testing across two authentic datasets and three simulated datasets underscores the efficacy of our model, suggesting marked enhancements in reconstruction quality. Specifically, our findings reveal enhancements in resolution by up to 20% compared to CryoDRGN. By harnessing the power of GNNs, our methodology shows promise for significant advancements in the fidelity and accuracy of output volumes, thereby contributing to the ongoing refinement of cryo-EM structural analysis methodologies.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

PDB: file format used to store 3D model information.

MRC: file format used to store 3D map information.

SNR: signal-to-noise ratio.

CTF: measure of how much the phase shift and microscope aberrations have corrupted the image.

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to Dr. Dong Si for his invaluable guidance, unwavering support, and mentorship throughout this project. Dr. Si's expertise and encouragement have been instrumental in shaping the direction and success of this endeavor. My gratitude also goes to the members of my defense committee: Dr. Renzhi Cao, Dr. Jie Hou, and Dr. Erika Parsons. Their willingness to share their expertise and time has been a cornerstone of my progress, providing me with much-needed direction and support during critical phases of my work.

I am immensely thankful to my colleagues from the DGL study group, Esha Gavali and Chloe Ma, whose collaborative spirit, insightful discussions, and contributions have enriched this project significantly. Their dedication and camaraderie have made this journey both productive and enjoyable. I also express my appreciation to the entire DAIS group for their support and feedback, which have fostered an environment conducive to growth and learning. Their collective support and encouragement have been indispensable in navigating challenges and achieving milestones in this research endeavor.

Chapter 1

INTRODUCTION

1.1 Motivation

Molecules and their intricate complexes exhibit a dynamic behavior that is indispensable to their biological functions, encompassing a continuous distribution of assembly, disassembly, and conformational changes. Deciphering these nuanced alterations is crucial for various stakeholders, including researchers striving to unravel the intricacies of biological processes, pharmaceutical companies seeking to develop targeted therapies, and educators advancing structural biology knowledge. This understanding not only provides invaluable insights into molecule functionality but also aids in the design of therapeutic drugs, enriching our comprehension of structural biology.

In the realm of structural biology, single-particle cryo-electron microscopy, described in Figure 1.1, has emerged as the technique of choice for resolving complex structures. Its advantage lies in its independence from crystallization and its minimal sample requirement [20]. Moreover, advancements in instrumentation and software have significantly streamlined the process, expediting the generation of density maps with near-atomic precision and fostering a surge in Cryo-EM depositions.

However, conventional reconstruction methodologies heavily reliant on extensive classification and particle averaging tend to yield static representations. Such approaches also necessitate substantial prior knowledge about the molecule, which may not always be readily available. Moreover, these methods often produce blurred or obscured structures that fail to elucidate the subtle variations [10].

Multiple deep and non-deep learning methods have been proposed to resolve these issues. These methods utilize the EMPIAR datasets of molecular particle images to discern the

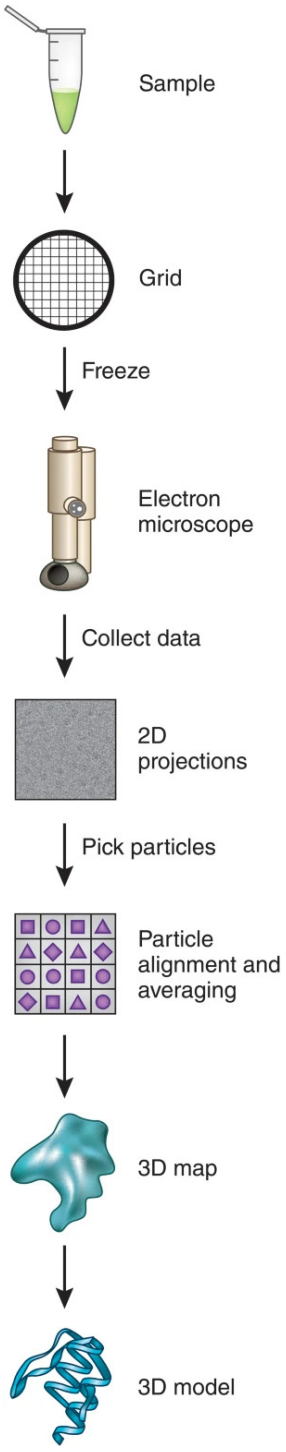


Figure 1.1: Single-Particle Cryo-EM process [7]. Molecule samples are flash-frozen in vitreous ice on special grids, protecting them from damage. Images of individual particles are then processed and aligned to create a detailed 3D model.

images belonging to the different conformations and model their dynamics. Solutions such as multi-body refinement in RELION [19] and 3D variability analysis in CryoSPARC [22] model the conformational changes using linear combinations of reaction coordinates and are effective at generating high-resolution strictures for rigid bodies. Meanwhile, deep-learning approaches use neural network models such as GANs and encoder-decoder models to fit the variations in the particle images to differentiate the conformations. Some examples of these methods are 3DFlex in CryoSPARC [23], CryoDRGN [34], and OPUS-DSD [15]. The former uses neural networks to fit the 3D displacement field of each particle while the latter two use neural networks to encode 2D images to latent spaces and then use those latent embeddings to generate 3D volumes. These solutions have proven that with minimal manual intervention and prior knowledge, the problem of heterogeneity can be addressed with significant success.

However, representing the electron density distribution of molecule particles in the form of images is inherently limiting due to the constriction of a pre-determined shape to the data. Recently, the burgeoning popularity of graph neural networks (GNNs) has offered a promising avenue, particularly for analyzing biological data represented in non-Euclidean domains like molecular structures and interactions [33]. Unlike conventional techniques, GNNs possess the capacity to capture latent information without rigid structural constraints, making them capable of navigating the intricate interconnections within molecular networks. Methods such as ModelAngelo [9] use GNNs to represent cryo-em maps using graphs based on their electron density information to generate protein structures and achieve significant improvements over prior solutions. This underscores the capabilities of graph neural network layers over traditional deep-learning layers to capture important information and relations for the task at hand.

1.2 Problem Statement

While CryoDRGN and OPUS-DSD stand out for their innovative use of an image encoder-volume decoder architecture based on the variational autoencoder (VAE) framework, we propose the exploration of alternative encoder models for several reasons:

- Different encoder architectures may offer unique advantages in capturing and representing complex structural variations inherent in cryo-EM images. By diversifying the range of encoder models available, researchers can tap into a broader spectrum of capabilities, potentially leading to improved reconstruction accuracy and efficiency.

- The inherent complexities of molecular structures demand adaptable and versatile encoder models. While CryoDRGN excels in certain aspects of heterogeneous molecule reconstruction, its latent space layout may not always yield interpretable features of energy landscapes, and structures generated from unoccupied regions may not align with physical reality [34]. Exploring alternative encoder models could shed light on novel approaches to mitigate these challenges, offering fresh perspectives on encoding structural heterogeneity and facilitating more nuanced interpretations of reconstructed volumes.

- The performance of cryo-EM reconstruction methods can be influenced by various factors, including dataset characteristics and imaging conditions. By experimenting with different encoder models, researchers can choose the best encoder based on the specific datasets and optimize performance under diverse conditions. This adaptability is essential for accommodating the wide-ranging complexities encountered in cryo-EM imaging, from variations in sample composition to fluctuations in imaging quality.

This thesis attempts to tackle the challenge of heterogeneity through the development of a GNN-based solution. By integrating GraphConv, a GNN model consisting of Graph Convolutional Layers, into the CryoDRGN framework, We aim to make use of the inherent expressiveness of graphs to improve the quality of conformation generation. We opted for CryoDRGN over OPUS-DSD due to the former's extensive suite of tools and programs within the framework, which streamlines pre-processing and analysis tasks.

CryoDRGN currently utilizes linear and residual layers as part of its encoder model to embed 2D particle images into a latent space. These latent space embeddings are then de-

coded using another model to generate 3D volumes corresponding to each embedding. While successful in identifying the different conformations in the input dataset, the quality of the generated volumes is at times poor with low resolutions and blurred structures [34]. Hence, improvements could be made to the encoder architecture to capture only the important information from the input images and discard the noise to generate better quality 3D volumes from their embeddings.

The primary goal of this research is to explore the impact of using graph neural networks to overcome the limitations of traditional modes. Through comprehensive evaluations of diverse real and simulated heterogeneous molecular datasets, employing various quality metrics, we seek to validate the efficacy of our proposed encoder model vis-à-vis CryoDRGN's existing encoder. This innovative approach could push the field of dynamic molecular structure resolution forward, facilitating a deeper understanding of these fundamental molecules and their intricate behaviors.

1.3 Structure of Thesis

In the upcoming chapters, we will discuss the research efforts in further detail. Starting with Chapter 2, we will dive deeper into the related works, especially CryoDRGN, to examine the current state-of-the-art solution, the tools used to quantify the results, and graph neural networks. In Chapter 3, we explore our methodology in-depth. This includes the architecture of the proposed encoder model and the datasets and metrics used for comparison. Chapter 4 will present the results comparing CryoDRGN's encoder model against the proposed model and discuss the findings. This chapter serves to highlight the improvements made over the existing solution. Chapter 5 will summarize the key findings and limitations of the work with potential enhancements to improve the results further.

Chapter 2

RELATED WORKS

In this chapter, we delve into the foundational works and software tools that have been instrumental in shaping our methodology and the subsequent results. We will place particular emphasis on CryoDRGN, examining its architecture and underlying principles as described by [34]. Additionally, we will explore the concept of Graph Neural Networks and their relevance to our research. Furthermore, we will provide an overview of the tools that have facilitated the simulation of datasets and the collection of results, including CryoSRPNT, EMAN2, ChimeraX, and Phenix, elucidating their roles in our investigations.

2.1 CryoDRGN

CryoDRGN [34] is a protein structure reconstruction method for cryo-electron microscopy (cryo-EM) images that addresses challenges posed by the heterogeneity of protein conformations. It employs an image encoder-volume decoder architecture, based on the variational autoencoder (VAE) framework (see Figure 2.1). Unlike conventional VAEs, it disentangles orientation and heterogeneity information within the latent space of the VAE to enhance accuracy. The generative decoder aims to produce electron density maps in 3D space using spatial coordinates as inputs. In contrast to other methods, it models the entire 3D space as a whole, capturing density information for any location.

The method's core components consist of an encoder for estimating a latent variable and a decoder for generating density maps from latent variables and positional encodings. The decoder optimizes output by maximizing the likelihood of generating electron density functions from input images, using latent variables and image alignment techniques within the VAE framework. It employs fixed sets of sine and cosine waves to encode pixel coordinates,

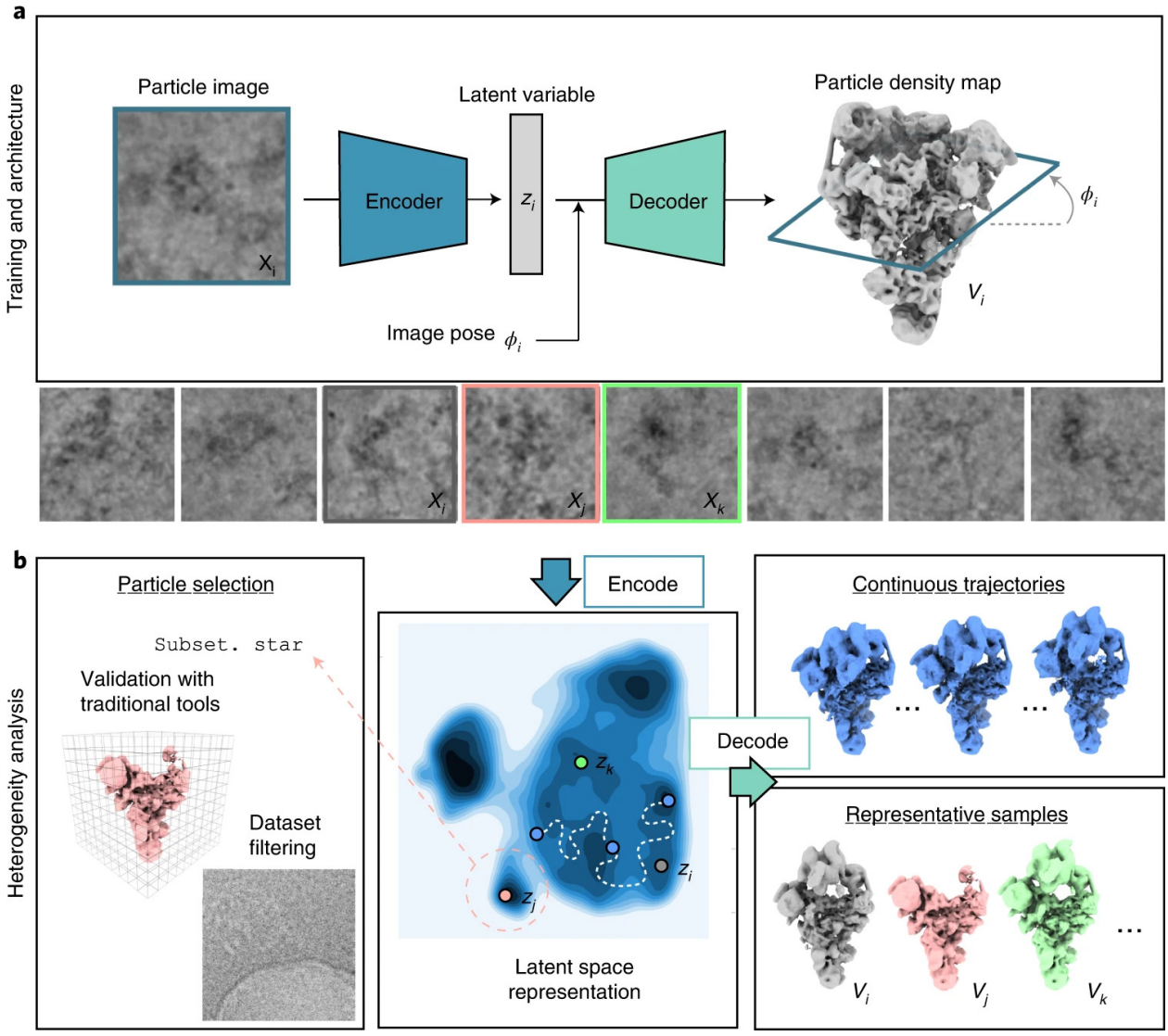


Figure 2.1: CryoDRGN Framework [34]. a, The CryoDRGN model consists of an encoder and decoder to convert particle images to an embedding and then convert those embeddings to 3D volumes. b, Once trained, the embeddings of the entire dataset can be traversed to generate the different conformations and also used to generate discrete volumes for validation using traditional tools.

adapting the encoding pattern for better results with noisy datasets.

