Machine Learning for Reconstructing Dynamic Protein Structures from Cryo-EM Images

by

Ellen D. Zhong

Submitted to the Computational & Systems Biology Graduate Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Computational & Systems Biology Graduate Program
May 24, 2022

Certified by . . . .
Bonnie Berger
Simons Professor of Mathematics
Thesis Supervisor

Certified by . . . .
Joseph H. Davis
Whitehead Assistant Professor of Biology
Thesis Supervisor

Accepted by . . . .
Christopher B. Burge
Director, Computational & Systems Biology Graduate Program

Machine Learning for Reconstructing Dynamic Protein Structures from Cryo-EM Images

by

Ellen D. Zhong

Submitted to the Computational & Systems Biology Graduate Program
on May 24, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Proteins and other biomolecules form dynamic macromolecular machines that carry out essential biological processes responsible for life. However, studying the mechanisms of these biomolecular complexes at relevant atomic-scale resolutions is an extraordinarily challenging task in structural biology. This thesis presents new algorithms that address the computational bottlenecks at the frontier of structure determination of dynamic biomolecular complexes via cryo-electron microscopy (cryo-EM).

In single particle cryo-EM, the central problem is to reconstruct the 3D structure of a target biomolecular complex from a set of noisy and randomly oriented 2D projection images, a challenging inverse problem especially when instances of the imaged biomolecular complex exhibit structural heterogeneity.

The main contribution of this thesis is a machine learning system, cryoDRGN, for reconstructing continuous distributions of biomolecular structures from cryo-EM images. Underpinning the cryoDRGN method is a deep generative model parameterized by a new neural representation of cryo-EM volumes and a learning algorithm to optimize this representation from unlabeled 2D cryo-EM images. Released as an open source software tool, cryoDRGN has been applied on real datasets to uncover heterogeneity in high resolution datasets, discover new conformations of large macromolecular machines and visualize continuous trajectories of their motion. This thesis also describes an extension, cryoDRGN2, for learning this model from unposed images, i.e. *ab initio* reconstruction. Finally, this thesis presents emerging directions in analyzing the learned manifold of cryo-EM structures and in incorporating atomic model priors into cryo-EM reconstruction.

Thesis Supervisor: Bonnie Berger
Title: Simons Professor of Mathematics

Thesis Supervisor: Joseph H. Davis
Title: Whitehead Assistant Professor of Biology

Acknowledgments

First, I would like to thank my advisors, Joey Davis and Bonnie Berger, who have been unbelievably supportive over the years. I am deeply grateful for the freedom that they have given me to pursue cross-disciplinary research, and for providing two distinct but equally enriching intellectual homes during my time at MIT. Through working with the both of them, I've been exposed to different research communities – broadly, computer science and life sciences – that have decidedly non-overlapping language and norms. I am lucky that I could count on Bonnie and Joey for guidance in navigating these differences between fields, which has been a hugely rewarding learning experience of my Ph.D.

I've had the privilege to work with fantastic peers and collaborators both at MIT and around the world. Tristan Bepler was an instrumental collaborator in the early days of my foray into machine learning. I am grateful to have the opportunity to work with Brian Hie to bring our skills to bear on fighting the COVID pandemic. I am lucky to have befriended and pipetted with Sam Rodriques during my first year. I've had the pleasure to collaborate on various projects with my labmates Ashwin Narayan, Laurel Kinman, and Barrett Powell. My additional thanks to Samantha Webster and Ben Demeo, and my other peers in the Davis and Berger labs for their feedback and support.

I thank all the mentors and colleagues I've had prior to MIT, in particular, my undergraduate advisor Michael Shirts, who originally instilled in me the values of rigorous and reproducible research. I learned many lessons that prepared me well for my Ph.D. after working with colleagues at D. E. Shaw Research, including Huafeng Xu, Thomas Weinreich, Kshitij Lauria, Justin Gullingsrud, and Robert McGibbon.

I knew immediately that I had made the right decision to pursue a Ph.D. and the right program when taking CSB.100 my first semester with Professor Chris Burge. Chris has been a role model for me in deeply engaging in the nitty-gritty technical details and the overarching biological questions. I thank my Ph.D. cohort in Computational and Systems Biology, Brian Trippe, Cameron Flower, Conner

Cummerlowe, Michael Murphy, and Adam Atanas, for the shared camaraderie in navigating interdisciplinary and diverse fields and subfields, and for the memories from that one time we reached full quorum on the GSC ski trip.

This thesis work has relied on support, guidance, and advice from many researchers in the broader cryo-EM, structural biology, and machine learning fields. I thank Roy Lederman and Amit Singer, whose work first inspired me to work on cryo-EM reconstruction; Pilar Cossio and Sonya Hanson at the Flatiron Institute; my committee members Thomas Schwartz and Niko Grigorieff; early cryoDRGN adopters Tanya Bodrug, David Haselbach, Miao Gui, and Alan Brown; Rado Enchev from the Crick Institute; Bridget Carragher from NYSBC; Navid Paknejad who plugged me into the cryo-EM network in NYC; Peers from MLSB, including Erika Alden DeBenedictus, Namrata Anand, Stephan Eismann, John Ingraham, Sergey Ovchinnikov, Roshan Rao, and Raphael Townshend; Colleagues from DeepMind, including John Jumper, Michael Figurnov, Alex Pritzel, Jonas Adler, Rob Fergus, Jason Yim, and Rishub Jain.

Thank you to the friends who have provided a counter-balance to the myopic nature of research, and truly made the last few years meaningful. In addition to many of those named above who I consider dear friends, thank you: Anton Bahktin, Sean Baxter, Jonathan Benezry, Davis Blalock, Noam Brown, Josh Chen, Irene Chen, Soumith Chintala, Tarun Chitra, Ariel Deutsch, Lawrence Diao, Thomas Georgiou, Kavya Kilari, Pallav Kosuri, Emma Kowal, Zeming Lin, Conor McMann, Sarah Nyquist, Thy Pham, Eitan Reich, Semon Rezchikov, Grace Riccardi, Reuben Saunders, Sahar Shahamatdar, Vincent Sitzmann, Elana Simon, Julie Ta, Jim Valcourt, Colvin Wang, Yuki Weber, Emily Xie, Franklin Yang, and Biqi Zhang.

Finally, this Ph.D. would not have been possible without the unfailing support of my partner (and occasional collaborator) Adam Lerer. All my thanks to my family who have been so supportive of my studies. This thesis is dedicated to Adam, my family, and my grandfather Zhong Ting Le.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins and protein complexes carry out many of the biological processes responsible for life, including DNA replication and repair, catalyzing reactions for metabolism and synthesis, as well as the degradation of aging or damaged cellular components. Built from 10s to 100s of individual protein, RNA, or lipid components, these nanometer scale machines are assembled into complex structures that enable their function. Direct observation of these structures at the atomic level has proven invaluable for understanding the inner workings of these complexes. However, many of these massive molecular machines have posed substantial challenges for traditional structural biology techniques, as such biomolecular machines are compositionally and conformationally dynamic, often undergoing large structural changes to perform their function.

Fueled by advances in both software and hardware, cryo-electron microscopy (cryo-EM) has emerged over the past decade as a transformative technology in structural biology, enabling the structure determination of large macromolecular complexes at near atomic level resolution [5]. In contrast to existing techniques for structure determination such as NMR spectroscopy which is limited to small, single proteins and X-ray crystallography which requires crystallization of the protein of interest, cryo-EM has succeeded in resolving high quality structures of large protein complexes, such as the molecular machinery that implement the central dogma of molecular biology [1, 6, 3, 10, 9]. A notable milestone was reached in 2015 with the publication of a cryo-EM structure below 3 Å resolution [2] – a threshold below which building atomic models is possible, enabling applications such as rational drug design and therapeutic development [7]. Exemplary of the rapid pace of progress in the field, just a few years later, cryo-EM broke the so-called atomic resolution barrier with the publication of 1.2 Å resolution structures, allowing for direct visualization of atom positions [11, 4].

Cryo-EM is however challenged by the computational task of analyzing the raw image data. A single particle cryo-EM imaging experiment captures on the order of $10^4 - 10^7$ two dimensional (2D) projections of the target biomolecular complex. Reconstruction algorithms combine these images to infer the underlying three dimensional (3D) structure, a challenging inverse problem. However, complications in reconstruction arise when instances of the molecule exhibit heterogeneity, a common occurrence for many protein complexes. Notably, this heterogeneity is of intrinsic interest to the structural biologist as it is intimately connected to function, and highlights the unique advantage of cryo-EM in studying dynamic molecular machinery. Unlike in crystallography however, where protein flexibility inhibits crystal formation, in cryo-EM, the bottleneck to structure determination for heterogeneous, flexible molecules is a largely computational task, termed the heterogeneity problem [8].

This thesis addresses a significant open problem in this field, namely reconstructing continuous forms of heterogeneity from single particle cryo-EM images. The heterogeneity problem has a certain open-endedness in that it challenges us to define precisely what the desired output of 3D reconstruction is – both from an algorithmic standpoint and in service of the structural biology practitioner. Indeed, existing approaches for heterogeneous cryo-EM reconstruction and heterogeneity analysis (described in Chapter 2) place often limiting assumptions on the model class of reconstructed structures. The approach taken here is to recast cryo-EM reconstruction as a (deep) generative modeling problem, and develop new function approximation techniques based on deep neural networks in order to learn continuous distributions of protein structures from cryo-EM data.

This works sits at an intersection of many disciplines including statistical inference, deep learning, computer vision, and structural biology. Chapter 2 contextualizes this work from the cryo-EM perspective, including a historical background of cryo-EM and a review of the theory of single particle cryo-EM data analysis, and from the standpoint of modern machine learning, with a brief overview of generative modeling techniques and related methods in computer vision.

Chapter 3 introduces the main contribution of this thesis: a neural method, cryoDRGN (Deep Reconstructing Generative Networks), for heterogeneous cryo-EM reconstruction. CryoDRGN leverages the expressive, flexible representation power of neural networks to reconstruct continuous distributions of protein structure from cryo-EM images. We propose a new architecture for modeling cryo-EM volumes using dense, fully-connected neural networks with sinusoidal featurization of input coordinates, and a learning algorithm to optimize this representation from unlabeled 2D cryo-EM images. A particular challenge in heterogeneous reconstruction is the joint inference of extrinsic projection direction (i.e. camera pose) and intrinsic structural heterogeneity of the imaged particle. We address this challenging optimization task using a combination of techniques: an explicit disentangling of pose variables and latent variables through our architecture, a branch and bound accelerated search algorithm for pose, and optionally, a proposed modification to cryo-EM imaging (i.e. tilt series pairs) to encourage representation learning.

As much of the preliminary work was performed on synthetic datasets, the next main thrust of this thesis was to extend cryoDRGN to work on real cryo-EM datasets. We demonstrated that the cryoDRGN neural architecture is capable of modeling high-resolution cryo-EM density maps from real data and used the method to uncover new structures and dynamics from previously published datasets. The bulk of these results are described in Chapter 4. Integral to this extension to real application settings was the release of a software package around this method. We established best practices for training the model and developed a suite of analysis tools to visualize the manifold of conformations, generate density maps for exploratory analysis, extract particle subsets for use with other tools, and generate trajectories to visualize molecular motions. CryoDRGN is open-source software freely available at `cryodrgn.csail.mit.edu`.

In the previous chapter, high-quality reconstructions on real data were achieved by modifying cryoDRGN to use fixed poses previously estimated from an upstream homogeneous reconstruction, hence eliding the difficult pose search procedure. However, by estimating the intrinsic structural heterogeneity separately from the extrinsic pose variables, cryoDRGN and other extant methods for heterogeneous reconstruction are limited to mildly heterogeneous conditions where pose inference remains accurate. In developing cryoDRGN2, we revisited the problem of *de novo* joint optimization of image pose and volumes – also termed *ab initio* heterogeneous reconstruction. In Chapter 5, we present techniques for pose estimation when optimizing neural models of volumes, enabling state of the art *ab initio* reconstruction of protein structures from real datasets. Our goal with cryoDRGN2 is to broaden the scope of single particle cryo-EM to new classes of heterogeneous proteins and protein complexes.

CryoDRGN, cryoDRGN2, and other contemporary heterogeneous cryo-EM reconstruction algorithms have provided new capabilities in modeling distributions of protein structures. However, interpreting the results of cryoDRGN models is challenging and open-ended, and largely based on expert-driven, manual inspection. In Chapter 6, we present an efficient and automated volume analysis framework for comprehensively characterizing the learned distribution of density maps reconstructed by cryoDRGN.