For pose estimation, CryoDRGN uses a global search approach combining rotations and translations. It efficiently refines poses using grid-based discretization and branch and bound pruning. The model's neural networks are designed with a customizable number of hidden layers, using ReLU activations. It is capable of unsupervised homogeneous and heterogeneous reconstructions, even automatically estimating poses and conformation ordering without requiring prior knowledge of the number of classes.

Their architecture is trained on downsampled images for efficient initial assessment and further fine-tuning on high-resolution reconstructions. Trajectories in latent space can be generated, providing insight into structural changes, and representative structures are created from regions with high particle density. CryoDRGN allows the filtering of impurities in particle stacks based on latent space representations and provides a tool for interactive exploration and interpretation of resulting structural ensembles.

It is important to note however that while CryoDRGN excels in reproducing the distribution of structures, its latent space layout may not necessarily produce interpretable features of energy landscapes; furthermore, the structures generated from unoccupied regions may not correspond to true physical structures. It should also be noted that poor-quality particles might impact the method's performance.

CryoDRGN has proven to be highly effective in analyzing complex protein ensembles by capturing both continuous and discrete heterogeneity from cryo-electron microscopy (cryo-EM) images. This approach addresses the challenges posed by structural variability in proteins, leading to precise and comprehensive reconstructions of protein structures. By utilizing the strengths of CryoDRGN's framework, we can optimize data processing and decoding while adapting the image encoder to enhance the overall methodology. This allows for a more robust and nuanced exploration of cryo-EM data, leading to deeper insights into protein structures and functions.

2.1.1 Pre-Processing

Prior to model training, the data undergoes a series of preprocessing steps to ensure it is suitable for analysis. First, particle images are scaled down to dimensions optimal for training. Then, consensus reconstructions are generated using tools such as cryoSPARC, yielding high-quality initial models for subsequent training. A crucial step involves applying positional encoding functions to Cartesian coordinates before feeding them into the neural networks. These functions comprise sinusoids with wavelengths adapted to the image size, a technique that has proven beneficial in enhancing performance, especially when dealing with noisy data.

2.1.2 Encoder

The encoder component in CryoDRGN, a neural network-based reconstruction method, plays a pivotal role in the unsupervised reconstruction of 3D volumes from unlabeled 2D images. Utilizing a variational autoencoder (VAE) framework, their image encoder-volume decoder architecture models protein structural heterogeneity in the latent variable. Unlike standard VAEs, which entangle all sources of image heterogeneity in the latent variable, their architecture distinguishes intrinsic volume heterogeneity from extrinsic orientation during imaging. This approach allows for the explicit modeling of the forward image formation process, relating 2D views to 3D volumes, and employs variational inference for unconstrained latent variables alongside a global search over $\mathrm{SO}(3) \times \mathrm{R}^2$ for image pose estimation.

During training, for each input image, the encoder produces statistics describing the latent variable, which is then broadcast to all pixels. A maximum likelihood pose, comprising rotation and translation, is inferred for each image using a branch and bound global optimization procedure. The decoder subsequently reconstructs the image pixel by pixel based on the positional encoding of the coordinates and the latent variable. The optimization objective is the variational lower bound of the model evidence, facilitating the learning process by comparing multiple 2D slices from the imaging dataset while denoising to prevent over-

fitting to noise. This approach captures structural variation within a single imaging dataset, with separate networks trained per experimental dataset to accommodate diverse protein or biomolecular complexes.

The image encoder is implemented using a neural network architecture tailored to learn a pose-invariant representation of molecule structural heterogeneity. This architecture utilizes a combination of linear layers and residual connections to capture the complex relationships between input images and latent variables. The encoder architecture consists of a stack of residual linear units organized into an MLP (Multi-Layer Perceptron). The MLP comprises several layers of residual connections interleaved with activation functions, such as ReLU (Rectified Linear Unit), which introduce non-linearity into the model. These residual connections enable the model to learn more complex representations by preserving information from earlier layers while allowing for the exploration of deeper network architectures. The output layer of the encoder MLP produces statistics parameterizing an approximate posterior distribution over the latent variable. This distribution captures the variability in protein structure present in the input image. By using a variational autoencoder framework, the encoder facilitates unsupervised learning by capturing the underlying structure of the data without the need for labeled training examples.

2.1.3 Decoder

The decoder component of CryoDRGN, a neural network-based reconstruction method for unsupervised 3D volume reconstruction from 2D images, is a crucial part of the architecture. Building upon the variational autoencoder (VAE) framework, their decoder is tasked with generating 3D volumes from latent variables representing protein structural heterogeneity. Unlike traditional VAEs that encapsulate all image heterogeneity in the latent space, their decoder disentangles intrinsic volume heterogeneity from extrinsic orientation, allowing for a more accurate representation of the underlying structural variations.

In their generative model, the decoder takes Cartesian coordinates and a continuous latent variable as inputs and outputs distribution parameters for a Gaussian distribution

over the Fourier domain of the reconstructed volume. Unlike conventional deconvolutional decoders, which operate on voxel-level distributions, their decoder functions over Cartesian coordinates, explicitly representing each pixel's location in 3D Fourier space. This approach follows the topological constraints between 2D views in 3D, enforced via the Fourier slice theorem [2], thereby enhancing reconstruction fidelity.

During training, each input image corresponds to an oriented central slice of the 3D volume in the Fourier domain. The decoder uses this information alongside latent pose variables and the unconstrained latent variable to reconstruct the image pixel by pixel. Instead of direct coordinates, the decoder uses a positional encoding with sine and cosine waves of varying frequencies. This encoding, defined within a sphere of radius 0.5, covers a range from wavelength 1 to the Nyquist limit, and excludes the highest 10% of frequencies for better results with noisy datasets

The pose inference process involves a global search over rotation and translation parameters to determine the maximum likelihood pose for each image. Techniques like grid discretization and frequency marching enhance the efficiency of this search, with a focus on low-frequency components due to their dominance in pose estimation. Once the pose is inferred, the decoder reconstructs the image using the positional encoding and the latent variable, with optional adjustments for phase shift and microscope characteristics. Additionally, it is worth noting that the decoder in CryoDRGN employs the same architecture as the encoder, ensuring consistency and symmetry between the encoding and decoding processes. This shared architecture facilitates seamless information flow and feature extraction throughout the reconstruction pipeline, ultimately contributing to the high-quality reconstruction of 3D volumes from unlabeled 2D images.

2.1.4 Loss Functions

The loss function utilized by CryoDRGN is designed to optimize the reconstruction accuracy of 3D volumes from 2D projection images. This function is derived from the relationship between 2D projection images and volumes in Fourier space, where 2D central slices correspond

to oriented 3D coordinates. Specifically, the loss function involves several steps:

1. **Rotation and Reconstruction**: Initially, oriented 3D coordinates for an image $\hat{X}_i$ are obtained by rotating a lattice spanning $(-0.5, 0.5)^2$ originally on the x-y plane by R_i, representing the orientation of the volume during imaging. The image encoder produces $\mu_{z|\hat{X}_i}$ and $\Sigma_{z|\hat{X}_i}$. A sampled value of the latent variable $z_i \sim \mathcal{N}(\mu_{z|\hat{X}_i}, \Sigma_{z|\hat{X}_i})$ is broadcast to all pixels. Given z_i and the current decoder, a Bayesian Neural Network (BNB) orientational search identifies the maximum likelihood rotation R_i and translation t_i for $\hat{X}_i$. The decoder p_θ then reconstructs the image pixel by pixel given the positional encoding of $R_i^\top c_0$ and z_i.

2. **Translation and Multiplication**: The reconstructed image is then translated by the image's in-plane shift t_i and multiplied by the contrast transfer function (CTF) $\hat{g}_i$ before being compared to the input image.

3. **Variational Lower Bound**: Following the standard Variational Autoencoder (VAE) framework, the optimization objective is a variational lower bound of the model evidence, represented as:

$$L(\hat{X}_i; \xi, \theta) = E_{q_\xi(z|\hat{X}_i)} \left[\log p_\theta(\hat{X}_i|z) \right] - KL(q_\xi(z|\hat{X}_i) \parallel p(z))$$

   The first term represents the reconstruction error, which is estimated with one Monte Carlo sample. The second term is the Kullback-Leibler (KL) divergence term, serving as a regularization term on the latent representation. β is an additional hyperparameter, set by default to $\frac{1}{|z|}$.

Variable Definitions:

- $\hat{X}_i$: Input 2D projection image.

- $\mu_{z|\hat{X}_i}$, $\Sigma_{z|\hat{X}_i}$: Mean and covariance of the latent variable z_i given the image $\hat{X}_i$.

- z_i: Latent variable sampled from the distribution $\mathcal{N}(\mu_{z|\hat{X}_i}, \Sigma_{z|\hat{X}_i})$.
- R_i: Rotation matrix representing the orientation of the volume.
- t_i: In-plane translation shift of the image.
- $\hat{g}_i$: Contrast transfer function (CTF) of the microscope.
- p_θ: Decoder function parameterized by θ.
- $q_\xi(z|\hat{X}_i)$: Variational distribution approximating the posterior $p(z|\hat{X}_i)$, parameterized by ξ.
- $KL(q_\xi(z|\hat{X}_i) \parallel p(z))$: Kullback-Leibler divergence between the approximate posterior $q_\xi(z|\hat{X}_i)$ and the prior $p(z)$.
- $L(\hat{X}_i; \xi, \theta)$: Variational lower bound of the model evidence.

Through training on numerous 2D slices with diverse orientations, CryoDRGN aims to learn the 3D volume by incorporating feedback from the 2D views.

2.2 Graph Neural Networks

Graphs are a versatile tool for representing entities and their interconnections across various domains such as social networks, biology, and economics. These networks model intricate relationships and interactions, providing insights into complex phenomena. However, challenges arise in dealing with graphs due to their complexity, large size, lack of spatial locality, and dynamic nature, which makes processing them more challenging compared to traditional data types.

Graph Neural Networks (GNNs) is a specialized solution for processing graph-structured data [35]. The layered architecture of GNNs involves message transformation and aggregation, addressing limitations faced by earlier methods. By using deep-learning techniques,

GNNs can incorporate features, adapt to evolving graphs, and capture structural similarities. The design of GNNs encompasses considerations such as input graph setup, architecture design, node embedding, and prediction head design, ensuring effective utilization across different applications.

GNNs are useful for a wide range of task categories, including node-level, edge-level, and sub-graph-level tasks. Techniques including neighborhood sampling and simplified Graph Convolutional Networks (GCN) are employed to handle large graphs efficiently. Neighborhood sampling operates on entire computation graphs within mini-batches, enhancing computational performance. Meanwhile, simplified GCN streamlines training by eliminating non-linear activations and considering matrix formulations.

One notable advantage of GNNs is their ability to capture local neighborhood structures, enabling nodes to gather information from varying ranges. However, the phenomenon of over-smoothing presents a challenge; specifically, deep GNNs can lead to the convergence of node embeddings, resulting in the loss of valuable information. Strategies to mitigate over-smoothing include designing more expressive layers, incorporating pre/post-processing layers, and integrating skip connections in order to preserve essential information and enhance the overall performance of GNNs in graph-related tasks.

The integration of GNNs in various studies across the field of biology has seen remarkable advances, particularly in molecular dynamics simulations, protein informatics, and functional residue identification. Li et al. [13] introduce a method that significantly improves traditional molecular dynamics simulations by using GNNs. By learning interactions from existing data, GNNs enable the direct prediction of atomic forces, reducing computational costs and enabling simulations of larger and more complex molecules. Additionally, GNNs focus on key information within molecular structures, making them more scalable and efficient, thereby paving the way for a deeper understanding of molecular behavior.

In Jamali et al.'s research [9] describing ModelAngelo, GNNs within the model automate atomic model-building and enhance protein identification in cryo-electron microscopy (cryo-EM) maps. By integrating information from cryo-EM maps with protein sequence

and structure data, ModelAngelo constructs atomic models with human-expert comparable quality and improves protein identification in complex biological structures. The incorporation of GNNs streamlines the model-building process, leading to significant time savings and potentially accelerating scientific discoveries in structural biology.

Tsuchiya et al. [28] explore the applications of neural networks, including GNNs, in protein informatics. Their model for predicting protein-compound interactions showcases the advantages of GNNs in capturing structural information and achieving robust learning from imbalanced datasets. It's incorporation of GNNs enhances prediction accuracy and enables the identification of important regions within proteins, thereby advancing drug discovery efforts and contributing to a deeper understanding of protein-ligand interactions.

The introduction of Graph Neural Network Force Field (GNNFF) in Li et al.'s work [13] represents a significant improvement in predicting atomic forces in molecular dynamics simulations. GNNFF demonstrates superior accuracy, efficiency, scalability, and versatility compared to existing models, demonstrating its potential to revolutionize the field of molecular dynamics simulations and facilitate atomistic-level studies of various molecular systems.

Lastly, Chiang et al. [5] introduce a novel approach for identifying functional residues in proteins using GNNs. By encoding dynamic information from molecular dynamics simulations, GNNs enhance predictive performance, scalability, interpretability, and flexibility in functional residue identification, thereby advancing our understanding of protein structure-function relationships and accelerating drug discovery efforts.

The integration of GNNs in biology research offers unprecedented opportunities for accelerating scientific discoveries, improving predictive performance, and gaining deeper insights into complex biological processes. These advances underscore the transformative potential of GNNs in advancing various fields within biology and related disciplines.

2.3 CryoSRPNT

CryoSRPNT is a software tool designed to simulate realistic cryo-EM particle datasets from

known 3D volumes, aiding in the development and benchmarking of cryo-EM reconstruction and analysis tools. Its `project3d` script generates noiseless 2D projections of a given 3D volume based on specified pose sampling schemes. Subsequently, the `acn` script introduces noise and CTF corruption to these projections, simulating the actual image formation process in cryo-EM. Using a PyTorch/CUDA framework for GPU acceleration, CryoSRPNT enables efficient simulation of large datasets, making it a valuable resource for advancing cryo-EM research and development.

2.4 EMAN2

EMAN2 is a comprehensive software package for processing and analyzing cryo-electron microscopy (cryo-EM) data [27]. It offers a wide array of tools, including utilities like `e2pdb2mrc` for converting protein structures from the PDB format into MRC density maps, and `e2proc2d` for performing general image processing tasks on cryo-EM images, such as CTF correction and image manipulation. EMAN2's extensive capabilities make it a valuable resource for researchers in the cryo-EM field, facilitating efficient and accurate structure determination from raw experimental data.

2.5 ChimeraX

ChimeraX is a powerful and versatile software for visualizing and analyzing molecular structures, favored by researchers, educators, and developers alike due to its user-friendly interface and extensive capabilities [21]. In the context of cryo-EM, ChimeraX offers specialized tools for fitting atomic models into density maps and assessing the quality of the fit. Users can perform both manual and automated fitting, adjusting the model's position and conformation to optimize its agreement with the experimental map. The software also provides features for map subtraction, allowing for the visualization of differences between the experimental map and the map calculated from the fitted model, aiding in the identification of potential errors or discrepancies in the fitting process.

2.6 Phenix

Phenix [14] is a comprehensive software suite designed to utilize reduced data from various diffraction techniques such as X-ray, electron, neutron, or cryo-EM, facilitating the determination of macromolecular structures. Tailored to the unique properties of each method, Phenix provides specialized tools aimed at extracting structural insights efficiently. A key strength of Phenix is its emphasis on automation, sparing users from repetitive, time-consuming, and error-prone tasks. Furthermore, its intuitive interface ensures accessibility for novices while retaining flexibility for experts. Continual enhancements and the development of new tools further refine the structure-solution workflow. Notably, recent efforts by Liebschner et al. [14] have yielded a suite of programs specifically catering to the analysis of cryo-EM maps and models, addressing the evolving needs of the cryo-EM community.

Specifically, we utilize the following functions from Phenix for the result collection following our methodology.

2.6.1 Mtriage

In cryo-EM experiments, factors like structural heterogeneity and sample movement can affect reconstruction quality. While resolution indicates map reliability, its assessment differs from crystallography. Mtriage [1], a new tool in Phenix, offers diverse approaches to estimate resolution and provides map statistics. Given its impact on decision-making, obtaining a reliable resolution estimate is crucial for subsequent processing.

2.6.2 Model-Map correlation and offset log

The Model-Map correlation and offset log function (`get_cc_mtz_pdb`) [14] is a tool designed to adjust the origin of a PDB file using space-group symmetry, aligning it with a supplied map to calculate the correlation between the model and map at each residue level. This tool offers various parameters for customizing correlation calculations. By default, it calculates model density using PDB file parameters, and users can choose additional options like scaling

and specifying reflections from the mtz file. The output includes an offset PDB file matched to the map's origin. Example usage involves providing map coefficients and PDB coordinates via the command line, with options to specify column names if needed.

Chapter 3

METHODOLOGY

This chapter delineates our approach to advancing the field of dynamic molecular structure prediction by blending an established method with a novel encoding model. Specifically, we explore the impact of using Graph Neural Networks to address the problem. In the upcoming sections, we discuss the steps we followed to develop the proposed encoder model architecture along with the datasets and metrics used for evaluation.

3.1 Model Architecture

In the pursuit of developing a Graph Neural Network (GNN) framework capable of identifying distinct molecular conformations within a dataset, a thorough evaluation of various architectures was conducted. In the initial phases of design, a comprehensive framework was devised to accept particle images as inputs and subsequently encode them into a latent space. The intent was for this latent space to exhibit clustering, facilitating the identification of diverse conformations based on the resultant clusters. However, despite exploring different architectures such as Graph U-Nets [8] and custom models employing iterative reduction of nodes and edges through multiple Graph Convolutional Network (GCN) layers, inspired by Ma et al. [16], the outcome consistently converged to a single point, rendering only one cluster discernible across all datasets. This demonstrated that using only particle images for the encoding process may not be enough. Additional data, such as pose information and diverse pre-processing techniques, may be necessary to reduce noise and standardize inputs.

Following this finding, CryoDRGN [34] emerged as a viable solution to the problem. It has a well-established framework proficient in data pre-processing and analysis with ongoing development of new features. These advantages dictated our choice of CryoDRGN as the

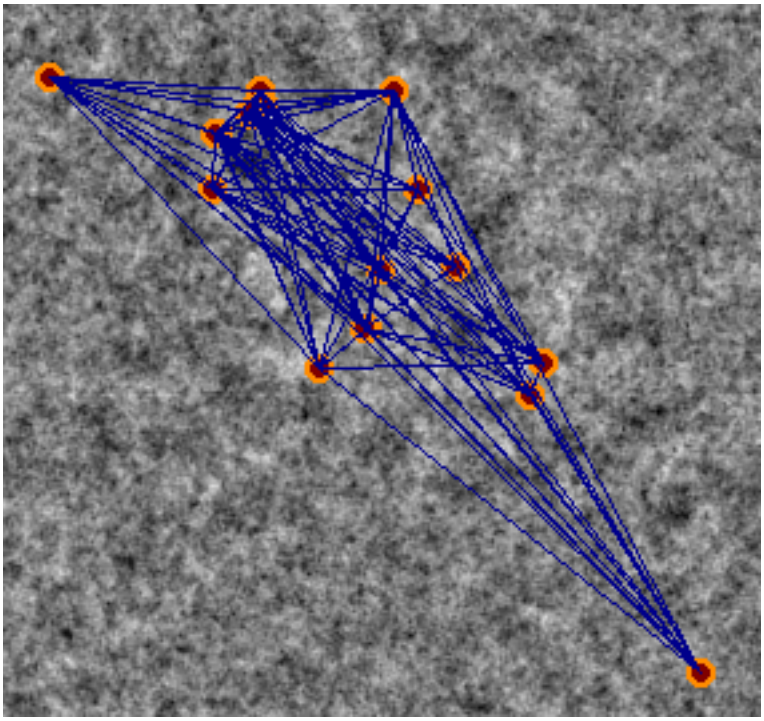


Figure 3.1: Sample electron density particle image with a few high-density points selected as nodes (maroon/orange) and edges connecting them (blue)

framework for this research.

Even with the adoption of CryoDRGN, employing off-the-shelf GNN layers like GCN-Conv [11], GATConv [29], GINConv [31], alongside pooling layers such as TopKPooling [4], SAGPooling [12], ASAPooling [24], consistently yielded a singular embedding for all input particle images. Extensive experimentation with various layers and tracking of value changes throughout the layers elucidated that the usage of pooling layers led to uniform embeddings across all inputs. To combine the structure analysis capabilities of graphs with the necessity to reduce nodes to an embedding, a hybrid approach was devised. This approach integrates GraphConv layers for initial data processing, followed by 1D Convolutional layers to condense the output from graph layers into an embedding vector.

The GNN encoder model (see Figure 3.2) is built to process graph-structured data, using

graph convolutional layers for feature extraction and dimensionality reduction. The following steps detail the conversion of input data images to graphs (see Figure 3.1):

1. A graph construction function iterates over each image in the input batch, filtering and selecting the nodes to form the graph. Points with higher electron density are assumed to carry the most information regarding changes in conformation. Hence we used a percentile threshold of 0.4 to select the top density points as nodes.

2. A graph is constructed by mapping edges between selected nodes using a k-nearest neighbors (k-NN) approach to connect each node to 20 of its nearest neighbors.

3. The constructed graphs are converted into a batch of graphs, corresponding to the input batch, using the PyTorch Geometric Batch class to enable efficient processing and parallelization.

Within the encoder module, multiple layers of graph convolution operations are applied to the input data, progressively extracting hierarchical representations of node features. These layers are each followed by a rectified linear unit (ReLU) activation function to introduce non-linearity and enhance model expressiveness. Further, the module incorporates 1D convolutional layers (Conv1d) to further refine extracted features and reduce dimensionality. These convolutional layers operate on the node feature matrix obtained from the graph convolutional layers, facilitating the extraction of high-level representations across different spatial dimensions. They also serve to reduce the data to the latent space embedding dimension. The command line arguments determine the number of layers and the latent space dimensions passed to the program.

Following feature extraction, the output is flattened into a one-dimensional vector using the Flatten module and passed to 2 fully connected layers (Linear) separately. The outputs of these Linear layers are the parameters of the latent variable distribution, specifically the mean (mu) and the standard deviation (sigma) used for Gaussian sampling during training and inference.