In Chapter 7, we end on a proof-of-concept study on exploiting prior information provided by an atomic model to reconstruct distributions of 3D structures from a cryo-EM dataset. Although reconstruction algorithms typically model the 3D volume as a generic function parameterized as a voxel array or neural network, in this chapter, we use the underlying atomic structure of the protein of interest to place well-defined physical constraints on the reconstructed structure. Although the reconstruction objective is highly non-convex when formulated in terms of atomic coordinates (similar to the protein folding problem), we show that gradient descent-based methods can reconstruct a continuous distribution of atomic structures when initialized from a structure within the underlying distribution.

This thesis has introduced a new paradigm for protein structure determination with cryo-EM, bridging machine learning with 3D reconstruction to resolve complex distributions of heterogeneous protein structures. Progress in both machine learning for protein (structure) problems and in cryo-EM technology is advancing at incredible speed and foreshadows incredible future progress at this intersection. We conclude with an outlook on the field and discussion of future research directions in Chapter 8.

Bibliography

[1] Alexey Amunts, Alan Brown, Xiao-Chen Bai, Jose L Llácer, Tanweer Hussain, Paul Emsley, Fei Long, Garib Murshudov, Sjors H W Scheres, and V Ramakrishnan. Structure of the yeast mitochondrial large ribosomal subunit. *Science*, 343(6178):1485–1489, March 2014.

[2] Melody G Campbell, David Veesler, Anchi Cheng, Clinton S Potter, and Bridget Carragher. 2.8 å resolution reconstruction of the thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *Elife*, 4, March 2015.

[3] Xizi Chen, Xiaotong Yin, Jiabei Li, Zihan Wu, Yilun Qi, Xinxin Wang, Weida Liu, and Yanhui Xu. Structures of the human mediator and mediator-bound preinitiation complex. *Science*, 372(6546), June 2021.

[4] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia M G E Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, Tomasz Uchański, Lingbo Yu, Dimple Karia, Evgeniya V Pechnikova, Erwin de Jong, Jeroen Keizer, Maarten Bischoff, Jamie McCormack, Peter Tiemeijer, Steven W Hardwick, Dimitri Y Chirgadze, Garib Murshudov, A Radu Aricescu, and Sjors H W Scheres. Single-particle cryo-EM at atomic resolution. *Nature*, 587(7832):152–156, October 2020.

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

[6] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, jun 2017.

[7] Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-Ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.*, 17(7):471–492, July 2018.

[8] Amit Singer and Fred J. Sigworth. Computational Methods for Single-Particle Electron Cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, jul 2020.

[9] Sameer Singh, Arnaud Vanden Broeck, Linamarie Miller, Malik Chaker-Margot, and Sebastian Klinge. Nucleolar maturation of the human small subunit processome. *Science*, 373(6560):eabj5338, September 2021.

[10] Chengyuan Wang, Vadim Molodtsov, Emre Firlar, Jason T Kaelber, Gregor Blaha, Min Su, and Richard H Ebright. Structural basis of transcription-translation coupling. *Science*, 369(6509):1359–1365, September 2020.

[11] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, November 2020.

Chapter 2

Background

This work is situated in the emerging area of machine learning for cryo-EM reconstruction, an interdisciplinary confluence of many different fields, including the many facets of cryo-EM methods (e.g. statistical inference, signal processing, electron optics), modern deep learning, computer vision, and structural biology. This chapter summarizes the relevant background upon which this work stands and contextualizes relevant related and contemporary work.

In Section 2.1, I overview the development of cryo-EM for protein structure determination, starting from the historical developments [12], through the recent "resolution revolution" [38], and up to the present "revolution evolution" [1] of the technique into the modern day aspirations of high-throughput, high-resolution cryo-EM.

The crux of structure determination with cryo-EM involves the 3D reconstruction of a density volume (or volumes) from an experimentally-derived dataset of projection images. In Section 2.2, I overview the theory behind single particle cryo-EM reconstruction, and the existing algorithms and software. While the homogeneous reconstruction task is well-defined, the identification and analysis of heterogeneity in cryo-EM is an open problem at the heart of this thesis. I review the existing and contemporary approaches to address structural heterogeneity.

We then switch gears. This thesis leverages major recent advances in the function approximation capabilities of deep neural networks, which can now learn complex distributions from large-scale high-dimensional data. In Section 2.3, I present a high level background of machine learning, then overview the generative modeling tools central to this thesis work, specifically Variational Autoencoders [36].

Finally, this thesis has proceeded concurrently with an explosion of attention and research in neural field techniques in computer vision. In Section 2.4, I overview related techniques for 3D modeling in graphics and other visual computing applications that have striking overlap with the parametric and differentiable forward model for cryo-EM introduced here.

2.1 Cryo-EM

What is the use of it? Everything under the electron beam would burn to a cinder!
-Dennis Gabor, 1928

2.1.1 What is cryo-EM?

What is cryo-electron microscopy? Broadly defined, cryo-electron microscopy refers to the imaging of biological specimens at cryogenic temperatures with a transmission electron microscope[1]. There are many branches of cryo-EM. In modern high-resolution biological imaging, cryo-EM can refer to one of several imaging techniques: cryo-electron tomography (cryo-ET), single particle cryo-EM, or electron crystallography. The difference between these sub-disciplines stems from the different contexts of the target specimen, e.g. imaging *in situ* cellular and subcellular structures in cryo-ET or crystalline proteins or small molecules in electron crystallography. The focus of this thesis is on *single particle cryo-EM*, where high resolution structures of macromolecular complexes are determined through imaging a purified solution of the target molecule[2] of interest.

We start by painting a picture of the incredible technology of modern day single particle cryo-EM. A Zirconium oxide-coated, tungsten-tipped field emission gun emits

[1] Scanning electron microscopy techniques are beyond the scope of this thesis.

[2] Since the target "molecule" is often a complex of many biomolecules, they are also refereed to as "particles".

a beam of electrons with incredible spatial and temporal coherence – around 10^{-5-6} radians [79], accelerated through hundreds of keV of potential energy, focused through water-cooled electromagnetic lenses – exquisitely sensitive to even the temperature fluctuation induced from a human entering the room – to illuminate the sample grid hole that is around 10^{-6} meters in width, where a collection of organic atoms (i.e. mostly Cs, Ns, Os) sit frozen in a slightly more ordered configuration than the surrounding H_2O molecules. The whole operation proceeds in vacuum, as air molecules would scatter the electron beam, and at liquid nitrogen temperatures, where radiation damage is minimized relative to room temperature conditions [17].

The sample itself (post-biochemical production and purification) consists of many copies of the target of interest suspended in a thin layer of amorphous (vitreous) ice. The ice layer must be thin (typically around 100 nm) to enable transmission electron microscopy, i.e. reasonable passage statistics without inelastic scattering events [65]. The aqueous sample is fixed for imaging by plunge freezing a thin film of the sample in liquid ethane, rapidly cooling the sample to form amorphous, non-crystalline ice.

Most electrons pass through the sample uninterrupted, contributing to the shot noise in the recorded images. Some are inelastically scattered, transferring some of their kinetic energy into the sample, and thus contribute to radiation damage. The inelastically scattered electrons are filtered out before reaching the detector. The observed contrast in recorded cryo-EM images stems from the elastically scattered electrons, whose wavefunction experiences a phase shift upon passing through the sample that can be (indirectly) measured by the detector. The degree of phase shift is determined not only by the sample itself, but is also affected by frequency-dependent interference due to the microscope's electron optics. Mathematically, this microscope-dependent signal transfer can be derived from weak phase-object approximation theory as the contrast transfer function (CTF) (see Section 2.2).

Ultimately, the imaging process magnifies the object of interest by 5-6 orders of magnitude, capturing snapshots of single particles where the spatial scale of each pixel spans around 1 Å. Note that this resolution is much lower than the theoretical limit given by the wavelength of the electron beam (0.02 Å at 300 kV [45]) due to limitations

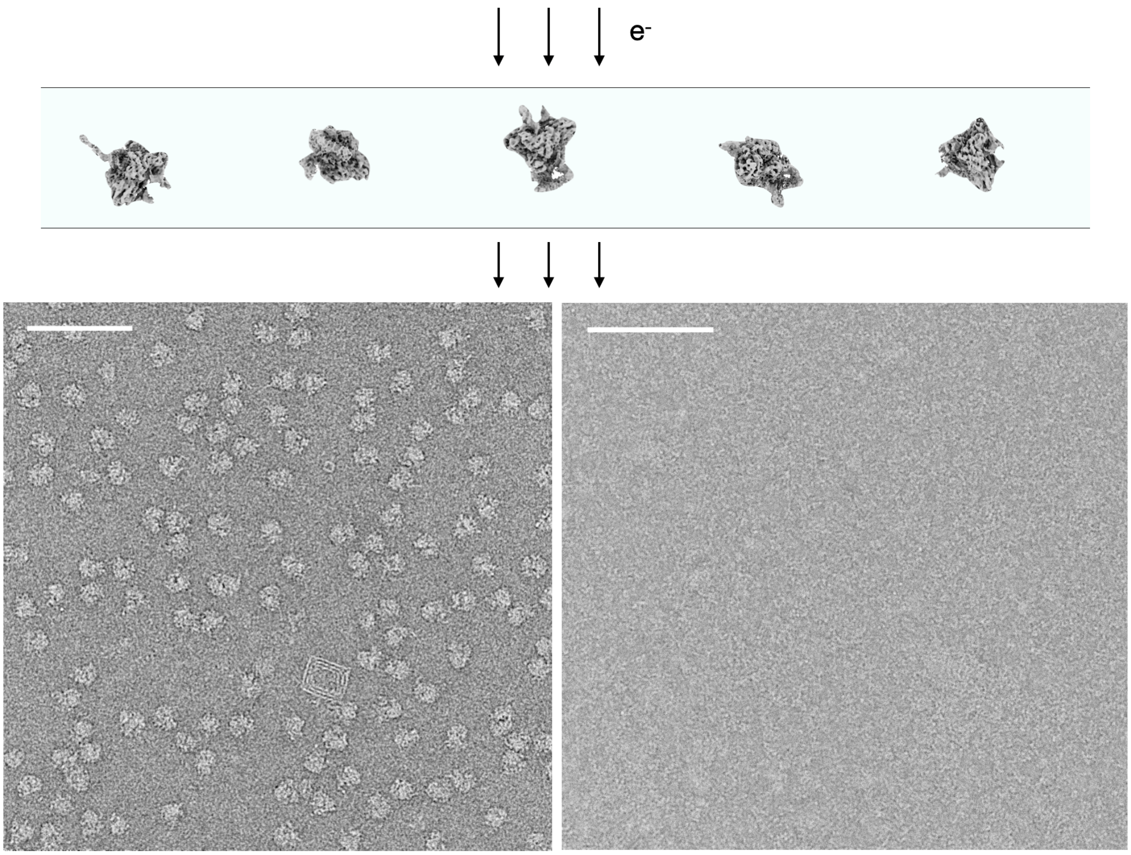


Figure 2-1: Schematic of single particle cryo-EM imaging (top) and example experimental micrographs (bottom). A purified sample of the molecule of interest is flash frozen in a thin layer of vitreous ice and imaged with a transmission electron microscope. The recorded 2D projection image contains many copies of the molecule captured from an unknown *pose*, or viewing direction, and in an unknown *conformational state*. Example micrographs of the 80S ribosome (EMPIAR-10028, 4.2 MDa, left) and the RAG1-RAG2 complex (EMPIAR-10049, 369 kDa, right) are from Wong et al. [93] and Ru et al. [70], respectively. The scale bar denotes 100 nm.

from radiation damage. Nevertheless, the resolving power of cryo-EM is much greater than that of light microscopy (where the resolution is limited by the wavelength of visible light to around 3000 Å), highlighting the ability of this technology at visualizing the details and intricacies of the molecular universe. After his original incredulity at the technology, many decades later, Dennis Gabor later recounted, "Who would have dared to believe that the cinder would preserve not only the structure of microscopic bodies but even the shapes of organic molecules?" [25].