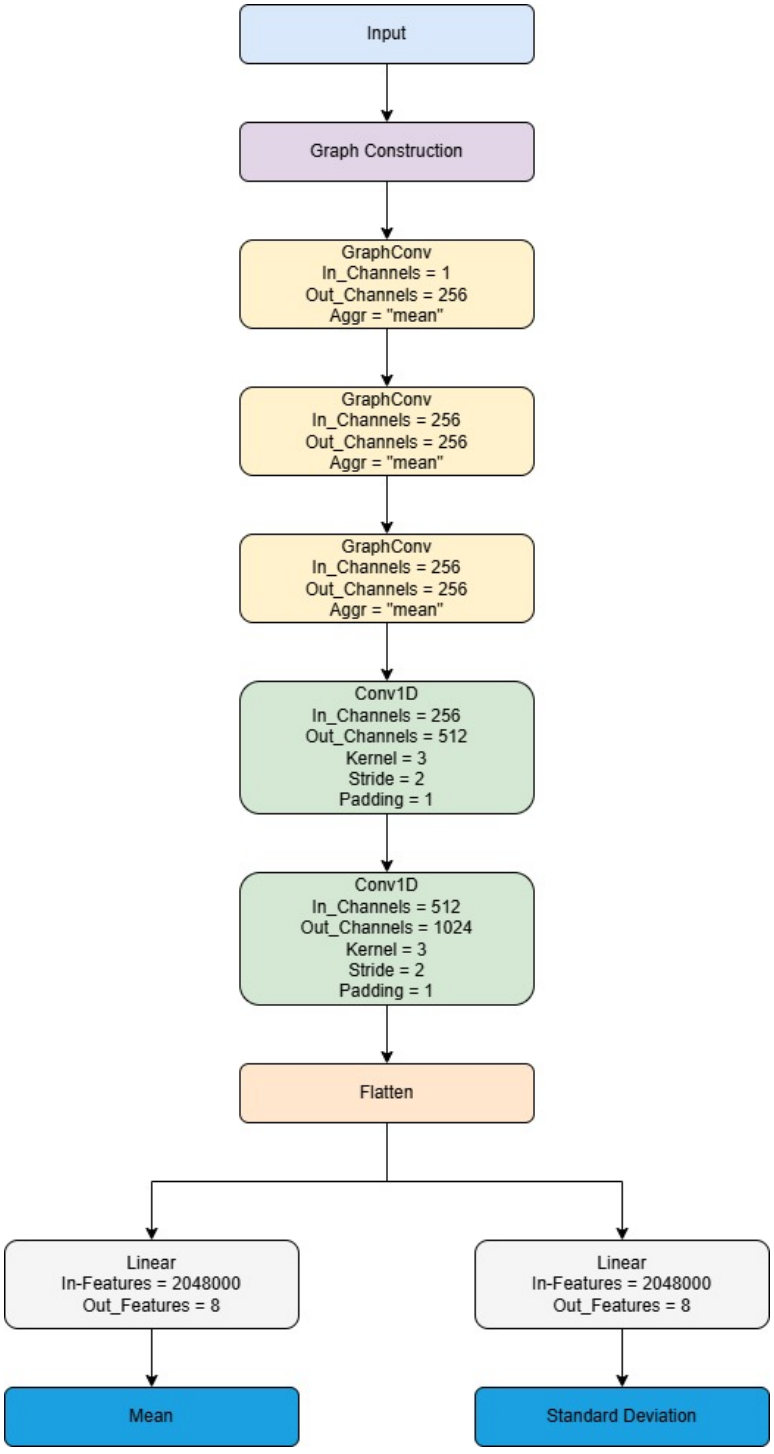


Figure 3.2: GraphConv Architecture. Assuming an input image size of 256 x 256, the data is used to construct a graph which is then fed to a series of GraphConv layers and a series of Conv1D layers to generate the mean and standard deviation for the Gaussian latent space.

3.2 Datasets

In our testing protocol, we employed 3 simulated and 2 real datasets to assess the efficacy of our method rigorously. This approach enabled a comprehensive validation utilizing simulated datasets for which ground truth conformations were available and ensured the testing of our model's performance with real datasets.

3.2.1 Simulated Datasets

Simulated datasets were built by following the below steps derived from the EMAN2 simulation steps and the CryoSRPNT repository:

1. Downloaded the PDB files associated with the conformations of the macromolecule being simulated.

2. Used the `e2pdb2mrc` program from EMAN2 [27] to convert the PDB files to MRC files. The MRC files were centered and generated with an angstrom/pixel value of 1.5, resolution of 4.5, and box size of 256.

3. The generated 3D volumes were then projected to a 2D plane using the `project3d` program from the CryoSRPNT repository. The angstrom/pixel value remained the same as the previous step with a 20-pixel image translation limit. The projections were sampled randomly from SO3 and the number of projections varied in the range of 10,000 to 15,000 for each conformation to simulate real dataset conditions. This step also produced the pose information files required for training.

4. These projection mrcs and pose files were combined using the `combine_particle_stacks` program published in this project's repository. A label file was also generated to map each image in the combined particle stack to its conformation, allowing for validation after training.

5. The CryoSRPNT repository provides the `acn` program to simulate ctf-corruption in input projection images but based on our testing, the results were incorrect. Hence, we used this program only to generate a ctf file for use in training whereas the actual ctf-corruption is performed by an EMAN2 program [27]. For the parameters of ctf-corruption, we used most of the values suggested by Zhong et al. [34]. Specifically, we set std1 to 0 for no structural noise, shot noise SNR to 0.1, angstrom/pixel value to 1.5, invert image data sign to True, defocus values to 15000 Å, voltage to 300 kV, spherical aberration to 2.0 mm, amplitude contrast ratio to 0.1, b-factor to 50 $Å^2$, and astigmatism to False.

6. Once the ctf file was generated, we used the `e2proc2d` program from EMAN2 [27] to perform the actual ctf-corruption. The `math.simulatectf` process was used to perform the simulation using the parameters mentioned in the previous step. This required the amplitude contrast ratio to be mentioned in percentage and the defocus value to be passed in µm. We also set the phaseflip value to 1 to perform CTF phase flipping.

By adhering to these steps, we simulated the following datasets:

RNA Polymerase I

This dataset contains the 6RQH, 6RQL, 6RRD, 6RUI, 6RUO, and 6RWE PDB files published by Sadien et al. [26]. These structures encompass significant variations in domain orientations, secondary structure elements, and active site configurations. For instance, domain movements between 6RQH and 6RQL highlight substantial shifts in the molecule's overall architecture, while changes in the active site loop between 6RUI and 6RUO demonstrate different functional states of the molecule. Additionally, the alteration in key catalytic residues observed in 6RWE compared to other structures underscores the dynamic nature of the molecule's functional mechanisms. This dataset evaluated the proposed solution's ability to detect diverse functional states (initiation, elongation, termination) and subunit interactions across different conformations.

SARS-CoV-2 Spike Protein

This dataset was built using the 8VKL and 8VKM PDB files published by Mannar et al. [17]. The protein undergoes significant conformational changes during viral fusion with host cells. These structures capture critical states along the transition from pre-fusion to post-fusion conformations. The pre-fusion conformation (8VKL) exhibits a closed state with the receptor-binding domains (RBDs) tucked in, while the post-fusion conformation (8VKM) reveals a dramatic rearrangement, with the RBDs displaced and the fusion peptide exposed. This dataset is ideal for testing our solution for large-scale structural changes in proteins. It evaluated the solution's ability to distinguish between major "open" and "closed" states, and potentially identify intermediate conformations while considering the impact of glycans and mutations.

Clostridium difficile Binary Toxin Translocase (CDTb)

For this dataset, we used the 6UWR and 6UWT PDB files published by Xu et al. [32]. These structures capture two distinct states: a symmetric tetradecamer (6UWT) and an asymmetric tetradecamer (6UWR). The transition between these conformations involves substantial rearrangement of the individual subunits, including significant changes in the relative orientations and interactions of the domains within each subunit. This conformational plasticity is crucial for CDTb's function in delivering toxins into host cells, as it facilitates pore formation in the cell membrane. These structures evaluate our solution's performance for multi-subunit protein assemblies. It evaluated the model's ability to capture conformational changes within the membrane environment, particularly those associated with toxin binding and translocation.

3.2.2 Real Datasets

The real cryo-electron microscopy (cryo-EM) datasets were procured from the Electron Microscopy Public Image Archive (EMPIAR). The ctf and pose files to be used for training

were gathered from the CryoDRGN-EMPIAR repository. We also downloaded the output files published by CryoDRGN from their database. These files required an adjustment to correct an anomaly in the MRC header data, which was accomplished using the `updata_header_from_data` command provided by the mrcfile library in Python [3].

EMPIAR 10028

Published by Wong et al. [30], this dataset contains the cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. It presents a unique challenge for our program. Ribosomes are large, complex macromolecular machines with inherent flexibility essential for their function. This dataset allowed us to test our program's ability to detect subtle conformational changes within the ribosome complex induced by drug binding.

Specifically, we evaluated our program's capacity to identify specific changes in the arrangement of ribosomal RNA and proteins upon emetine binding. Additionally, we can assess our program's sensitivity in detecting any potential heterogeneity within the dataset, as the ribosome may adopt multiple conformations even within a single state.

EMPIAR 10076

Released by Davis et al. [6], this dataset features the human mitochondrial ribosome in different functional states. It provided a valuable opportunity to evaluate our program's ability to analyze conformational heterogeneity. Mitochondrial ribosomes, distinct from their cytoplasmic counterparts, exhibit unique structural features and dynamics. This dataset enabled us to test our program's capacity to distinguish between various functional states of the ribosome, such as the initiation, elongation, and termination phases of protein synthesis.

Moreover, we assessed our program's ability to identify potential conformational changes induced by the presence of different tRNA and mRNA molecules within the ribosome complex. This dataset also includes structures with bound antibiotics, that allowed us to evaluate

our program's capacity to detect drug-induced conformational changes and potentially identify allosteric effects.

The major and minor class label files for this dataset were downloaded from the CryoDRGN database and the reference maps were collected from EMDB. To ensure consistency in the comparison, the reference maps were resampled using the `vop resample #ID1 onGrid #ID2` command. This adjustment allowed the reference maps to match the box size and voxel size of the generated maps. In some cases, the maps needed to be flipped on the z-axis using the `volume flip #ID axis z` command. Alignment was also necessary and was performed in ChimeraX [18]. The maps were aligned manually first, followed by fine-tuning with the `Fit` function.

3.3 Training

The following steps were performed for each dataset using the proposed model. We also ran CryoDRGN using the below steps on the simulated datasets to generate outputs for comparison.

1. The training was conducted using the `train_vae` command with the default hyperparameters recommended by the CryoDRGN team for 50 epochs.

2. Once the training finished, we verified that the model had converged using the `analyze_convergence` function available in the CryoDRGN framework by following their tutorial [10].

3. We then used CryoDRGN's analyze command which provided a Jupyter Notebook to generate the final conformations. All hyperparameters used for the analyze command were defaulted except for the Angstrom/pixel value. For the simulated datasets, this was set to 1.5 whereas for the real datasets, this was set to the value used by CryoDRGN's output mrc files available in their database.

3.4 Evaluation Criteria

After training the model on various datasets and generating conformations linked to the provided labels, the next step involved comparing the quality of these reconstructions by our model with those from CryoDRGN.

3.4.1 Unmasked Resolution

Unmasked resolution in cryo-EM maps is a fundamental indicator of the level of detail present in the reconstruction. When comparing our proposed solution to CryoDRGN, a higher unmasked resolution in our generated maps would indicate a superior ability to capture fine structural details and potentially reveal subtle conformational differences that CryoDRGN might miss. This is crucial for accurately identifying and characterizing the different conformations present in a dataset. This metric was calculated using the mtriage function available in Phenix.

3.4.2 Percentage of Residues not found in Acceptable Density

The percentage of residues not found in acceptable density within the reference models is a critical metric for evaluating the accuracy of our solution in capturing the conformational heterogeneity within a dataset. This metric is calculated by comparing the reference atomic models with the density in the maps generated by our solution and CryoDRGN. A lower percentage in our solution compared to CryoDRGN would suggest that the reference models better fit the density in our maps and strengthen our confidence in the ability of our solution to generate maps that accurately represent the various conformations present. This metric was calculated using the Model-Map correlation and offset log function available in Phenix.

3.4.3 Masked FSC (Map - Reference Map)

The Masked Fourier Shell Correlation (FSC) between Map and Reference Map plots is a key metric used to evaluate the quality of cryo-EM reconstructions. By plotting the correlation

between the reconstructed and reference maps against spatial frequency, the FSC curve reveals the resolution at which the reconstruction accurately reflects the true structure. Higher correlation values indicate better agreement and thus higher confidence in the accuracy of the reconstruction.

The 0.143 cutoff on the FSC curve is particularly important as it denotes the resolution limit of the reconstruction. This is the point at which the correlation drops below a significant level, suggesting that any features beyond this frequency are more likely due to noise than actual structural information. A higher frequency value at the 0.143 cutoff indicates a higher resolution reconstruction with more reliable fine details. This metric was plotted using a custom program available, `dataset_fsc`, in this project's repository.

3.4.4 Map – Reference Model Correlation

Map - Reference Model correlation assesses the overall similarity between the generated maps and reference atomic models. A higher correlation value for our solution compared to CryoDRGN would imply a closer resemblance to the reference structure, suggesting a more faithful reconstruction of the underlying structural information. This would further support the reliability and accuracy of our proposed solution in identifying and characterizing the diverse conformations within a dataset. This metric was calculated using the Model-Map correlation and offset log function available in Phenix.

3.4.5 Map - Reference Map Difference

Generating difference maps by subtracting the output maps of both our solution and CryoDRGN from the reference map provides a visual representation of the discrepancies between each method and the ground truth. Analyzing these difference maps allows us to identify patterns of errors, revealing potential biases or systematic issues in the reconstruction algorithms. Furthermore, by comparing the difference maps of our solution and CryoDRGN, we can assess the relative strengths and weaknesses of each method, pinpointing regions where

one approach outperforms the other. The difference maps were generated using the subtract function in ChimeraX and at the same threshold level for the maps being compared.

Chapter 4

RESULTS

To rigorously assess the efficacy of our proposed graph encoder model in heterogenous conformation reconstruction, we conducted a comprehensive evaluation using a combination of simulated and real datasets. This multifaceted approach allowed us to utilize the ground truth conformations available in the simulated datasets while also validating our model's performance on authentic cryo-EM data. The results presented in the following subsections offer insights into the capabilities of our model and provide a direct comparison with the established CryoDRGN method.

Specifically, we first analyze the resolution of our reconstructed cryo-EM maps compared to those generated by CryoDRGN. We then delve deeper into the conformational heterogeneity captured by each method, utilizing metrics such as the percentage of residues not found in acceptable density and the Fourier shell correlation (FSC) between the reconstructed and reference maps. Finally, we visualize the differences between the reconstructed maps and the reference maps to gain a qualitative understanding of the strengths and weaknesses of each approach.

4.1 Unmasked Resolution

Figure 4.1 presents a comprehensive comparison of unmasked resolutions achieved by GraphConv and CryoDRGN across a diverse set of datasets, encompassing both simulated and real cryo-EM data. Across all evaluated datasets, GraphConv consistently demonstrates superior performance compared to CryoDRGN, achieving lower unmasked resolution values and thus revealing a higher level of detail in the reconstructed maps.

In the simulated datasets (RNA Polymerase I, SARS-CoV-2, and CDTb), GraphConv's

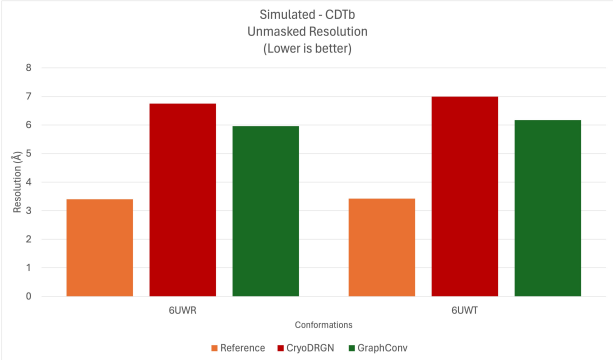
advantage is evident, with the most significant improvement observed in the CDTb dataset (Figure 4.1a), where differences range from 0.79 Å to 0.82 Å. This highlights GraphConv's effectiveness in resolving conformational changes within complex membrane environments. For the simulated SARS-CoV-2 (Figure 4.1b) and RNA Polymerase I datasets (Figure 4.1c), the improvements are more subtle, ranging from 0.1-0.2 Å and 0.2-0.3 Å, respectively. This suggests that the specific characteristics of the dataset, such as the nature and magnitude of conformational changes, may influence the degree of GraphConv's performance improvement.

Similarly, GraphConv consistently outperforms CryoDRGN in the real datasets (EMPIAR 10076 and EMPIAR 10028). In EMPIAR 10076 (Figure 4.1e), GraphConv's advantage is more pronounced, with differences of approximately 1 Å across various conformations, showcasing its ability to capture conformational heterogeneity in the human mitochondrial ribosome. For EMPIAR 10028, while the improvements are smaller, they remain consistent, suggesting a slight advantage of GraphConv over CryoDRGN in resolving drug-induced conformational changes in the Plasmodium falciparum 80S ribosome. Notably, in the EMPIAR datasets, GraphConv's reconstructions frequently approach the reference resolution, particularly in EMPIAR-10076, where differences range from 0.5 to 1.0 Å, indicating a remarkable ability to capture fine structural details.

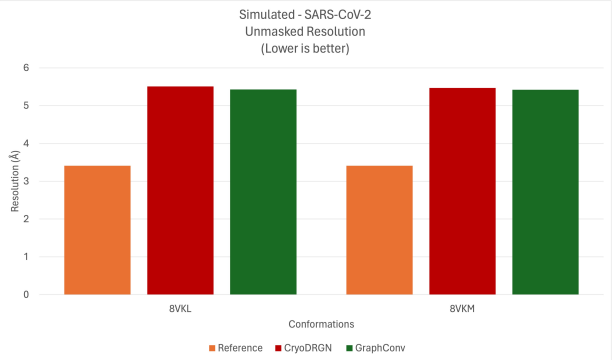
The observed trend suggests that GraphConv is particularly adept at capturing large-scale conformational rearrangements, as evidenced by the most significant improvements achieved for the CDTb data, which exhibits substantial rearrangements of subunits compared to the subtler changes observed in other datasets.

4.2 Percentage of Residues not found in Acceptable Density

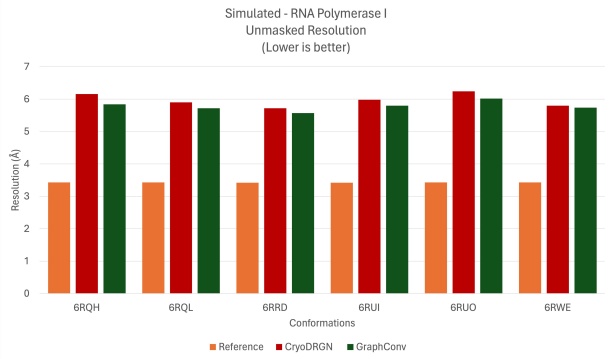
Figure 4.2 demonstrates the superior performance of GraphConv compared to CryoDRGN in accurately representing protein structures and capturing conformational changes across most simulated datasets. This is evidenced by a consistent reduction in the percentage of residues not found in acceptable density (AD), a key metric reflecting the agreement between the generated map and the reference structure.



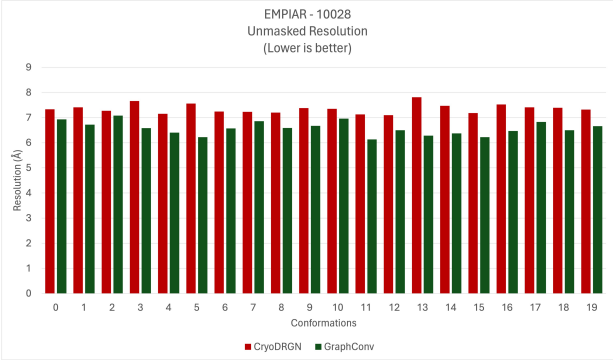
(a) CDTb



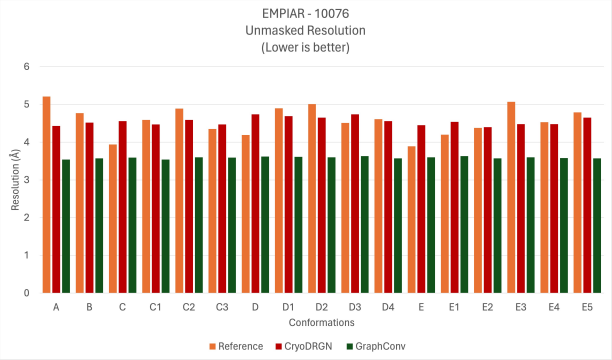
(b) SARS-CoV-2



(c) RNA Polymerase I



(d) 10028



(e) 10076

Figure 4.1: Unmasked Resolution. The bar plots measure the resolution of the reconstructions generated by CryoDRGN and GraphConv. The results demonstrate that GraphConv consistently generates better quality volumes as compared to CryoDRGN.

In the RNA Polymerase I dataset (Figure 4.2c), GraphConv moderately improves the representation of diverse functional states by reducing the percentage of residues not found in AD by 1.16% to 2.16% across all conformations. For the SARS-CoV-2 dataset (Figure 4.2b), GraphConv exhibits slight improvements of 0.61% and 0.26% for the two conformations, respectively.

The most substantial improvement of GraphConv is observed in the CDTb dataset (Figure 4.2a), where the percentage of residues not found in AD is significantly reduced by 1.99% for the 6UWR conformation and 2.1% for the 6UWT conformation. This highlights GraphConv's superior ability to accurately model complex structural transitions in membrane-bound proteins, particularly given the substantial rearrangement of subunits and conformational plasticity inherent to CDTb.

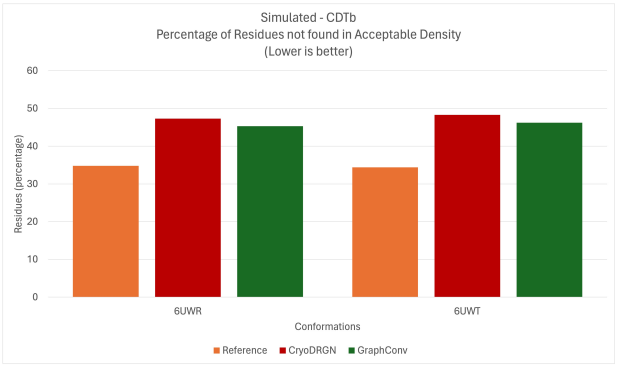
4.3 Masked FSC (Map - Reference Map)

The Fourier Shell Correlation (FSC) curves in Figure 4.3 reveal a consistent trend of GraphConv achieving slightly higher resolution than CryoDRGN across all three simulated datasets. This improvement, while subtle for the RNA Polymerase I and SARS-CoV-2 Spike Protein datasets, is more pronounced in the CDTb dataset.

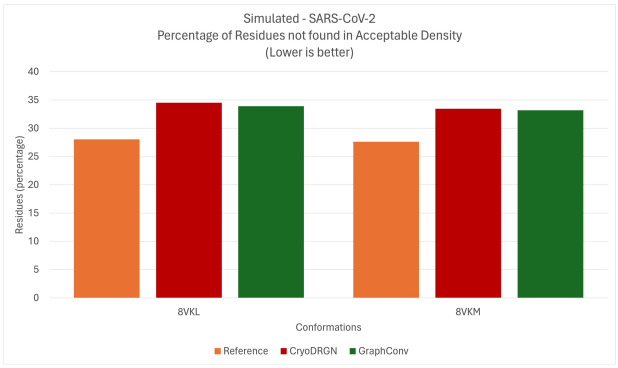
For RNA Polymerase I (Figures 4.3e and 4.3f), GraphConv's slight edge at the 0.143 cutoff suggests a better ability to capture fine structural details. Similarly, for the SARS-CoV-2 Spike Protein dataset (Figures 4.3c and 4.3d), GraphConv's minor improvement, particularly at higher frequencies, hints at its potential advantage in resolving certain aspects of the protein's conformational changes.

The most significant improvement is observed in the CDTb dataset, where GraphConv (Figure 4.3b) notably outperforms CryoDRGN (Figure 4.3a). The approximately 0.5Å resolution improvement at the 0.143 cutoff for both conformations underscores GraphConv's exceptional ability to resolve the complex subunit rearrangements and conformational plasticity inherent to CDTb.

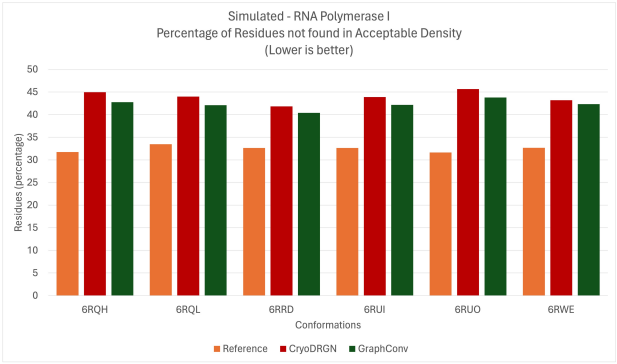
Since reference maps were available for EMPIAR 10076, we also plotted the FSC curves



(a) CDTb

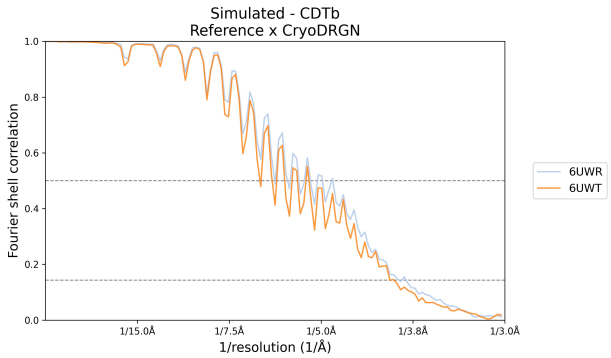


(b) SARS-CoV-2

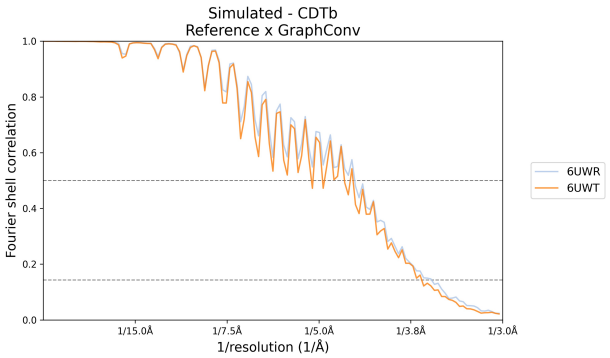


(c) RNA Polymerase I

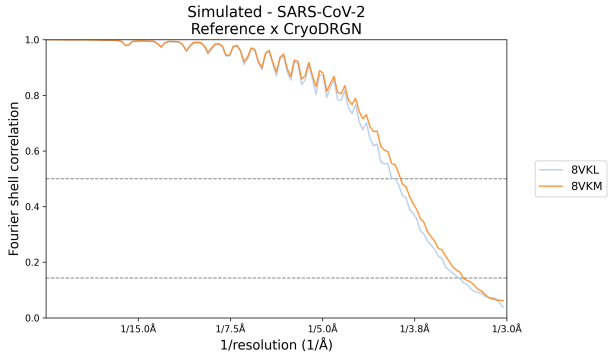
Figure 4.2: Residues not found in Acceptable Density. The bar plots measure the residues not found in regions of the generated map with acceptable density. The results demonstrate the our method, GraphConv, performed better than CryoDRGN consistently with an average 2% improvement.