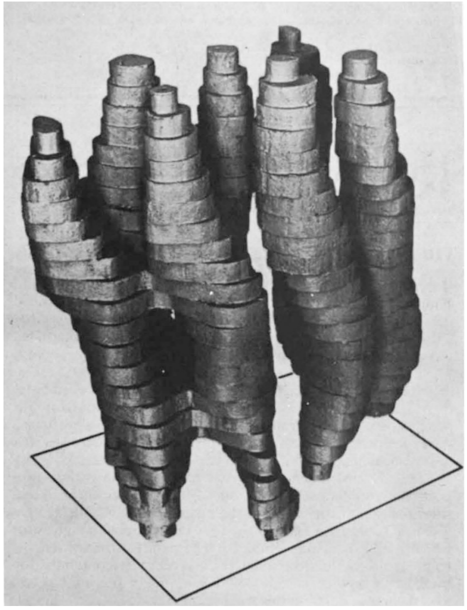
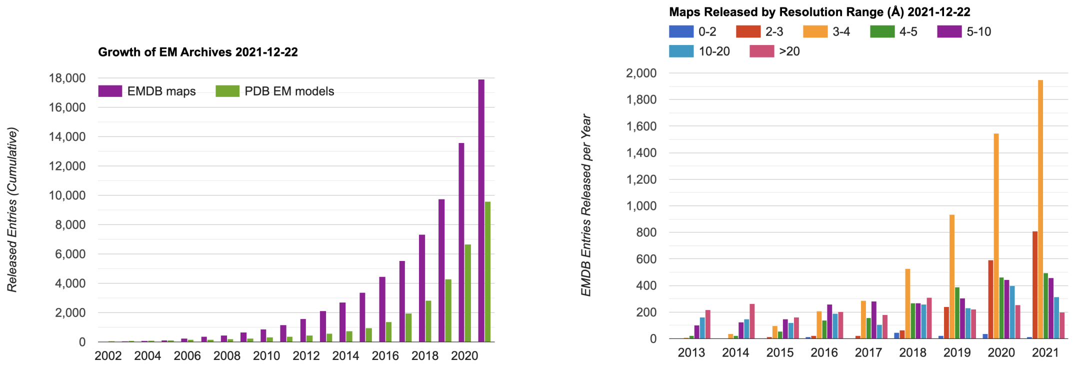


Figure 2-2: Model of the first protein structure solved at 7 Å resolution by electron microscopy of a crystal of bacteriorhodopsin by Richard Henderson and Nigel Unwin. Figure adapted from Henderson and Unwin [27].

2.1.2 How did we get here?

The development of cryo-electron microscopy from a specialized technique pioneered by a few groups into a mainstream structural biology tool has proceeded over several decades, starting from the first electron microscope developed by M. Knoll and E. Ruska in 1931[3] [71, 37], and culminating in the 2017 Nobel Prize in Chemistry to Jacques Dubochet, Richard Henderson, and Joachim Frank "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution" [52]. We summarize a few of the milestones in the field that culminated with this recognition. The earliest demonstration of a 3D protein structure solved with cryo-EM was the 7 Å structure of bacteriorhodopsin (Fig. 2-2), published in 1975 by Richard Henderson and Nigel Unwin [27]. This first cryo-EM derived protein structure is exemplary of electron crystallography, as the sample was fixed in a crystalline lattice. To realize the potential of cryo-EM at visualizing the structure of these biomolecules in their hydrated, aqueous form, Ken Taylor and Robert Glaeser proposed fixing the

[3] Ernst Ruska received the 1986 Nobel Prize in Physics "for his fundamental work in electron optics, and for the design of the first electron microscope."[51]



(a) EMDB deposition statistics over time.

(b) EMDB deposition statistics over time striated by map resolution.

Figure 2-3: Growth of cryo-electron microscopy. Accessed from https://www.emdataresource.org/statistics.html on December 22nd, 2021.

sample in amorphous ice for cryo-EM imaging without crystallization [85]. Later work by Jacques Dubochet and colleagues developed a practical method for rapid vitrification by plunge freezing the sample in liquid ethane [18, 17]. Without the diffraction signal originating from an ordered crystalline lattice, new algorithms were required to combine the resulting cryo-EM images of randomly oriented particles; Joakim Frank and colleagues pioneered these image processing algorithms that have since created the field of single particle reconstruction. Together these advances enabled the first near-atomic resolution protein structure solved with cryo-EM published in 1990 by Richard Henderson and his colleagues [26], demonstrating the potential of cryo-EM for structural biology.

The pace of progress has yet to slow since these seminal contributions. Advances in cryo-EM technology in the past decade have led to a marked increase in the speed, quality, and adoption of the technique (Figure 2-3), transforming the field from medium-to-low resolution "blob-ology" to a state where collecting high resolution density maps sufficient for atomic model building (3-4 Å) is now routine (for certain classes of well-behaved targets).

This new era of cryo-EM, dubbed the "resolution revolution" [38], has been primarily driven by the introduction of direct electron detector (DED) cameras. With this new camera technology, more high resolution information is captured compared to the

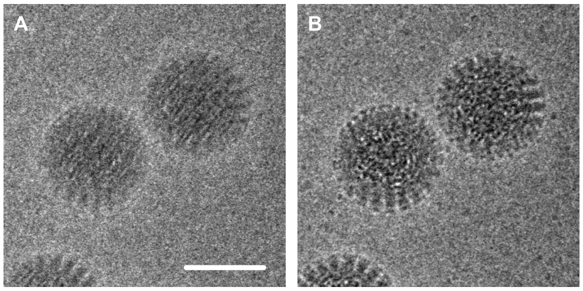


Figure 2-4: Direct electron detectors allow image collection in movie mode at high frame rates (here, 40 frames per second), which can be used to correct for beam-induced motion of the particles within the ice. 60 frames are averaged together without (left) and with (right) motion correction. The scale bar denotes 50 nm. Figure adapted from Brilot et al. [9].

previous generation of charge coupled device (CCD) cameras, which must convert the incoming electron signal to photons via a scintillator. Moreover, DEDs are also capable of imaging at much faster frame rates; This new "movie mode" of imaging allows for the computational correction of beam-induced motion of the particles within the ice, de-blurring the collected micrographs, and drastically improving the underlying signal captured in the raw images (Figure 2-4). There have been many subsequent hardware improvements that continue to improve the quality of the images, recently enabling true atomic resolution structure determination where atom positions are directly visualized (1.2 Å) as opposed to inferred from the known geometry of amino acids in the 3-4 Å regime [95, 48].

Many algorithmic and software developments have gone hand in hand with these hardware improvements, driving the speed and automation, and thus the accessibility and adoption of cryo-EM. We note a few major algorithmic innovations to the reconstruction process here and save the technical details for Section 2.2. In 2013, Sjors Scheres proposed a Bayesian formulation of cryo-EM reconstruction [74] implemented in the RELION software package [75], which posed reconstruction as optimization of a single statistical model, thus eliminating or formalizing many of the heuristics in previous protocols. Brubaker et al. [10] proposed an algorithm for *ab initio* reconstruction based on stochastic gradient descent (SGD), which was later accelerated with a branch and bound algorithm by Punjani et al. and implemented in the cryoSPARC software package [58]. The improved global convergence properties of SGD-based reconstruction could successfully reconstruct structures from random initialization across a broad range of systems, providing a data-driven method for producing (low-resolution) initial models for refinement instead of needing to rely on homology models or negative stain EM. As no biomolecular complex is truly static, multiclass reconstruction algorithms were proposed to sort heterogeneity into a few discrete structures [73]. These algorithms, also referred to as “3D classification” or “heterogeneous refinement”, have been implemented in many cryo-EM software packages and have become the main workhorse for processing heterogeneous datasets. Finally, the rise of GPU compute has significantly accelerated image processing, e.g. reconstruction workflows that previously took weeks to months on CPU can now be processed in a few hours on a single GPU workstation.

2.1.3 The computational pipeline

Computational processing of single particle cryo-EM datasets can be broken down into several distinct stages: 1) micrograph preprocessing, 2) 3D reconstruction, and 3) atomic model building. We briefly summarize these processing steps. The reader is referred to Singer and Frank [79] for further details on the single particle cryo-EM image processing pipeline.

First, the noisy micrograph image is preprocessed and then segmented, where bounding boxes containing individual molecules (i.e. particles) are identified and extracted (i.e. particle picking). Preprocessing steps include beam-induced motion correction [2, 22] and estimation of CTF parameters [67, 96]. There are numerous particle picking algorithms, including methods based on templates [50, 32, 90, 76] and new deep learning-based detection algorithms [92, 99, 91, 4]. Once particles are extracted, they are typically inspected and optionally filtered with 2D classification [77], a clustering algorithm which sorts images into common viewing directions. Particles imaged from consistent viewing directions will be averaged together and can be visually inspected for quality and structural features of interest; clusters that exhibit obvious junk (e.g. false positives from particle picking) can be removed.

Next, the stack of extracted single particle images is reconstructed to yield a 3D cryo-EM density volume (*homogeneous reconstruction*) or a distribution of volumes (*heterogeneous reconstruction*). The theory describing this procedure is described in the next section. In practice, this is typically a multi-stage procedure, in particular to deal with heterogeneity. Processing often involves multiple rounds of 3D classification on subsets of the images, returning to the micrographs to re-pick or re-extract particles, and to re-estimate CTF parameters and motion correction ("particle polishing") (see Figure 2-6 for an example workflow). Finally, once a suitable structure is obtained, it may be post-processed to facilitate interpretation, e.g. sharpening high resolution features with B-factor correction [69] or deep learning models [72] and local resolution estimation.

Finally, given a sufficiently high resolution reconstruction, an atomic model is built into the resulting 3D volume either *de novo* or fit starting from a related model, or more recently, from an AlphaFold prediction [33], typically with the help of various automated or interactive tools [19, 42, 87, 34, 78, 13, 55].

2.2 Single particle reconstruction

The task in single particle cryo-EM reconstruction is to recover the 3D structure (also called a *density map* or *volume*) of the target protein or biomolecular complex of interest, defined as the function $V : \mathbb{R}^3 \to \mathbb{R}$, where $V(\mathbf{x})$ gives the electron scattering potential at a point in space $\mathbf{x} \in \mathbb{R}^3$. V is estimated from a set of noisy projection images $X_1, ..., X_N$, each containing a copy of V projected in some unknown orientation (Figure 2-5).

2.2.1 Image Formation Model

In the idealized, homogeneous scenario, many *identical* copies of the molecule of interest are rapidly frozen in a thin layer of vitreous ice. As the molecules are tumbling randomly in solution before being flash frozen, the recorded images $X_1, ..., X_N$ each capture a 2D projection of V in an unknown orientation relative to the imaging axis

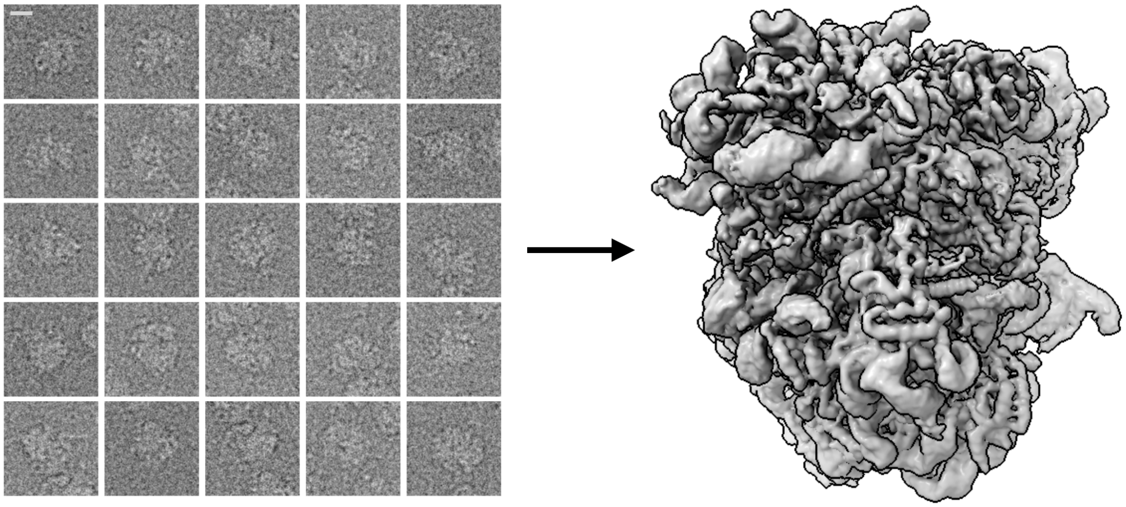


Figure 2-5: Single particle cryo-EM reconstruction algorithms tackle the inverse problem of determining a 3D density volume or a distribution of volumes from a dataset of $10^4 - 10^7$ recorded projection images. Each image is a noisy projection of a unique instance of the molecule (particle) suspended in ice at a random orientation. *Homogeneous* reconstruction algorithms must jointly learn the density volume V and the pose of each image $\phi_i \in SO(3) \times \mathbb{R}^2$. *Heterogeneous* reconstruction algorithms take into account structural variation between the particles and aim to reconstruct a distribution of volumes. *Left:* Example cryo-EM images of the *Plasmodium falciparum* 80S ribosome from EMPIAR-10028 [93]. *Right:* A density volume reconstructed by cryoDRGN from this dataset, visualized as an isosurface contour. The scale bar denotes 100 Å, and the size of the 80S ribosome is approximately 250 Å in length.