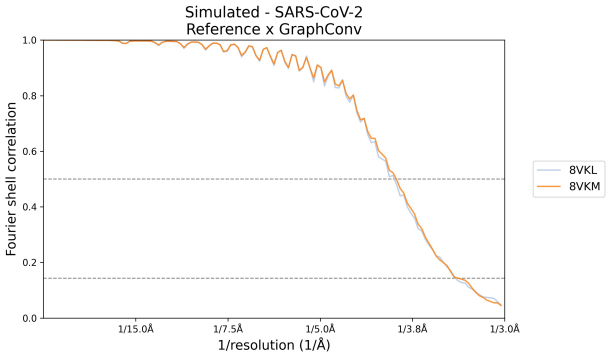
(a) CDTb - CryoDRGN



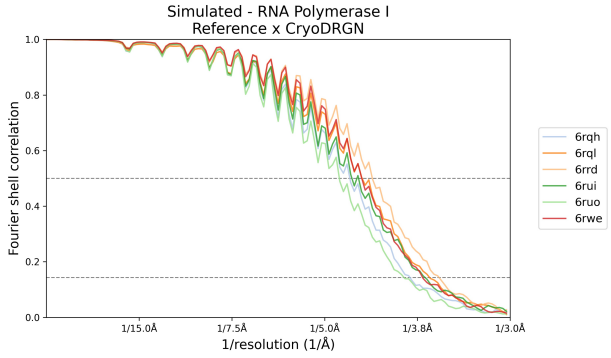
(b) CDTb - GraphConv



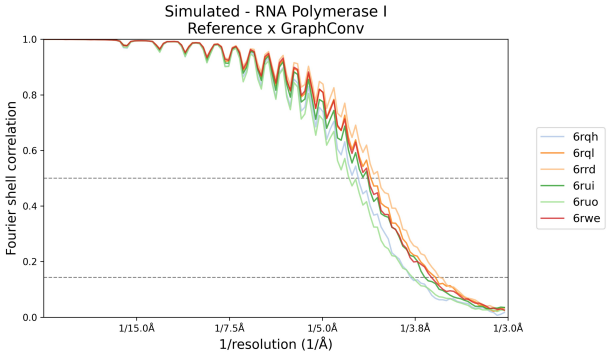
(c) SARS-CoV-2 - CryoDRGN



(d) SARS-CoV-2 - GraphConv

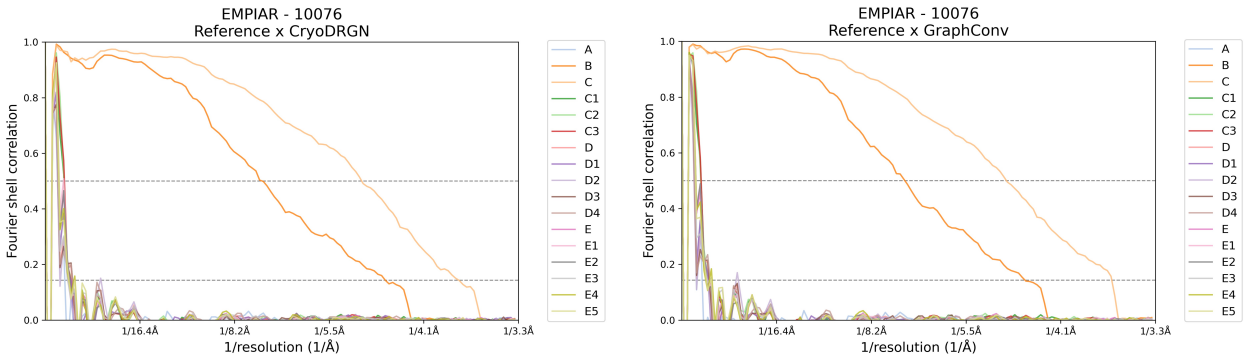


(e) RNA Polymerase I - CryoDRGN



(f) RNA Polymerase I - GraphConv

Figure 4.3: Masked FSC (Reference x Generated). These plots show the FSC values between the reference and generated maps at different resolutions. (a) GraphConv consistently has higher values than CryoDRGN, indicating better alignment with the reference maps. (b) Both methods have similar values, indicating comparable performance. (c) GraphConv outperforms CryoDRGN, showing slightly better agreement with the reference maps.



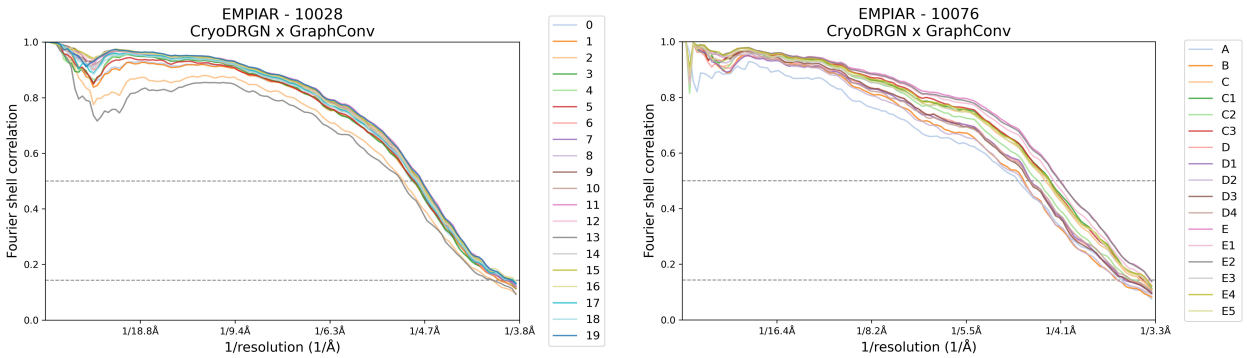
(a) CryoDRGN

(b) GraphConv

Figure 4.4: Masked FSC (10076 - Reference x Generated). These plots display the FSC values between the maps at various resolutions for the EMPIAR 10076 dataset. Due to an unidentified issue, the plots show unusual values across most conformations for both methods, rendering the results uninterpretable.

between them and the generated maps (see Figure 4.4). However, the curves for most of the volumes exhibited unusual oscillations and sudden drops. While some fluctuation is expected, the erratic nature of these curves suggests potential issues with either the reference maps or the generated volumes. But, during alignment in ChimeraX, the volumes had more than 0.9 correlation with each other which makes these plots more confusing. Furthermore, the plot comparing Reference maps and CryoDRGN's maps published by Zhong et al. [34] does not have these issues. Hence, we decided to compare our outputs with CryoDRGN's outputs to ascertain that the generated volumes match the expected structures.

The Fourier Shell Correlation (FSC) plots in Figure 4.5 show the agreement between the maps generated by CryoDRGN and GraphConv for the two real datasets, EMPIAR-10028 (Figure 4.5a) and EMPIAR-10076 (Figure 4.5b). For both datasets, the FSC curves between CryoDRGN and GraphConv are high at low frequencies (large scales) but decrease with increasing frequency (smaller scales). This is expected, as it is more difficult to achieve



(a) 10028 (b) 10076

Figure 4.5: Masked FSC (CryoDRGN x GraphConv). For the real datasets, we use CryoDRGN's generated volumes as the ground truth (validated by the CryoDRGN team [34]) to compute the FSC curve with GraphConv's generated volumes. The values are generally high and pass the cutoffs at around 4.5Å and 3Å.

agreement at higher resolutions. In the absence of a reference map for comparison, it is difficult to definitively determine which method produces maps that better reflect the true structure. However, the high degree of agreement between the CryoDRGN and GraphConv maps suggests that both methods are producing maps that are similar to each other.

4.4 Map – Reference Model Correlation

The map-to-model correlation analysis across the three simulated datasets in Figure 4.6 reveals varying performance for GraphConv compared to CryoDRGN. Higher correlation values indicate a closer resemblance to the reference map. Both solutions exhibit virtually identical performance across all six conformations for the RNA Polymerase I dataset (see Figure 4.6c). This indicates that both methods were equally capable of capturing the diverse functional states and subtle structural variations within this complex molecule. protein.

In contrast, GraphConv displays a slight edge over CryoDRGN for the SARS-CoV-2 spike

protein dataset (see Figure 4.6b), with improvements of 0.01 for both the 8VKL (pre-fusion) and 8VKM (post-fusion) conformations. These marginal gains might indicate GraphConv's ability to slightly better capture the large-scale conformational changes associated with viral fusion.

However, for the CDTb dataset (see 4.6a), GraphConv shows a slight decrease in correlation for both conformations (6UWR and 6UWT) by 0.05 and 0.04, respectively, suggesting that CryoDRGN was marginally better at capturing the substantial subunit rearrangements in this complex.

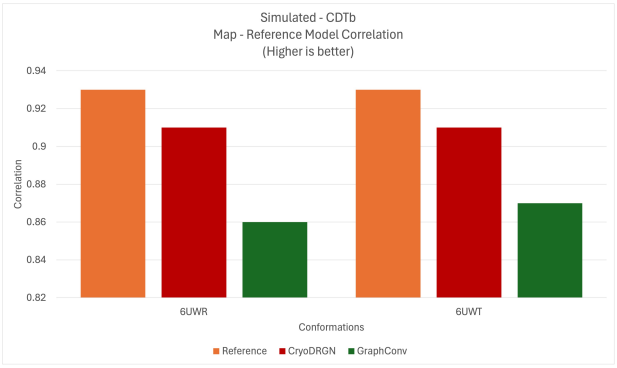
4.5 Map - Reference Map Difference

In Figure 4.8, the CryoDRGN results for the SARS-CoV-2 Spike protein exhibit a noticeable blob of density in the center of the protein for both the 8VKL and 8VKM conformations. However, the GraphConv results do not show this density blob. This suggests that CryoDRGN might be over-interpreting certain regions or introducing artifacts. GraphConv, on the other hand, produces a map that more closely aligns with the reference map, indicating superior accuracy in capturing the true structure.

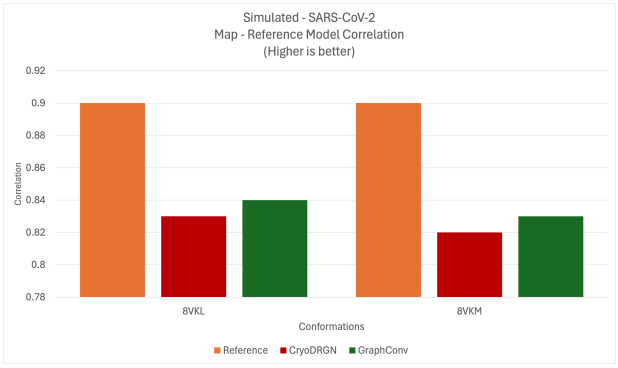
Similarly, in Figure 4.7b for the CDTb 6UWT conformation, a density blob is observed in the center of the CryoDRGN result but is absent in the GraphConv result. This further supports the idea that CryoDRGN might be prone to generating spurious densities in the interior regions of the protein structures.

In contrast, it is important to note that for the remaining volumes (see Figures 4.7a and 4.9), both CryoDRGN and GraphConv produce almost identical results, and the volumes are correctly represented in both.

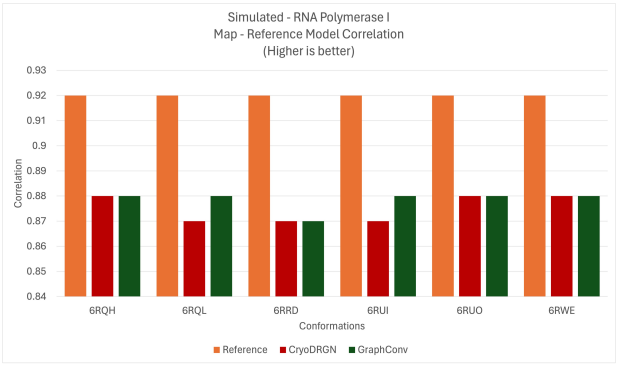
The absence of this density blob in the GraphConv results for three of the generated volumes suggests that GraphConv is better at capturing density details, specifically from the inner parts of a volume, compared to CryoDRGN. This is a significant advantage, as it implies that GraphConv might be more reliable in identifying and characterizing conformational changes that occur within the core of a protein structure.