(Figure 2-1). The generation of image $X : \mathbb{R}^2 \to \mathbb{R}$ can be modeled as:

$$X(r_x, r_y) = g * T_t \int_{\mathbb{R}} V(R^T \mathbf{r})\ dr_z + noise \qquad \mathbf{r} = (r_x, r_y, r_z)^T \tag{2.1}$$

where V is the density volume, $R \in SO(3)$, the 3D rotation group, is an unknown orientation of the volume, and T_t is an unknown in-plane translation by $t \in \mathbb{R}^2$, corresponding to imperfect centering of the volume within the image. The image signal is convolved with g, the point spread function for the microscope before being corrupted with additive noise and registered on a discrete grid of size $D \times D$, where D is the size of the image along one dimension.

2.2.2 Fourier Slice Theorem

Many reconstruction algorithms use the *Fourier slice theorem* in three dimensions [7], which states that the Fourier transform of a 2D projection of V is a 2D slice through the origin of V in the Fourier domain. The slice is perpendicular to the axis of projection. Mathematically, the theorem is written as:

$$\mathcal{F}_2 P_2 = \mathcal{S}_2 \mathcal{F}_3 \tag{2.2}$$

where $\mathcal{F}_n$ is the Fourier transform in n dimensions, P_2 is the projection operator from three to two dimensions, and $\mathcal{S}_2$ is the slice operator, which extracts a 2D central slice perpendicular to the axis of projection.

In the Fourier domain, the generative process for image $\hat{X}$ from volume $\hat{V}$ can thus be written:

$$\hat{X} = \hat{g} S_t A_R \hat{V} + \epsilon \tag{2.3}$$

where $\hat{g} = \mathcal{F} g$ is the contrast transfer function (CTF), S_t is a phase shift operator corresponding to image translation by t in real space, and $A_R \hat{V} = \hat{V}(R^T(\cdot, \cdot, 0)^T)$ is a linear slice operator corresponding to rotation by R and linear projection along the z-axis in real space. The frequency-dependent noise ϵ is typically modelled as independent, zero-centered Gaussian noise in Fourier space. Under this model, the probability of of observing an image $\hat{X}$ with pose $\phi = (R, t)$ from volume $\hat{V}$ is:

$$p(\hat{X}|\phi, \hat{V}) = p(\hat{X}|R, t, \hat{V}) = \frac{1}{Z} \exp\left(\sum_{l} \frac{-1}{2\sigma^2} \left| \hat{g}^{(l)} A_R^{(l)} \hat{V} - S_t^{(l)} \hat{X}^{(l)} \right|^2 \right) \tag{2.4}$$

where l is a two-component index over Fourier coefficients for the image, σ is the width of the Gaussian noise, and Z is a normalization constant.

2.2.3 Contrast Transfer Function

The contrast transfer function (CTF) $\hat{g} : R^2 \rightarrow \mathbb{R}$ is a two-dimensional, real-valued function of the spatial frequency vector $\mathbf{k} = (k_x, k_y)$ that approximates the signal transfer as an electron beam travels through the microscope and passes through a weak-phase object. It describes the phase shift incurred by the microscope's abberations, including defocus settings $(z1, z2, z_\theta)$ and spherical aberration C_s. A derivation is provided in [79]. A commonly used form of the CTF is written as:

$$\hat{g}(\mathbf{k}) = \sqrt{(1 - w^2)} \sin(\gamma(\mathbf{k})) - w * \cos(\gamma(\mathbf{k})) \tag{2.5}$$

where w is the amplitude contrast ratio, and

$$\gamma(\mathbf{k}) = 2\pi \left(-\frac{z(\mathbf{k})\lambda|\mathbf{k}|^2}{2} + \frac{C_s \lambda^3 |\mathbf{k}|^2}{4} \right) - \phi_0 \tag{2.6}$$

where λ is the wavelength of the electron, C_s is the spherical aberration of the objective lens, ϕ_0 is any additional phase shift from a phase plate, and $z(\mathbf{k})$ describes the defocus of the microscope:

$$z(\mathbf{k}) = 0.5 * (z_1 + z_2 + (z_1 - z_2)\cos(2 * (\theta - z_\theta)); \theta = \arctan^2(k_y/k_x) \tag{2.7}$$

which is parameterized by the defocus parameters, z_1 and z_2, and the defocus astigmatism, z_θ. Finally, an envelope function models the attenuation of the signal at high frequencies:

$$\hat{g}'(\mathbf{k}) = \hat{g}(\mathbf{k})e^{(-B|\mathbf{k}|/4)} \tag{2.8}$$

which is parameterized by the B-factor B. These parameters are determined by the microscopy settings, some of which are set and others are estimated by function fitting software prior to 3D reconstruction. In the high resolution regime ($<$ 2.5 Å), higher order terms of the CTF must be modeled [100].

2.2.4 Homogeneous cryo-EM reconstruction

To recover the desired structure, cryo-EM reconstruction methods must jointly solve for the unknown volume V and image poses $\phi_i = (R_i, t_i)$. Most algorithms use Expectation Maximization [74] or simpler variants of coordinate ascent to find a *maximum a posteriori* estimate of V marginalizing over the posterior distribution of ϕ_i's, i.e.:

$$V^{\text{MAP}} = \arg\max_V \sum_{i=1}^{N} \log \int p(X_i|\phi, V)p(\phi)d\phi + \log p(V) \tag{2.9}$$

Intuitively, given $V^{(n)}$, the estimate of the volume at iteration n, images are first aligned with $V^{(n)}$ (E-step), then with the updated alignments, the images are backprojected to yield $V^{(n+1)}$ (M-step).

This procedure, termed iterative refinement, is a local optimization procedure whose result depends on the initial model $V^{(0)}$. In practice, reconstruction typically proceeds in two stages: 1) generation of a low-resolution initial model $V^{(0)}$, and 2) iterative refinement of the initial model to high resolution.

Generation of an initial model

Initial structures can be obtained experimentally [41], inferred based on homology to complexes with known structure, or via *ab-initio* reconstruction with stochastic gradient descent [10, 58]. Punjani et al. found that SGD was able to avoid errant local minima while optimizing the highly nonconvex objective of Equation 2.9 [58]. This approach provides a data-driven method for determining an initial model, though optimization can still fail to converge to the correct local minima, especially for datasets with significant structural heterogeneity.

2.2.5 Resolution estimation and validation

As 3D reconstruction tackles an inverse problem with no ground truth, a major challenge for the field is validation of the final density map(s) [28]. A particular worry stems from bias from the initial model and producing artifacts in the volume due to the high noise level in the images and the local convergence properties of iterative refinement. While validation remains an open problem (especially for heterogeneous distributions), it has been somewhat mitigated by entering a higher resolution regime where the final density map may be checked against our prior knowledge on the known geometry of the constitutive amino acids.

The current standard in the field is a two-fold cross validation approach based on the Fourier Shell Correlation (FSC) curve. The FSC curve measures correlation between volumes as a function of spherically-averaged radial shells in Fourier space:

$$FSC(k) = \frac{\sum_{s \in S_k} \hat{U}_s \hat{V}_s^*}{\sqrt{(\sum_{s \in S_k} |\hat{U}_s|^2)(\sum_{s \in S_k} |\hat{V}_s|^2)}} \tag{2.10}$$

where $\hat{U}$ and $\hat{V}$ are the two volumes being compared and S_k is the set of voxels in a spherical shell at distance k from the origin.

The gold standard FSC (GSFSC) metric for validation and resolution estimation is to independently reconstruct random halves of the dataset and compute the FSC curve between the aligned structures [28]. The FSC curve is used as a diagnostic towards consistency of the reconstructed structures; Resolution is often reported as $1/k_0$ where $k_0 = \arg\max_k FSC(k) < C$ and C is some fixed threshold ($C = 0.143$ for the GSFSC). Resolution can also estimated by computing the FSC curve between the final reconstructed structure (using all the data) and a density map simulated from the final atomic map with the threshold $C = 0.5$. Typically the background is zeroed out before computing the FSC curve to avoid spurious correlations from any background solvent density.

2.2.6 Heterogeneous cryo-EM reconstruction

The ability to image conformationally and compositionally heterogeneous biomolecular complexes is a major advantage of cryo-EM compared to other techniques for structure determination, such as X-ray crystallography, which requires structural homogeneity for crystallization, and NMR spectroscopy, which produces an ensemble measurement. However, structural heterogeneity poses a significant challenge in computational image processing[4], termed *the heterogeneity problem* [79].

Heterogeneous reconstruction algorithms relax the assumption in the image formation model that each image captures an identical, static structure, and instead aim to recover a distribution of structures. To make the problem well-posed, heterogeneous reconstruction algorithms impose some model class on the reconstructed distribution. Methods can be roughly divided into those that model the imaged structures with a *discrete* model for heterogeneity, and those that parameterize some *continuous* model for heterogeneous structures.

Multiclass refinement

Given sample heterogeneity, the standard approach in the cryo-EM field is to simultaneously reconstruct K independent volumes, a *discrete* model of conformational heterogeneity. Termed *multiclass refinement* (or 3D classification), the image formation model is extended to assume that images are generated from $V_1, ..., V_K$ independent volumes, with inference now requiring marginalization over ϕ_i's and class assignment probabilities π_j's:

$$\underset{V_1,...,V_K}{\arg\max} \sum_{i=1}^{N} \log \sum_{j=1}^{K} \left(\pi_j \int p(X_i|\phi, V_j) p(\phi) d\phi \right) + \sum_{j=1}^{K} \log p(V_j) \tag{2.11}$$

[4] sometimes requiring the practitioner to revisit sample preparation to experimentally rigidify the system (e.g. with chemical crosslinking agents or by genetically modifying the target biomolecule)

While this formulation is sufficiently descriptive when the structural heterogeneity consists of a small number of discrete conformations, it suffers when the heterogeneity is complex or when conformations lie along a continuum of states.

In practice, resolving such heterogeneity is handled through a hierarchical approach refining subsets of the imaging dataset with manual choices for the number of classes and the initial models for refinement. Many rounds of expert-guided processing are typically performed with manual selection of K and the initial volumes for refinement (an example is given in Figure 2-6). Because the number and nature of the underlying structural states are unknown, multiclass refinement is error-prone, and often leads to overfitting of some conformations, completely missing other states, or producing blurring artifacts from averaging together continuous forms of heterogeneity. Furthermore, there exists an upper limit to the number of resolvable structures in this hierarchical approach as each class much have a minimum number of images to provide sufficient views for 3D reconstruction and to produce high resolution structures (i.e. a bias-variance trade-off).

Masked refinement

A class of techniques called “focused refinement” or “focused classification” use masking to focus on specific regions of the structure: poses are estimated in the (E)-step of iterative refinement by aligning images to the masked volume. These methods require a user-defined mask on the region of interest from an upstream consensus homogeneous reconstruction. This method encodes the assumption that the region of focus is rigid, though its orientation may change relative to the rest of the complex. In practice there are SNR limitations in focusing on regions that are too small to provide enough signal for alignment.

In the standard focused refinement approach, there is also a mismatch between the image, which contains the full complex, and the estimate of the volume, which only contains the masked region. To address this mismatch, “focused refinement with signal subtraction” was introduced in Bai et al. [3]. A second mask is defined around the region of the volume that should be *excluded* in alignment, e.g. a micelle in

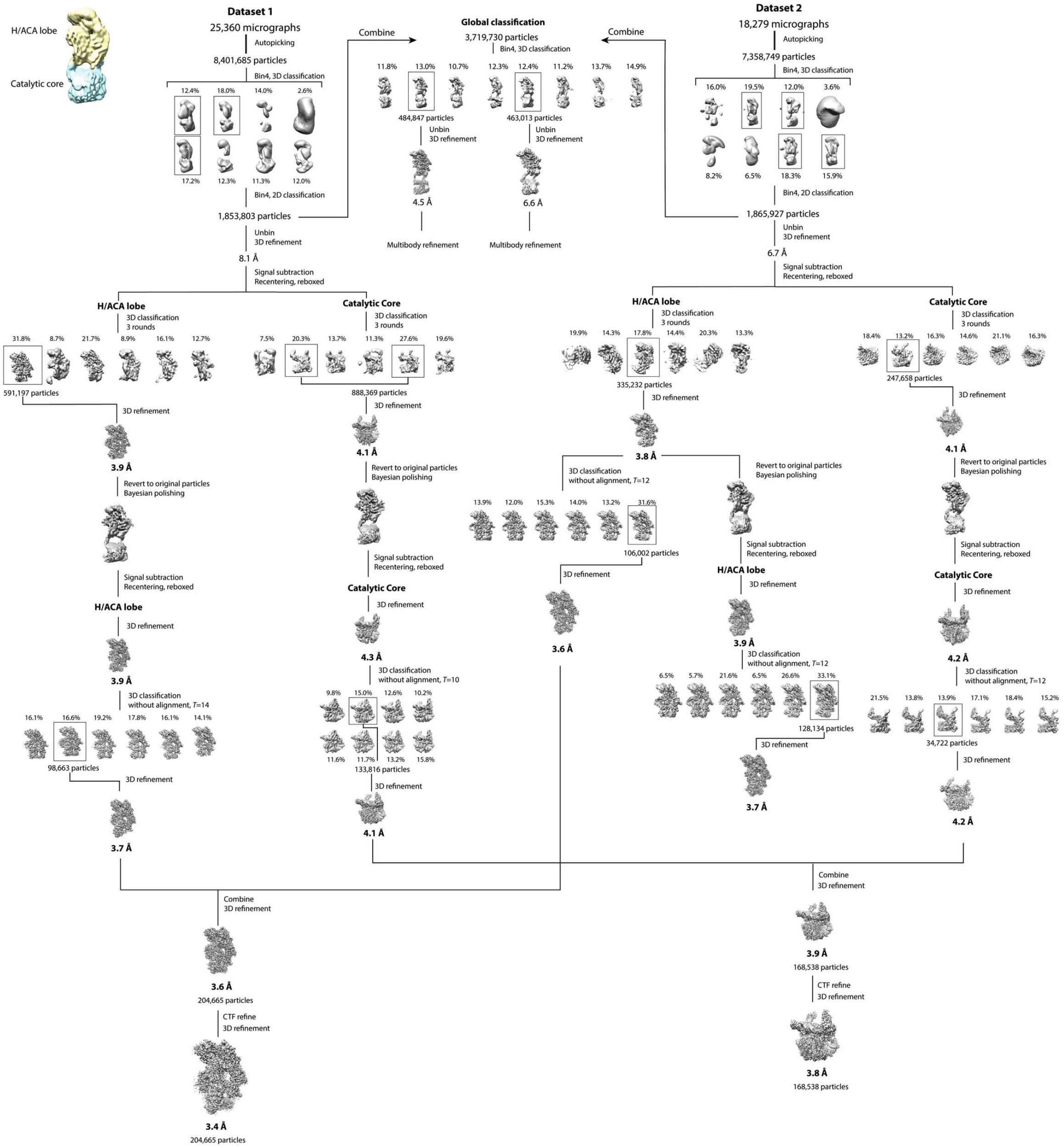


Figure 2-6: An example cryo-EM reconstruction workflow for a heterogeneous dataset of human telomerase holoenzyme involving many rounds of 3D classification. The initial dataset contained 3,719,730 images and the final processed dataset yielded 2 structures from a filtered set of 373,203 images. Figure adapted from Ghanim et al. [21].

a membrane protein complex. The density defined within this second mask from the consensus reconstruction is then computationally deleted from the images (after projection). This procedure also assumes that the region to remove is relatively rigid and well-estimated by the consensus reconstruction.

Continuous heterogeneity

Despite their demonstrated utility, discrete approaches are not well suited to datasets where there are continuous motions in the macromolecule of interest. Recent works have proposed methods for modeling continuous heterogeneity in cryo-EM reconstruction. Many of these methods extend the image formation model to model the volume as some function of a continuous latent variable z.

Nakane et al. [47] propose multibody refinement, where flexible structures are modeled as the sum of b rigid bodies. In this case, the inferred latent z corresponds to the relative position and translation of the b bodies. Rigid bodies are initially defined from a homogeneous reconstruction, thus imposing specific structural assumptions on the exhibited heterogeneity. Dashti et al. [20] learn a continuous embedding of images with diffusion maps on the 2D images. Their approach requires grouping images by projection direction before manifold estimation. Theoretical work for continuous heterogeneous reconstruction includes expansion of discrete 3D volumes in a basis of Laplacian eigenvectors [46] and a general framework for modelling hyper-volumes [39] e.g. as a tensor product of spatial and temporal basis functions [40].

Many methods have sought to characterize the variance-covariance matrix of the 3D volumes [43, 54, 83], recently popularized with the development of 3D Variability Analysis (3DVA) [57]. In 3DVA, the density map is discretized as a D^3 voxel array, and the space of conformations is restricted to a rank-N subspace of $\mathbb{R}^{D^3}$. Thus, 3DVA simultaneously fits a low-rank latent z and a linear mapping $f(z) : \mathbb{R}^N \to \mathbb{R}^{D^3}$ from this latent to discretized volumes. 3DVA has since been used extensively as it is straightforward to use and interpret, runs quickly, and is integrated as a tool within the cryoSPARC software package [58]. However, since 3DVA fits a linear model of continuous heterogeneity, the underlying motions of the system are approximated as linear interpolations between eigenvolumes, which can produce non-physical artifacts (e.g. motions approximated by mass appearing and disappearing). To avoid artifacts from the linear interpolations, images may be binned along the inferred reaction coordinates for a traditional homogeneous refinement, however this approach is not guaranteed to capture the underlying conformation and suffers from a bias-variance tradeoff.

Since the development of cryoDRGN, several deep learning approaches have been proposed for heterogeneous cryo-EM reconstruction, including e2gmm [11], and 3D-flex [56], which propose alternative neural parameterizations of the forward model. Yet other deep learning methods explore alternate paradigms for reconstruction, for example with adversarial learning [23, 24], learning poses with autoencoders [68, 49], and incorporating atomic models into reconstruction [98, 68, 35]. Bepler et al. propose a coordinate-based VAE architecture for modeling 2D continuous motions in negative stain EM images [5]. All the methods described in this section require previously estimated image poses. See Chapter 5 for an overview of related work in *ab initio* heterogeneous reconstruction, and see Chapter 7 for an overview of related work in including atomic models in 3D reconstruction.

2.3 Machine Learning

In this section, we provide a brief background on the trends and techniques in machine learning that are relevant to this thesis. The recent rise of cryo-EM has paralleled the huge growth of deep learning methods and applications. This growth has been driven by improvements in the capabilities of deep neural networks and optimization techniques to model complex functions, the availability of ever-larger training datasets, and the rise of GPU compute and easy to use software frameworks around deep learning. At its core, neural network-based models (i.e. neural models) can now be stably trained on large-scale, high-dimensional data via gradient descent using backpropagation. While a detailed review is beyond the scope of this thesis, a key trend is that deep learning has become the dominant paradigm for machine learning across many data modalities relevant to our interaction with the natural world, e.g. vision and natural language.

2.3.1 Supervised vs. Unsupervised Learning

Within machine learning, a distinction exists between supervised and unsupervised learning tasks. Supervised learning refers to the setting where the goal is to predict the value of a target variable y given some observed variable x from a dataset of (x, y) pairs. This problem is denoted *regression* if the target variable y is continuous or *classification* if y is discrete. Examples of supervised learning tasks include predicting a person's likelihood of a heart attack, classifying images, or picking particles in a cryo-EM micrograph. For all these tasks, a dataset of labeled examples is required to train a model. Many of the successes of deep learning are in applications where there exist large corpuses of labeled data, where a supervised learning task can be posed as learning some function $f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^R$ parameterized by a deep neural network, (e.g. images to probability of belonging to some class).

In unsupervised learning, the goal is to infer some unknown, or latent, characteristic from observed data, x, but without access to any ground truth or assigned labels. Some classical examples include clustering (inferring cluster labels) and dimensionality reduction (inferring a lower-dimensional representation). Intuitively, these algorithms encode some assumptions about the nature of the data, either implicitly or explicitly, and learn common patterns or features in a data-driven fashion. Often times, an unsupervised learning problem can be posed as a generative modeling task, e.g. specifying a probabilistic formulation of the data generation process $p_\theta(x)$, and fitting the parameters θ to approximate the observed data. Research in the field of deep generative models is especially active, which we describe next.

2.3.2 Generative modeling

We are interested in modeling our data x as random variables which follow some *unknown*, underlying distribution $p(x)$.

$$x \sim p(x)$$

Generative models are a family of machine learning methods that define a model $p_\theta(x)$ of the data generation process and aim to learn its parameters θ to best approximate $p(x)$.

$$p_\theta(x) \approx p(x)$$

There are many reasons why this is desirable and a wide variety of possible use cases and downstream applications. If the generative model is structured to reflect the underlying data generating process, the parameters and latent variables may be directly informative, e.g. the topic clusters in LDA [6], or the generated volumes in cryoDRGN. Second, accurately fitting such a model may reveal low-dimensional, latent structure of the data; or most eponymously, a generative model may be used to generate new samples of x.

The literature on generative modeling, encompassing both classical and modern deep-learning based approaches, is vast with rapid ongoing development. Here we highlight two broad classes of likelihood-based approaches to modeling the data distribution $p(x)$, autoregressive models and latent variable models[5]

Auto-regressive models decompose distributions over sequences $\{x_i\}$ auto-regressively:

$$p_\theta(x) = \prod_{i=0}^{N} p_\theta(x_i|x_{<i})$$

Each term in this product is a discrete classification task (e.g. over words), so the negative log likelihood of $p_\theta(x)$ can be directly optimized to maximize the likelihood

[5] Other classes of generative models include likelihood-free approaches, such as Generative Adversarial Networks (GANs) [59], and energy-based models (EBM) which learn unnormalized $p(x)$ (and can be framed as latent variable models).

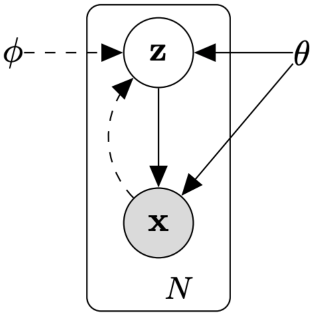


Figure 2-7: Graphical model of a VAE describing the generation of observed data x from unknown latent variable z. Figure from Kingma and Welling [36]. A generative model $p_\theta(x|z)$ and an inference model $q_\phi(x|z)$ parameterized by nonlinear neural networks can be jointly trained to approximate complex data distributions.

of the data.

$$\mathcal{L}(\theta) = -\log p_\theta(x) = \sum_{i=0}^{N} -\log p_\theta(x_i|x_{<i})$$

Auto-regressive models are most commonly used for natural language tasks [60], but have also been used for modeling other forms of sequence data, e.g. biological sequences [66, 30], and more recently, images [88, 16].

Another class of approach is *latent-variable* models, which formulate $p(x)$ in terms of conditional likelihoods $p_\theta(x|z)$ given unobserved latent variables z drawn from a prior distribution $p(z)$. Examples of latent-variable models include Hidden Markov models (HMMs), Bayesian mixture models, as well as deep generative models such as VAEs [36], normalizing flows [14, 15] and diffusion or score-based models [31, 81]. At the intersection of deep learning and generative modeling, *deep generative models* have been of huge interest lately due to their expressive modelling capabilities on large-scale datasets and efficient, tractable inference.