(a) CDTb

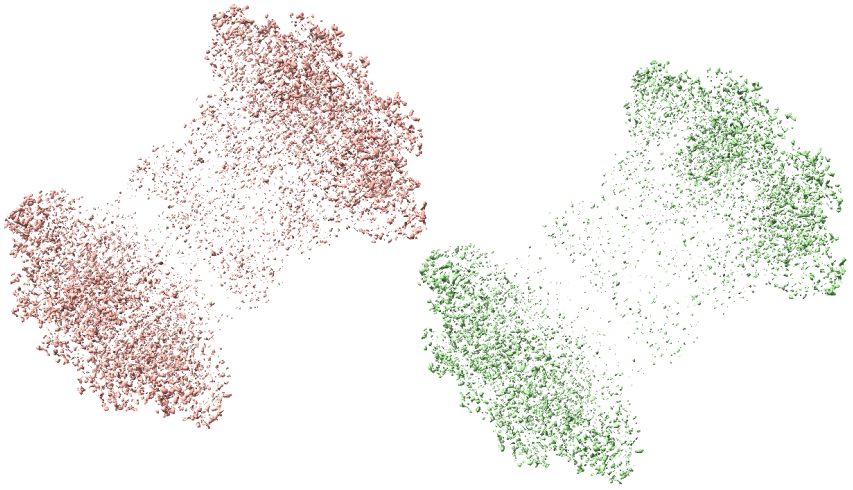


(b) SARS-CoV-2

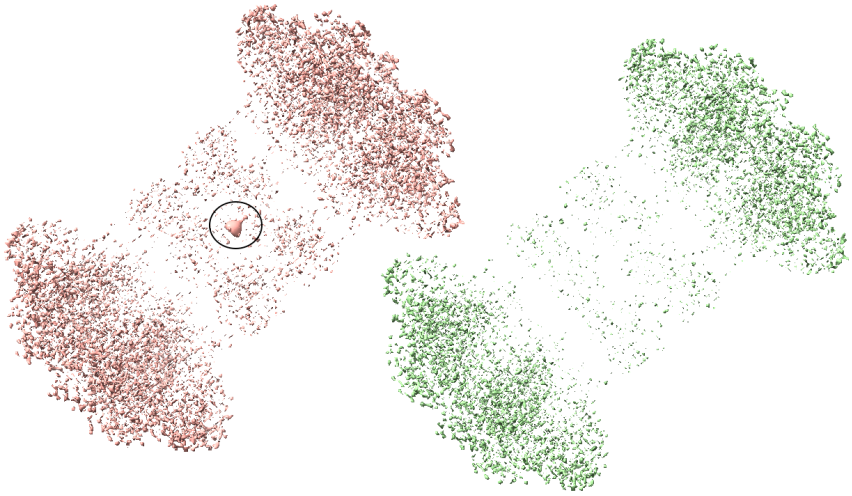


(c) RNA Polymerase I

Figure 4.6: Map – Reference Model Correlation. These bar plots assess the fit between the generated maps and the reference models. (a) GraphConv performed 5% worse than CryoDRGN, indicating a better correlation with the reference model for CryoDRGN. (b) GraphConv showed a slight 1% improvement, but this difference is not significant. (c) GraphConv performed as well as or slightly better than CryoDRGN in all cases, but the improvements are still not significant.

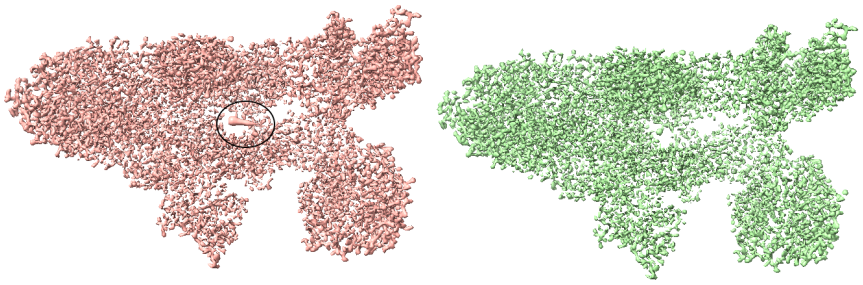


(a) 6UWR

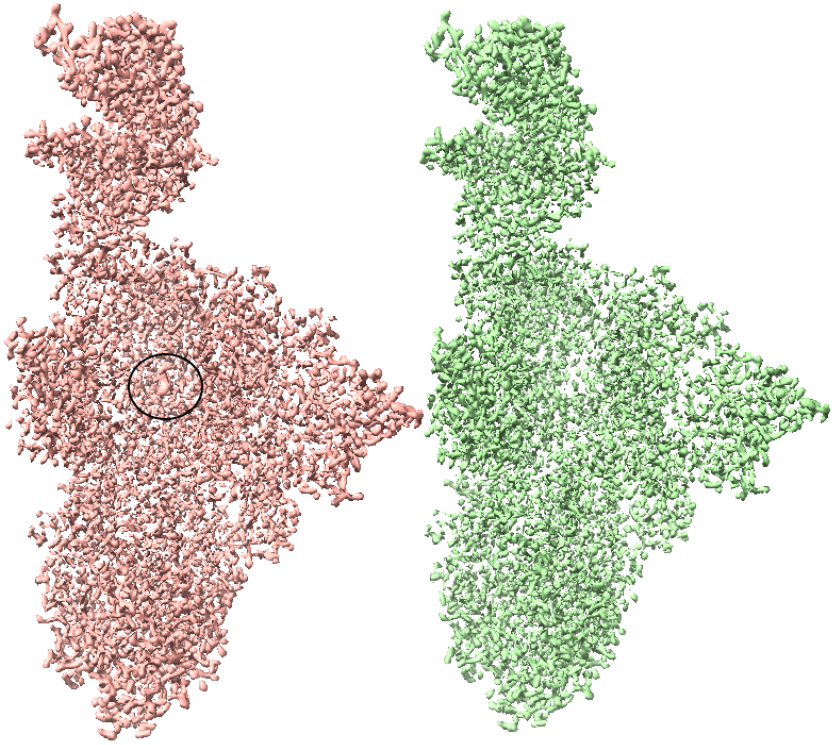


(b) 6UWT

Figure 4.7: Difference Map - CDTb (Red: CryoDRGN and Green: GraphConv). (a) Both methods generated similar maps. (b) CryoDRGN has generated an additional density blob in the central region (highlighted by a black oval) indicating a potential problem in its structure encoding process.

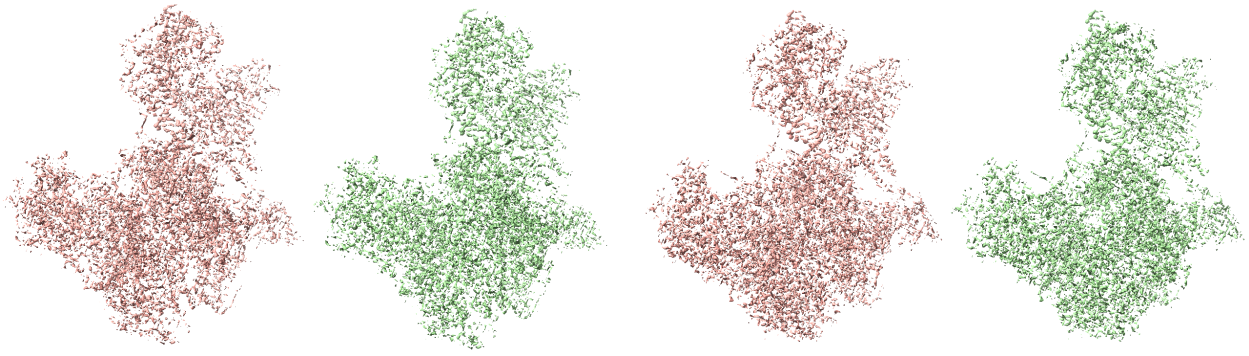


(a) 8VKL



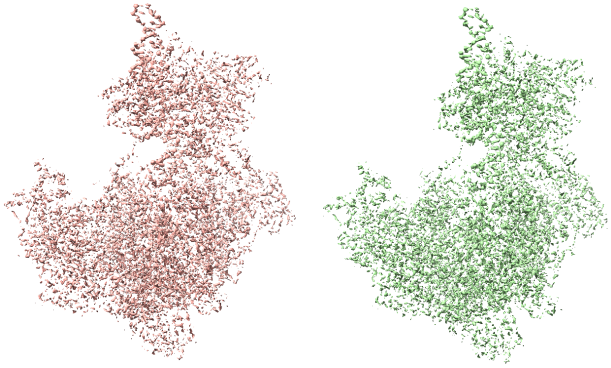
(b) 8VKM

Figure 4.8: Difference Map - SARS-CoV-2 (Red: CryoDRGN and Green: GraphConv). Both (a) and (b) highlight the issue with the generated volumes from CryoDRGN with an additional density blob in the central region.

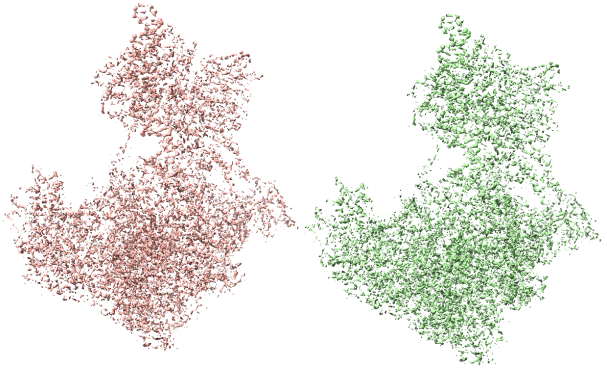


(a) 6RQH

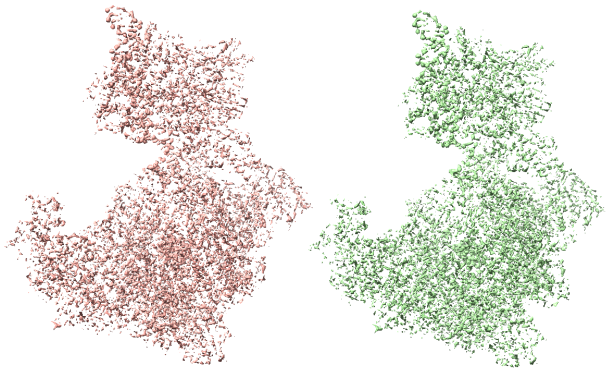
(b) 6RQL



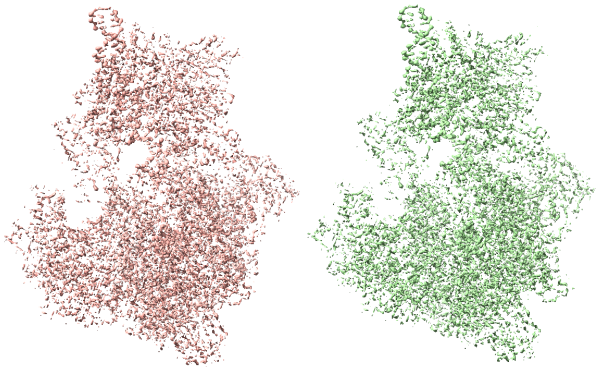
(c) 6RRD



(d) 6RUI



(e) 6RUO



(f) 6RWE

Figure 4.9: Difference Map - RNA Polymerase I (Red: CryoDRGN and Green: GraphConv). All of the difference maps are similar without any issues in any of the generated volumes.

Chapter 5

CONCLUSION

In this study, we proposed a novel graph neural network (GNN) encoder model and integrated it into the CryoDRGN framework to enhance the reconstruction of 3D volumes from 2D cryo-electron microscopy (cryo-EM) images. By replacing CryoDRGN's existing encoder with our custom GNN-based model, we aimed to improve the quality and accuracy of the generated volumes. Our approach involved converting input particle images into graph structures using an on-the-fly conversion function, which was then processed by graph convolutional layers and 1D convolutional layers to extract features and reduce dimensionality. The resulting embeddings were then utilized by CryoDRGN's decoder to generate the corresponding 3D volumes. We also developed a process to simulate datasets using open-source tools and a custom Python script to evaluate our model with ground-truth validation and compare it with CryoDRGN.

Our results, detailed in Chapter 4, demonstrated the effectiveness of our GNN encoder model in improving the quality of 3D volume reconstructions. We observed consistent enhancements in unmasked resolution across various datasets, particularly in the EMPIAR datasets, where GraphConv's reconstructions often closely approached the reference resolution. This indicates a superior ability to capture fine structural details compared to the original CryoDRGN model. Additionally, our model demonstrated improved accuracy in detecting residues within acceptable density thresholds for the CDTb dataset, further highlighting its effectiveness in representing the underlying molecular structures. The Masked Fourier Shell Correlation (FSC) plots also revealed a clear improvement in resolution achieved by GraphConv over CryoDRGN, particularly for the CDTb dataset.

A key factor in the observed improvements could be attributed to the use of graphs to

represent the particle image data. Unlike traditional grid-like representations, graphs can capture the inherent relationships between different regions of the image without imposing a fixed structure. This flexibility allows the model to better adapt to the unique characteristics of each particle image, potentially leading to more accurate and detailed reconstructions.