Variational Autoencoders

Variational autoencoders (VAEs) are a class of deep generative model consisting of a nonlinear latent variable model parameterized with neural networks and an associated learning algorithm using variational inference and stochastic gradient estimation. A graphical model overview of a VAE is shown in Figure 2-7. Briefly, the observed data

x is described as being generated by an unobserved continuous latent variable z, i.e.

$$p_\theta(x) = \int p_\theta(x, z)dz$$

$$p_\theta(x, z) = p_\theta(x|z)p(z)$$

The conditional probability $p_\theta(x|z)$ for a particular value of z is a simple factorized distribution, e.g. a Gaussian for continuous data or Bernoulli for discrete data, whose parameters are output by a neural network, i.e. a stochastic decoder with parameters θ. The prior on the latent variables $p(z)$ is a predefined, simple distribution, typically, $p(z) = N(z; 0, I)$.

Given a dataset $\mathbf{x} = x^{(0)}, ..., x^{(N)}$ of i.i.d. samples from $p(x)$, we wish to maximize the marginal (log) likelihood of the data $p_\theta(\mathbf{x})$. However, optimizing the log likelihood function directly is intractable. Instead, a variational approximation to the true posterior is introduced, $q_\phi(z|x) \approx p(z|x)$, parameterized by a Gaussian inference network, i.e. a stochastic encoder with parameters ϕ.

With this variational approximation, the training objective can now be formulated as a lower bound of the log-likelihood, i.e. the Evidence Lower Bound (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|X)}[\log p(z|x)] - D_{KL}(q_\phi(z|x)||p(z)) \triangleq \mathcal{L}(x; \theta, \phi)$$

The expectation in the first term of the ELBO can be computed using Monte Carlo estimation, typically using one sample, and deterministic gradients with respect to θ and ϕ can be computed with the reparameterization trick [36, 64]. Stated simply, the objective function for the VAE consists of a reconstruction error (as in the standard autoencoder) and a regularization term on the latent embedding.

A downstream application of VAEs is that they provide an embedding or feature representation of the data from approximate samples from the posterior distribution over the latent variables $p(z|x)$. These low-dimensional embeddings can be more descriptive than linear embeddings (e.g. from PCA) and more robust and potentially more easily separable than non-generative autoencoders due to the use of the independent latent Gaussian posteriors.

2.4 Computer Vision

In this last section, we briefly overview a related research area in computer vision that has largely developed in parallel with cryo-EM reconstruction – neural rendering for novel view synthesis.

2.4.1 Novel View Synthesis

In computer vision, the task of *novel view synthesis* aims to learn a representation of a 3D scene or object given a dataset of 2D images in order to render novel views from this scene. A fundamental difference between this task and cryo-EM reconstruction is the imaging modality: instead of reconstructing noisy integral projections from an electron microscope, a "reconstruction" is performed by synthesizing natural RBG images, typically of the *same* object with *known* camera poses. There are many possible parameterizations to represent the 3D scene or object (e.g. signed distance functions, occupancy fields, or point clouds). Several research directions aim to improve the scene representation, for example, to enable high quality photorealistic representations, fast rendering, and to include geometric equivariances. There are also many downstream extensions and related applications to novel view synthesis (e.g. semantic understanding of the underlying scene or generalization between scenes). See Tewari et al. for a review of neural rendering methods and their applications [86]. A research direction that is particularly related to the methods described in this thesis is in developing implicit neural representations of the scene, which we highlight next.

2.4.2 Coordinate-based representations

Neural fields was recently proposed as an overarching descriptor of the coordinate-based neural network architecture used in NeRF, cryoDRGN, and other modeling tasks [94]. A neural field uses a neural network to directly model a continuous function that maps spatial coordinates to values, e.g. for a 3D density volume, a function $f_\theta : \mathbb{R}^3 \to \mathbb{R}$.

Many of the early examples of neural fields model functional representations of images or distribution of images (i.e. a 2D field of RGB values). Examples include compositional pattern producing networks (CPPNs) [82], CocoNet [8], and spatial-VAE [5]. In CocoNet, the network is used to memorize the image, which can then be used for various tasks such as denoising and upsampling. In spatial-VAE, latent geometric variables (e.g. 2D rotations) can be explicitly modeled as their geometric transformation of the input coordinates in order to learn latent image factors and help disentangle positional information from image content. In other graphics applications, a coordinate representation was used in learning generative models over texture fields [29, 53, 62, 61] and radiance functions [63]. For modeling 3D scenes from 2D images, Sitzmann et al. proposed Scene Representation Networks (SRN), which used a coordinate-based representation of 3D scenes learned from posed 2D images, capable of generalizing across scenes [80].

Until recently, continuous coordinate-based representations were unable to model high resolution, photorealistic features competitively with discretized voxel-based representations. To address this bottleneck, in cryoDRGN, we modified a coordinate MLP representation for the density volume by introducing a positional encoding of the input coordinates, which significantly improved the representation capacity of the neural model [97]. This representation was also proposed in Neural Radiance Fields for novel view synthesis of natural scenes [44], which we describe in the next section.

2.4.3 Neural Radiance Fields

Mildenhall, Srinivasan, Tancik, et al. introduced Neural Radiance Fields (NeRF), a simple fully-connected neural network capable of representing a single 3D scene with photorealistic quality [44]. More specifically, scenes are represented as a 5-D function, $f_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ mapping from 3D location $\mathbf{x}$ and 2D viewing direction $\mathbf{d}$ to an emitted color $\mathbf{c} = (r, g, b)$ and volume density σ. Differentiable volume rendering is used to train and to render new scenes from this representation. As standard MLPs are poorly suited to modeling complex functions with low-dimensional domain (e.g. a function over x-y-z Cartesian coordinates), they proposed featurizing the input coordinates using a positional encoding function (a sinusoidal basis) inspired from the transformer literature [89]. Follow up work leverages neural tangent kernel theory to explain how this sinusoidal feature mapping allows neural networks to learn high frequency functions in low dimensional domains [84]. Since the publication of NeRF, the field has exploded with NeRF-related follow-up work, exploring many downstream problems and applications, perhaps a testament to the surprising simplicity and effectiveness of the model architecture. Many new research directions explore challenges in applying these methods in real application settings (e.g. unknown camera poses, poor lighting, low compute budget, and dynamic scenes) [86].

Bibliography

[1] Alexey Amunts. The revolution evolution. *Nat Plants*, December 2021.

[2] Xiao-Chen Bai, Israel S Fernandez, Greg McMullan, and Sjors H W Scheres. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife*, 2:e00461, February 2013.

[3] Xiao-Chen Bai, Eeson Rajendra, Guanghui Yang, Yigong Shi, and Sjors H W Scheres. Sampling the conformational space of the catalytic subunit of human γ-secretase. *Elife*, 4, December 2015.

[4] T Bepler, A Morin, M Rapp, J Brasch, and L Shapiro. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 2019.

[5] Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from rotation and translation with spatial-VAE. *Neural Informational Processing Systems (NeurIPS)*, 2019.

[6] D M Blei, A Y Ng, and M I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

[7] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

[8] Paul Andrei Bricman and Radu Tudor Ionescu. CocoNet: A deep neural network for mapping pixel coordinates to color values. *arXiv.org*, May 2018.

[9] Axel F Brilot, James Z Chen, Anchi Cheng, Junhua Pan, Stephen C Harrison, Clinton S Potter, Bridget Carragher, Richard Henderson, and Nikolaus Grigorieff. Beam-induced motion of vitrified specimen on holey carbon film. *J. Struct. Biol.*, 177(3):630–637, March 2012.

[10] Marcus A Brubaker, Ali Punjani, and David J Fleet. Building proteins in a day: Efficient 3D molecular reconstruction. April 2015.

[11] Muyuan Chen, Steven Ludtke, and Verna Marrs. Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. *arXiv*, 2021.

[12] Daniel Cressey and Ewen Callaway. Cryo-electron microscopy wins chemistry nobel. *Nature*, 550(7675):167, October 2017.

[13] Tristan Ian Croll. ISOLDE: A physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology*, 2018.

[14] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. October 2014.

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. November 2016.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. October 2020.

[17] J Dubochet, M Adrian, J J Chang, J C Homo, J Lepault, A W McDowall, and P Schultz. Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.*, 21(2):129–228, May 1988.

[18] J Dubochet, J Lepault, R Freeman, J A Berriman, and J-C Homo. Electron microscopy of frozen water and aqueous solutions. *J. Microsc.*, 128(3):219–237, December 1982.

[19] P Emsley, B Lohkamp, W G Scott, and K Cowtan. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, April 2010.

[20] Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods (San Diego, Calif.)*, 100:61–67, May 2016.

[21] George E Ghanim, Adam J Fountain, Anne-Marie M van Roon, Ramya Rangan, Rhiju Das, Kathleen Collins, and Thi Hoang Duong Nguyen. Structure of human telomerase holoenzyme with bound telomeric DNA. *Nature*, 593(7859):449–453, May 2021.

[22] Nikolaus Grigorieff. Direct detection pays off for electron cryo-microscopy. *Elife*, 2:e00573, February 2013.

[23] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *bioRxiv*, 2020.

[24] Harshit Gupta, Thong H Phan, Jaejun Yoo, and Michael Unser. Multi-CryoGAN: Reconstruction of continuous conformations in cryo-EM using generative adversarial networks. In *Computer Vision – ECCV 2020 Workshops*, Lecture notes

in computer science, pages 429–444. Springer International Publishing, Cham, 2020.

[25] P.W. Hawkes and M. Hytch. *The Beginnings of Electron Microscopy - Part 1*. ISSN. Elsevier Science, 2021.

[26] R Henderson, J M Baldwin, T A Ceska, F Zemlin, E Beckmann, and K H Downing. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.*, 213(4):899–929, June 1990.

[27] R Henderson and P N Unwin. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, 257(5521):28–32, September 1975.

[28] Richard Henderson, Andrej Sali, Matthew L Baker, Bridget Carragher, Batsal Devkota, Kenneth H Downing, Edward H Egelman, Zukang Feng, Joachim Frank, Nikolaus Grigorieff, Wen Jiang, Steven J Ludtke, Ohad Medalia, Pawel A Penczek, Peter B Rosenthal, Michael G Rossmann, Michael F Schmid, Gunnar F Schröder, Alasdair C Steven, David L Stokes, John D Westbrook, Willy Wriggers, Huanwang Yang, Jasmine Young, Helen M Berman, Wah Chiu, Gerard J Kleywegt, and Catherine L Lawson. Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2):205–214, February 2012.

[29] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Learning a neural 3D texture space from 2D exemplars. December 2019.

[30] Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[32] Zhong Huang and Pawel A Penczek. Application of template matching technique to particle detection in electron micrographs. *J. Struct. Biol.*, 145(1-2):29–40, January 2004.

[33] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, July 2021.

[34] Daisuke Kihara, Genki Terashi, and Sai Raghavendra Maddhuri Venkata Subramaniya. De Novo Computational Protein Tertiary Structure Modeling Pipeline for Cryo-EM Maps of Intermediate Resolution. *Biophysical Journal*, 118(3):292a, feb 2020.

[35] Dari Kimanius, Gustav Zickert, Takanori Nakane, Jonas Adler, Sebastian Lunz, C-B Schönlieb, Ozan Öktem, and Sjors HW Scheres. Exploiting prior knowledge about biological macromolecules in cryo-em structure determination. *IUCrJ*, 8(1), 2021.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations (ICLR)*, 2013.

[37] M Knoll and E Ruska. Das elektronenmikroskop. *Zeitschrift für Physik*, 78(5):318–339, May 1932.

[38] Kühlbrandt, Werner. Biochemistry. The resolution revolution. *Science*, 343(6178):1443–1444, March 2014.

[39] Roy R Lederman, Joakim Andén, and Amit Singer. Hyper-Molecules: on the Representation and Recovery of Dynamical Structures, with Application to Flexible Macro-Molecular Structures in Cryo-EM. *arXiv.org*, July 2019.

[40] Roy R Lederman and Amit Singer. Continuously heterogeneous hyper-objects in cryo-EM and 3-D movies of many temporal dimensions. *arXiv.org*, April 2017.

[41] Andres E Leschziner and Eva Nogales. The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *Journal of structural biology*, 153(3):284–299, 2006.

[42] Dorothee Liebschner, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkoczi, 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, oct 2019.

[43] W Liu and J Frank. Estimation of variance distribution in three-dimensional reconstruction. i. theory. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 12(12):2615–2627, December 1995.

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. March 2020.

[45] Jacqueline L S Milne, Mario J Borgnia, Alberto Bartesaghi, Erin E H Tran, Lesley A Earl, David M Schauder, Jeffrey Lengyel, Jason Pierson, Ardan Patwardhan, and Sriram Subramaniam. Cryo-electron microscopy–a primer for the non-microscopist. *FEBS J.*, 280(1):28–45, January 2013.

[46] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *arXiv.org*, July 2019.

[47] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[48] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia M G E Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, Tomasz Uchański, Lingbo Yu, Dimple Karia, Evgeniya V Pechnikova, Erwin de Jong, Jeroen Keizer, Maarten Bischoff, Jamie McCormack, Peter Tiemeijer, Steven W Hardwick, Dimitri Y Chirgadze, Garib Murshudov, A Radu Aricescu, and Sjors H W Scheres. Single-particle cryo-EM at atomic resolution. *Nature*, 587(7832):152–156, October 2020.

[49] Youssef S G Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-end simultaneous learning of single-particle orientation and 3D map reconstruction from cryo-electron microscopy data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, October 2021.

[50] W V Nicholson and R M Glaeser. Review: automatic particle detection in electron microscopy. *J. Struct. Biol.*, 133(2-3):90–101, February 2001.

[51] NobelPrize.org. The nobel prize in physics 1986, 1986.

[52] NobelPrize.org. The nobel prize in chemistry 2017, 2017.

[53] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. May 2019.

[54] Pawel A Penczek, Marek Kimmel, and Christian M T Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure*, 19(11):1582–1590, November 2011.

[55] 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. *Journal of Computational Chemistry*, 25(13):1605–1612, oct 2004.

[56] Ali Punjani and David J Fleet. 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. *bioRxiv*, 2021.

[57] Ali Punjani and David J Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.*, 213(2):107702, June 2021.

[58] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, March 2017.

[59] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[60] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[61] Gilles Rainer, Abhijeet Ghosh, Wenzel Jakob, and Tim Weyrich. Unified neural encoding of BTFs. *Comput. Graph. Forum*, 39(2):167–178, May 2020.

[62] Gilles Rainer, Wenzel Jakob, Abhijeet Ghosh, and Tim Weyrich. Neural BTF compression and interpolation. *Comput. Graph. Forum*, 38(2):235–244, May 2019.

[63] Peiran Ren, Jiaping Wang, Minmin Gong, Stephen Lin, Xin Tong, and Baining Guo. Global illumination with radiance regression functions. *ACM Trans. Graph.*, 32(4):1–12, July 2013.

[64] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 2014. PMLR.

[65] William J Rice, Anchi Cheng, Alex J Noble, Edward T Eng, Laura Y Kim, Bridget Carragher, and Clinton S Potter. Routine determination of ice thickness for cryo-EM grids. *J. Struct. Biol.*, 204(1):38–44, October 2018.

[66] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15), April 2021.

[67] Alexis Rohou and Nikolaus Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.*, 192(2):216–221, November 2015.

[68] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch,

S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes, 2021.

[69] Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, October 2003.

[70] H Ru. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell*, 163, 2015.

[71] Ernst Ruska and M Knoll. Die magnetische sammelspule für schnelle elektronenstrahlen. *The magnetic concentrating coil for fast electron beams.) Z. techn. Physik*, 12:389–400, 1931.

[72] R Sánchez-García, J Gomez-Blanco, A Cuervo, J M Carazo, COS Sorzano, and J Vargas. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *bioRxiv*, 2020.

[73] Sjors H W Scheres. Chapter eleven - classification of structural heterogeneity by Maximum-Likelihood methods. In Grant J Jensen, editor, *Methods in Enzymology*, volume 482, pages 295–320. Academic Press, January 2010.

[74] Sjors H W Scheres. A Bayesian view on cryo-EM structure determination. *Journal of molecular biology*, 415(2):406–418, January 2012.

[75] Sjors H W Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3):519–530, December 2012.

[76] Sjors H W Scheres. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.*, 189(2):114–122, February 2015.

[77] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.

[78] Dong Si, Spencer A Moritz, Jonas Pfab, Jie Hou, Renzhi Cao, Liguo Wang, Tianqi Wu, and Jianlin Cheng. Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. *Scientific Reports*, 10(1):1–22, mar 2020.

[79] Amit Singer and Fred J. Sigworth. Computational Methods for Single-Particle Electron Cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, jul 2020.

[80] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[81] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based generative modeling through stochastic differential equations. In *The International Conference on Learning Representations (ICLR)*, 2021.

[82] Kenneth O Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genet. Program. Evolvable Mach.*, 8(2):131–162, June 2007.

[83] Hemant D Tagare, Alp Kucukelbir, Fred J Sigworth, Hongwei Wang, and Murali Rao. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *J. Struct. Biol.*, 191(2):245–262, August 2015.

[84] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. June 2020.

[85] K A Taylor and R M Glaeser. Electron diffraction of frozen, hydrated protein crystals. *Science*, 186(4168):1036–1037, December 1974.

[86] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering. November 2021.

[87] Leonardo G Trabuco, Elizabeth Villa, Kakoli Mitra, Joachim Frank, and Klaus Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure (London, England : 1993)*, 16(5):673–683, May 2008.

[88] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. January 2016.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[90] N R Voss, C K Yoshioka, M Radermacher, C S Potter, and B Carragher. DoG picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.*, 166(2):205–213, May 2009.

[91] Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, Dennis Quentin, Daniel Roderer, Sebastian Tacke, Birte Siebolds, Evelyn Schubert, Tanvir R Shaikh, Pascal Lill, Christos Gatsogiannis, and Stefan Raunser. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol*, 2:218, June 2019.

[92] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng. DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. *J. Struct. Biol.*, 195(3):325–336, September 2016.

[93] Wilson Wong, Xiao-Chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H W Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e01963, June 2014.

[94] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. November 2021.

[95] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, November 2020.

[96] K Zhang. Gctf: real-time CTF determination and correction. *J. Struct. Biol.*, 193, 2016.

[97] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

[98] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-em reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

[99] Yanan Zhu, Qi Ouyang, and Youdong Mao. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics*, 18(1):348, July 2017.

[100] Jasenko Zivanov, Takanori Nakane, and Sjors H W Scheres. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ*, 7(Pt 2):253–267, March 2020.

Chapter 3

CryoDRGN: A neural model for cryo-EM reconstruction

Single particle cryo-EM is uniquely poised to visualize the dynamics of large macromolecular machines (Chapter 2). However, the structural variability of imaged biomolecular complexes complicates the 3D reconstruction process and is typically addressed using discrete clustering approaches that fail to capture the full range of protein dynamics.

In this chapter, we introduce the main contribution of this thesis: a neural method, cryoDRGN (Deep Reconstructing Generative Networks), for heterogeneous cryo-EM reconstruction. We propose a deep generative model of volumes for 3D cryo-EM reconstruction. The model is trained from unlabelled 2D images, and we show that it can learn can learn continuous deformations in protein structure.

This chapter presents work described in [13] performed jointly with Tristan Bepler, Joey Davis, and Bonnie Berger and presented at the 2020 International Conference of Learning Representations.

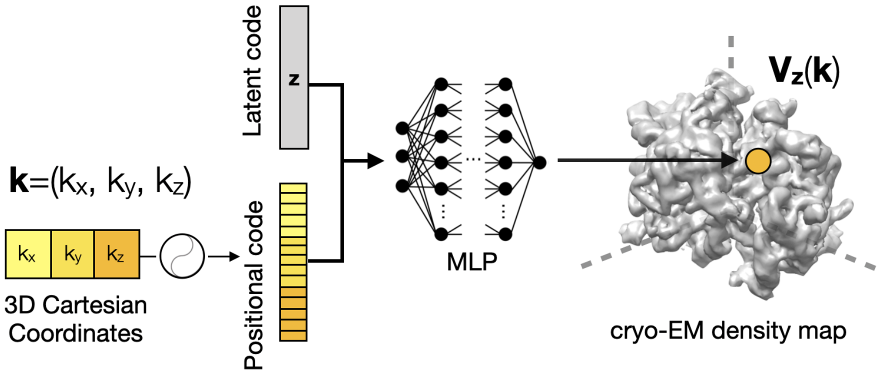


Figure 3-1: Architecture of cryoDRGN's neural model of cryo-EM density.

3.1 Methods

3.1.1 Volume representation

The cryoDRGN method proposes a new representation for modeling cryo-EM volumes using neural networks (Figure 3-1). Instead of the typical voxel-based representation for volumes, which discretizes an inherently continuous function, here, a neural network is used to model the continuous density function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$. Given 3D Cartesian coordinates $\mathbf{x}$ (i.e. x-y-z values), the model outputs a prediction of the density at that location in space. This function is approximated using a dense, fully-connected neural network, also called a multi-layer perceptron (MLP).

Instead of directly providing 3D Cartesian coordinates $\mathbf{x}$ to the MLP, the input coordinates are first featurized using a fixed positional encoding function consisting of D sinusoids at varying frequency. This choice of parameterization was originally inspired from the transformer literature [10], which uses this featurization to provide contextual information on the position of words within sequences.

$$pe^{(2d)}(k) = cos(\gamma(d)k); d = 0, ...D/2 - 1 \tag{3.1}$$

$$pe^{(2d+1)}(k) = sin(\gamma(d)k); d = 0, ...D/2 - 1 \tag{3.2}$$

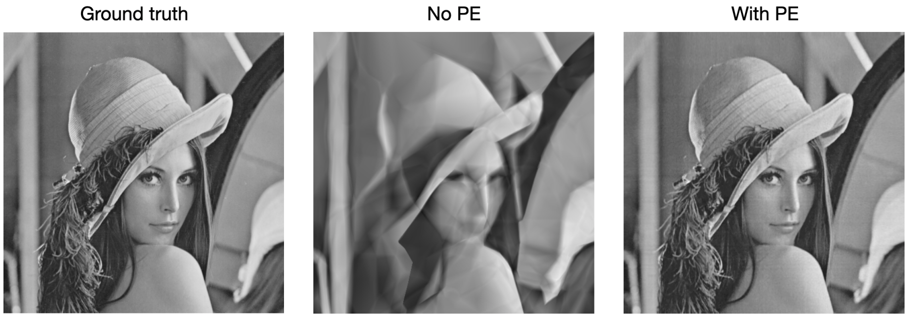


Figure 3-2: Result after regressing a coordinate MLP modeling $f_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}$ to the ground truth signal with and without the positional encoding (PE) of input coordinates.

$$\gamma(d) = D\pi \left(\frac{2}{D}\right)^{d/(D/2-1)} \tag{3.3}$$

Without this featurization, the coordinate MLP is poorly suited to learning high frequency signals. We empirically show this through a simple toy experiment where we regress a coordinate MLP to memorize a black and white image (i.e. a 2D function) with and without the positional encoding of input coordinates (Figure 3-2).

In the cryo-EM reconstruction setting, the coordinate MLP is used to model volumes in the frequency domain. We investigated both a Fourier or Hartley domain representation. In the Fourier domain, input coordinates k are wavevectors, and the outputs are coefficients of the real and imaginary components of the signal at k. Modeling the volume in the frequency domain allows us to efficiently relate 2D images as slices out of the 3D volume via the Fourier slice theorem. We also explore a Hartley domain representation of the volume function, which is closely related to the Fourier representation as the real minus the imaginary component for real-valued functions.

Without loss of generality, we assume a length scale of the volume representation that restricts the support of the volume to a sphere of radius 0.5. The wavelengths of the positional encoding in Equations 3.1 and 3.2 thus follow a geometric series spanning the Fourier basis from wavelength 1 to the Nyquist limit $(2/D)$ of the image data. While this encoding empirically works well for noiseless data, we obtain better results with a slightly modified featurization for noisy datasets consisting of a geometric series that excludes the top 10 percentile of highest frequency components of the noiseless positional encoding.

The computer vision literature has since popularized this model architecture through the introduction of neural radiance fields (NeRF) and the subsequent explosion of literature using, adapting, and improving upon NeRF models (See Section 2.4). Follow up work leverages neural tangent kernel theory to explain how this sinusoidal feature mapping allows neural networks to learn high frequency functions in low dimensional domains [9]. In a later version of cryoDRGN (software version 1.0), instead of geometrically-spaced, axis-aligned Fourier features, we follow Tancik et al. [9] and use frequencies sampled from a Gaussian distribution.