5.1 Limitations

However, it is important to acknowledge the limitations of our approach. While our modified CryoDRGN framework showed improvements in volume reconstruction, it did not significantly enhance the detection of diverse conformational states beyond the capabilities of the original model. This suggests that while our GNN encoder model excels at capturing structural details, it may not be sufficient to address the inherent challenges in identifying all distinct conformations within a dataset. Additionally, the improvements in resolution and residue detection were not uniform across all datasets, indicating that the effectiveness of our approach may be influenced by the specific characteristics of the dataset, such as the nature and extent of conformational changes. Furthermore, the evaluation metrics used in this study, while informative, may not fully capture all aspects of reconstruction quality and conformational heterogeneity. For instance, the unmasked resolution metric may not be sensitive to subtle differences in local resolution, and the percentage of residues not found in acceptable density may not fully reflect the accuracy of the model in capturing the full range of conformational variability. Finally, our model takes longer to train as compared to the CryoDRGN model with training times being approximately 10 times longer due to the on-the-fly graph creation process. Based on our experimentation, creating the graphs before training is significantly slower due to the file conversion and storage process and takes at least 10 days for all datasets.

5.2 Future Work

To further enhance the capabilities of our approach, future research could explore the integration of GNNs throughout the entire encoding and decoding process, rather than just

in the encoder. This would involve developing a GNN-based decoder that can effectively utilize the graph representations generated by the encoder to produce more accurate and informative 3D volumes. By incorporating GNNs in both the encoding and decoding stages, we can potentially achieve a more seamless and comprehensive utilization of graph-based information, leading to further improvements in reconstruction quality and conformational state detection.

In addition to expanding the use of GNNs, future work could also focus on refining the graph construction methods. The current approach of converting particle images into graphs based on density thresholds and k-nearest neighbors may not be optimal for all datasets. Exploring alternative graph construction techniques, such as those based on learned edge features or incorporating domain-specific knowledge, could lead to more informative graph representations and, consequently, better reconstruction results.

Furthermore, incorporating attention mechanisms into the GNN architecture could be a promising avenue for future research. Attention mechanisms allow the model to focus on the most relevant parts of the graph during processing, potentially improving its ability to capture subtle conformational differences and distinguish between diverse states. By selectively attending to informative nodes and edges, the model can better utilize the graph structure to generate more accurate and detailed 3D volumes.

Finally, evaluating the framework on a wider range of datasets with diverse characteristics would be crucial for assessing its robustness and generalizability. By testing the model on datasets with varying levels of complexity, conformational heterogeneity, and image quality, we can gain a deeper understanding of its strengths and weaknesses and identify areas for further improvement. This would also help in developing a more comprehensive and versatile tool for cryo-EM structure determination, capable of handling a wide range of biological systems and experimental conditions.

BIBLIOGRAPHY

[1] Pavel V. Afonine, Bruno P. Klaholz, Nigel W. Moriarty, Billy K. Poon, Oleg V. Sobolev, Thomas C. Terwilliger, Paul D. Adams, and Alexandre Urzhumtsev. New tools for the analysis and validation of cryo-em maps and atomic models. *Acta Crystallographica Section D Structural Biology*, 74(9):814–840, September 2018.

[2] Rn Bracewell. Strip Integration in Radio Astronomy. *Australian Journal of Physics*, 9(2):198, 1956.

[3] Tom Burnley, Colin M. Palmer, and Martyn Winn. Recent developments in the ccp-em software suite. *Acta Crystallographica Section D Structural Biology*, 73(6):469–477, June 2017.

[4] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. Towards sparse hierarchical graph classifiers. 2018.

[5] Yuan Chiang, Wei-Han Hui, and Shu-Wei Chang. Encoding protein dynamic information in graph representation for functional residue identification. *Cell Reports Physical Science*, 3(7):100975, July 2022. arXiv:2112.12033 [cond-mat, physics:physics, q-bio].

[6] Joseph H. Davis, Yong Zi Tan, Bridget Carragher, Clinton S. Potter, Dmitry Lyumkis, and James R. Williamson. Modular assembly of the bacterial large ribosomal subunit. *Cell*, 167(6):1610–1622.e15, December 2016.

[7] Allison Doerr. Single-particle cryo-electron microscopy. *Nature Methods*, 13(1):23–23, January 2016.

[8] Hongyang Gao and Shuiwang Ji. Graph u-nets. 2019.

[9] Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors H. W. Scheres. Automated model building and protein identification in cryo-em maps. *Nature*, February 2024.

[10] Laurel F. Kinman, Barrett M. Powell, Ellen D. Zhong, Bonnie Berger, and Joseph H. Davis. Uncovering structural ensembles from single-particle cryo-em data using cryodrgn. *Nature Protocols*, November 2022.

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2016.

[12] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. 2019.

[13] Zijie Li, Kazem Meidani, Prakarsh Yadav, and Amir Barati Farimani. Graph neural networks accelerated molecular dynamics. *The Journal of Chemical Physics*, 156(14):144103, April 2022. arXiv:2112.03383 [physics].

[14] Dorothee Liebschner, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkóczi, Vincent B. Chen, Tristan I. Croll, Bradley Hintze, Li-Wei Hung, Swati Jain, Airlie J. McCoy, Nigel W. Moriarty, Robert D. Oeffner, Billy K. Poon, Michael G. Prisant, Randy J. Read, Jane S. Richardson, David C. Richardson, Massimo D. Sammito, Oleg V. Sobolev, Duncan H. Stockwell, Thomas C. Terwilliger, Alexandre G. Urzhumtsev, Lizbeth L. Videau, Christopher J. Williams, and Paul D. Adams. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D Structural Biology*, 75(10):861–877, October 2019.

[15] Zhenwei Luo, Fengyun Ni, Qinghua Wang, and Jianpeng Ma. Opus-dsd: deep structural disentanglement for cryo-em single-particle analysis. *Nature Methods*, 20(11):1729–1738, November 2023.

[16] Tengfei Ma and Jie Chen. Unsupervised learning of graph hierarchical abstractions with differentiable coarsening and optimal transport. 2019.

[17] Dhiraj Mannar, James W. Saville, Chad Poloni, Xing Zhu, Alison Bezeruk, Keith Tidey, Sana Ahmed, Katharine S. Tuttle, Faezeh Vahdatihassani, Spencer Cholak, Laura Cook, Theodore S. Steiner, and Sriram Subramaniam. Altered receptor binding, antibody evasion and retention of t cell recognition by the sars-cov-2 xbb.1.5 spike protein. *Nature Communications*, 15(1):1854, February 2024.

[18] Elaine C. Meng, Thomas D. Goddard, Eric F. Pettersen, Greg S. Couch, Zach J. Pearson, John H. Morris, and Thomas E. Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, November 2023.

[19] Takanori Nakane and Sjors H. W. Scheres. *Multi-body Refinement of Cryo-EM Images in RELION*, volume 2215 of *Methods in Molecular Biology*, page 145–160. Springer US, New York, NY, 2021.

[20] Eva Nogales. The development of cryo-em into a mainstream structural biology technique. *Nature Methods*, 13(1):24–27, January 2016.

[21] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, Jan 2021.

[22] Ali Punjani and David J. Fleet. 3d variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-em. *Journal of Structural Biology*, 213(2):107702, June 2021.

[23] Ali Punjani and David J. Fleet. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870, June 2023.

[24] Ekagra Ranjan, Soumya Sanyal, and Partha Pratim Talukdar. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. 2019.

[25] Heng Ru, Melissa G. Chambers, Tian-Min Fu, Alexander B. Tong, Maofu Liao, and Hao Wu. Molecular mechanism of v(d)j recombination from synaptic rag1-rag2 complex structures. *Cell*, 163(5):1138–1152, November 2015.

[26] Yashar Sadian, Florence Baudin, Lucas Tafur, Brice Murciano, Rene Wetzel, Felix Weis, and Christoph W. Müller. Molecular insight into rna polymerase i promoter recognition and promoter melting. *Nature Communications*, 10(1):5543, December 2019.

[27] Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. Eman2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, Jan 2007.

[28] Yuko Tsuchiya and Kentaro Tomii. Neural networks for protein structure and function prediction and dynamic analysis. *Biophysical Reviews*, 12(2):569–573, April 2020.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. 2017.

[30] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors Hw Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, June 2014.

[31] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? 2018.

[32] Xingjian Xu, Raquel Godoy-Ruiz, Kaylin A. Adipietro, Christopher Peralta, Danya Ben-Hail, Kristen M. Varney, Mary E. Cook, Braden M. Roth, Paul T. Wilder, Thomas Cleveland, Alexander Grishaev, Heather M. Neu, Sarah L. J. Michel, Wenbo Yu, Dorothy Beckett, Richard R. Rustandi, Catherine Lancaster, John W. Loughney, Adam Kristopeit, Sianny Christanti, Jessica W. Olson, Alexander D. MacKerell, Amedee Des Georges, Edwin Pozharski, and David J. Weber. Structure of the cell-binding component of the clostridium difficile binary toxin reveals a di-heptamer macromolecular assembly. *Proceedings of the National Academy of Sciences*, 117(2):1049–1058, January 2020.

[33] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 12:690049, July 2021.

[34] Ellen D. Zhong, Tristan Bepler, Bonnie Berger, and Joseph H. Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 18(2):176–185, February 2021.

[35] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. 2018.

Appendix A

FULL TEST DATA

Table A.1: Unmasked Resolution

(Lower is better)

Dataset	Conformation	Reference (Å)	CryoDRGN (Å)	GraphConv (Å)
RNA Polymerase I	6RQH	3.43	6.16	**5.84**
	6RQL	3.43	5.9	**5.72**
	6RRD	3.42	5.72	**5.57**
	6RUI	3.42	5.98	**5.8**
	6RUO	3.43	6.24	**6.02**
	6RWE	3.43	5.8	**5.74**
SARS-CoV-2	8VKL	3.41	5.51	**5.43**
	8VKM	3.41	5.47	**5.42**
CDTb	6UWR	3.4	6.75	**5.96**
	6UWT	3.42	6.99	**6.17**

Table A.1: Unmasked Resolution (Continued)

(Lower is better)

Dataset	Conformation	Reference (Å)	CryoDRGN (Å)	GraphConv (Å)
EMPIAR 10076	A	5.21	4.43	**3.54**
	B	4.77	4.52	**3.57**
	C	3.94	4.56	**3.59**
	C1	4.59	4.47	**3.54**
	C2	4.89	4.59	**3.6**
	C3	4.35	4.47	**3.59**
	D	4.19	4.74	**3.62**
	D1	4.9	4.69	**3.61**
	D2	5.01	4.65	**3.6**
	D3	4.51	4.74	**3.63**
	D4	4.61	4.56	**3.57**
	E	3.89	4.45	**3.6**
	E1	4.2	4.54	**3.63**
	E2	4.38	4.4	**3.57**
	E3	5.07	4.48	**3.6**
	E4	4.53	4.48	**3.58**
	E5	4.79	4.65	**3.57**

Table A.1: Unmasked Resolution (Continued)

(Lower is better)

Dataset	Conformation	Reference (Å)	CryoDRGN (Å)	GraphConv (Å)
EMPIAR 10028	0	-	7.33	**6.93**
	1	-	7.41	**6.72**
	2	-	7.27	**7.08**
	3	-	7.66	**6.58**
	4	-	7.15	**6.4**
	5	-	7.56	**6.22**
	6	-	7.24	**6.57**
	7	-	7.23	**6.86**
	8	-	7.2	**6.59**
	9	-	7.38	**6.67**
	10	-	7.35	**6.96**
	11	-	7.13	**6.13**
	12	-	7.1	**6.5**
	13	-	7.81	**6.28**
	14	-	7.47	**6.37**
	15	-	7.18	**6.22**
	16	-	7.52	**6.47**
	17	-	7.41	**6.83**
	18	-	7.39	**6.5**
	19	-	7.32	**6.66**

Table A.2: Map-Model Correlation and % of Residues Not Found in Acceptable Density (AD)

Dataset	Conformation	Map-Model Correlation (Higher is better)		% of Residues not found in AD (Lower is better)	
		CryoDRGN	GraphConv	CryoDRGN	GraphConv
RNA Polymerase I	6RQH	0.88	0.88	44.95	**42.75**
	6RQL	0.87	**0.88**	44.00	**42.09**
	6RRD	0.87	0.87	41.82	**40.39**
	6RUI	0.87	**0.88**	43.88	**42.16**
	6RUO	0.88	0.88	45.66	**43.79**
	6RWE	0.88	0.88	43.19	**42.31**
SARS-CoV-2	8VKL	0.83	**0.84**	34.50	**33.89**
	8VKM	0.82	**0.83**	33.43	**33.17**
CDTb	6UWR	**0.91**	0.86	47.29	**45.30**
	6UWT	**0.91**	0.87	48.31	**46.21**