3.1.2 Generative model

Our volume representation extends naturally to modeling continuous generative factors of structural heterogeneity. Instead of approximating a single volume, $\hat{V} : \mathbb{R}^3 \to \mathbb{R}$, we propose a deep generative model to approximate the function, $\hat{V} : \mathbb{R}^{3+n} \to \mathbb{R}$, representing a continuous n-dimensional manifold of 3D density volumes in the Fourier domain.

Specifically, the volume $\hat{V}$ is modelled as a probabilistic decoder $p_\theta(\hat{V}|k, z)$, where θ are parameters of a multilayer perceptron (MLP). Given Cartesian coordinates $k \in \mathbb{R}^3$ and continuous latent variable z, the decoder outputs distribution parameters for a Gaussian distribution over $\hat{V}(k, z)$, i.e. the density of volume $\hat{V}_z$ at frequency k in Fourier space. Here, these coordinates are explicitly treated as each pixel's location in 3D Fourier space and thus enforce the topological constraints between 2D views in 3D via the Fourier slice theorem.

By the image formation model, each image corresponds to an *oriented* central slice of the 3D volume in the Fourier domain (Section 2). During training, the 3D coordinates of an image's pixels can be explicitly represented by the rotation of a $D \times D$ lattice initially on the x-y plane. Under this model, the log probability of an image, $\hat{X}$, represented as a vector of size $D \times D$, given the current MLP, latent pose variables $R \in SO(3)$ and $t \in \mathbb{R}^2$, and unconstrained latent variable, z, is:

$$\log p(\hat{X}|R,t,z) = \log p(\hat{X}'|R,z) = \sum_i \log p_\theta(\hat{V}|R^T c_0^{(i)}, z) \tag{3.4}$$

where i indexes over the coordinates of a fixed lattice c_0. Note that $\hat{X}' = S(-t)\hat{X}$ is the centered image, where S is the phase shift operator corresponding to image translation in real space. We define c_0 as a vector of 3D coordinates of a fixed lattice spanning $[-0.5, 0.5]^2$ on the x-y plane to represent the unoriented coordinates of an image's pixels.

3.1.3 Inference

We employ a standard VAE for approximate inference of the latent variable z, but use a global search to infer the pose $\phi = (R, t)$ using a branch and bound algorithm.

Variational encoder: As each cryo-EM image is a noisy projection of an instance of the volume at a random, unknown pose (viewing direction), the image encoder aims to learn a *pose-invariant* representation of the protein's structural heterogeneity. Following the standard VAE framework, the probabilistic encoder $q_\xi(z|\hat{X})$ is a MLP with variational parameters ξ and Gaussian output with diagonal covariance. Given an input cryo-EM image $\hat{X}$, represented as a $D \times D$ vector, the encoder MLP outputs $\mu_{z|\hat{X}}$ and $\Sigma_{z|\hat{X}}$, statistics that parameterize an approximate posterior to the intractable true posterior $p(z|\hat{X})$. The prior on z is a standard normal, $\mathcal{N}(0, \mathbf{I})$.

Pose inference: We perform a global search over $SO(3) \times \mathbb{R}^2$ for the maximum-likelihood pose for each image given the current decoder MLP and a sampled value of z from the approximate posterior. Two techniques are used to improve the efficiency of the search over poses: (1) discretizing the search space on a uniform grid and sub-dividing grid points after pruning candidate poses with *branch and bound* (BNB), and (2) band pass limiting the objective to low frequency components and incrementally increasing the k-space limit at each iteration (*frequency marching*). The pose inference procedure encodes the intuition that low-frequency components dominate pose estimation, and is fully described in Section 3.4.

In summary, for a given image $\hat{X}_i$, the image encoder produces $\mu_{z|\hat{X}_i}$ and $\Sigma_{z|\hat{X}_i}$. A

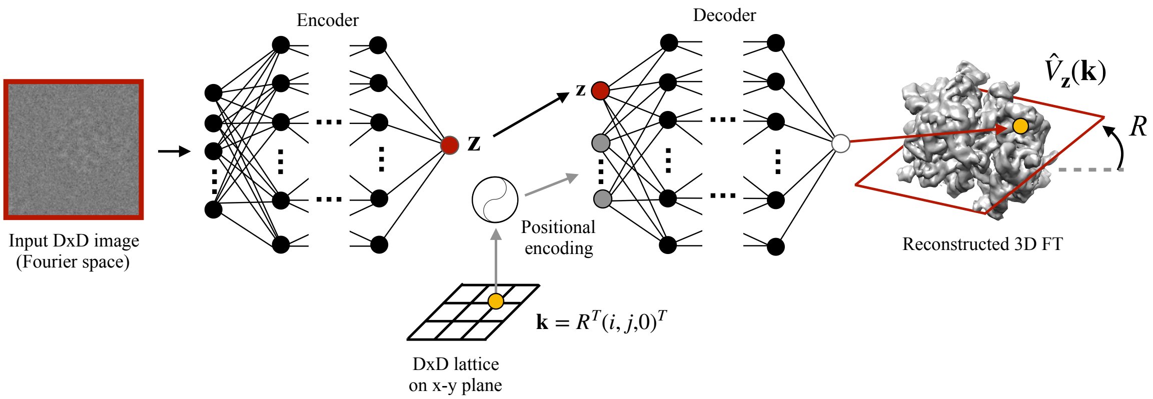


Figure 3-3: CryoDRGN model architecture. We use a VAE to perform approximate inference for latent variable z denoting image heterogeneity. The decoder reconstructs an image pixel by pixel given z and $pe(k)$, the positional encoding of 3D Cartesian coordinates. The 3D coordinates corresponding to each image pixel are obtained by rotating a $D \times D$ lattice on the x-y plane by R, the image orientation. The latent orientation for each image is inferred through a branch and bound global optimization procedure (not shown).

sampled value of the latent $z_i \sim \mathcal{N}(\mu_{z|\hat{X}_i}, \Sigma_{z|\hat{X}_i})$ is broadcast to all pixels. Given z_i and the current decoder, 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 $R_i^T c_0$ and z_i. The phase shift corresponding to t_i and optionally the microscope CTF $\hat{g}_i$ is then applied on the reconstructed pixel intensities. Following the standard VAE framework, the optimization objective is the variational lower bound of the model evidence:

$$\mathcal{L}(\hat{X}_i; \xi, \theta) = \mathbb{E}_{q_\xi(z|\hat{X}_i)}[\log p_\theta(\hat{X}_i|z)] - KL(q_\xi(z|\hat{X}_i)||p(z)) \tag{3.5}$$

where the expectation of the log likelihood is estimated with one Monte Carlo sample. By comparing many 2D slices from the imaging dataset, the volume can be learned through feedback from these single views. Furthermore, this learning process is denoising as overfitting to noise from a single image would lead to higher reconstruction error for other views. We note that the distribution of 3D volumes models heterogeneity within a single imaging dataset, capturing structural variation for a particular protein or biomolecular complex, and that a separate network is trained per experimental dataset. Unless otherwise specified, the encoder and decoder networks are both MLPs containing 10 hidden layers of dimension 128 with ReLU activations for the results presented in this chapter. Further architecture and implementation details are given in Section 3.4.

3.2 Results

In this section, we present both qualitative and quantitative results for 1) homogeneous cryo-EM reconstruction, validating that cryoDRGN reconstructed volumes match those from existing tools; 2) heterogeneous cryo-EM reconstruction with pose supervision, demonstrating automatic learning of the latent manifold that previously required many expert-guided rounds of multiclass refinement; and 3) *ab initio* reconstruction of continuous distributions of 3D protein structures, a capability not provided by any existing tool at the time of publication.

3.2.1 *Ab initio* homogeneous reconstruction

We first evaluate cryoDRGN on homogeneous datasets, where existing tools are capable of reconstruction. We create two synthetic datasets following the cryo-EM image formation model (image size D=128, 50k projections, with and without noise), and use one real dataset from EMPIAR-10028 consisting of 105,247 images of the 80S ribosome downsampled to image size D=90. The encoder network is not used in homogeneous reconstruction. As a baseline for comparison, we perform homogeneous *ab-initio* reconstruction followed by iterative refinement in cryoSPARC [8]. We compare against cryoSPARC as a representative of traditional state-of-the-art tools. Further dataset preprocessing and training details are given in Section 3.4.

We find that cryoDRGN inferred poses and reconstructed volumes match those from state of-the-art tools. The similarity of the volumes to the ground truth can be quantified with the with the Fourier shell correlation (FSC) curve[1]. Reconstructed

[1] The FSC curve measures correlation between volumes as a function of radial shells in Fourier space. The field currently lacks a rigorous method for measuring the quality of reconstruction. In practice, however, resolution is often reported as $1/k_0$ where $k_0 = \arg\max_k FSC(k) < C$ and C is some fixed threshold.

Method	**Dataset** No Noise	 SNR=0.1
cryoSPARC	0.0009 / 0.47	0.002 / 0.64
cryoDRGN	0.0004 / 0.27	0.003 / 0.38

Table 3.1: Homogeneous reconstruction pose accuracy quantified by median rotation/translation error to the ground truth image poses. Rotation/translation error is defined as the squared Frobenius/L2 norm after a global alignment of the reconstructed volumes.

volumes and quantitative comparison with the FSC curve is given in Figure 3-6. Pose error to the ground truth image poses are given in Table3.1. For the real cryo-EM dataset (no ground truth), the median pose difference between cryoDRGN and cryoSPARC reconstructions is 0.002 for rotations and 1.0 pixels for translations, and the resulting volumes are correlated above a FSC cutoff of 0.5 across all frequencies.

3.2.2 Heterogeneous reconstruction with pose supervision

Next, we evaluate cryoDRGN for heterogeneous cryo-EM reconstruction on EMPIAR-10076, a real dataset of the *E. coli* large ribosomal subunit (LSU) undergoing assembly (131,899 images, downsampled to D=90) [3]. Here, poses are obtained through alignment to an existing structure of the LSU and treated as known during training. In the original analysis of this dataset, multiple rounds of discrete multiclass refinement with varying number of classes followed by human comparison of similar volumes were used to identify 4 major structural states of the LSU. We train cryoDRGN with a 1-D latent variable treating image pose as fixed to skip BNB pose inference. As a baseline, we reproduce the published structures originally obtained through multiclass refinement with cryoSPARC. Further baseline and training details are given in Section 3.4.

We find that cryoDRGN automatically identifies all 4 major states of the LSU (Figure 3-4a). Quantitative comparison with FSC curves[1] and additional volumes along the latent space are shown in Figure 3-8. We compare the cryoDRGN latent encoding $\mu_{z|X}$ for each image to the MAP cluster assignment in cryoSPARC and

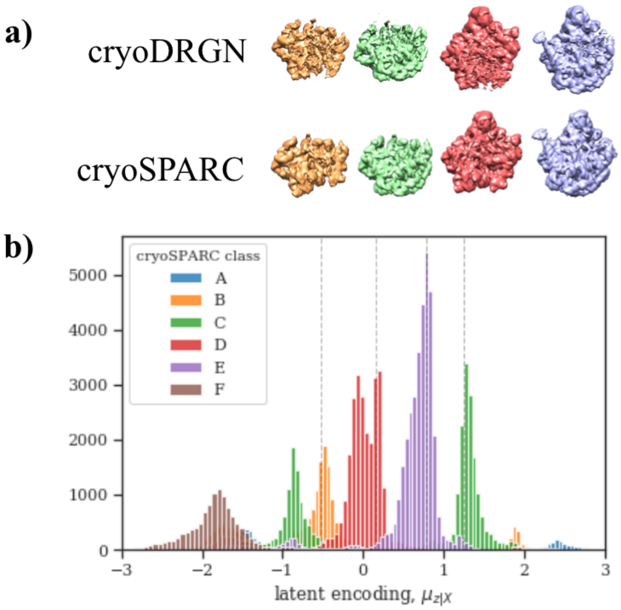


Figure 3-4: a) Volumes generated at values of the latent encoding (at the dashed lines) and the corresponding published volumes of the 4 major states of the LSU. b) Histogram of latent encodings from cryoDRGN, colored by cluster assignment from a discrete multiclass reconstruction in cryoSPARC.

find that the learned latent manifold aligns with cryoSPARC clusters (Figure 3-4b). CryoDRGN identifies subpopulations in some of the cryoSPARC clusters (e.g. Class D), which is partitioned by a subsequent round of cryoSPARC multiclass refinement (Figure 3-9). Published structures A and F correspond to impurities in the sample. CryoDRGN correctly assigns images from these impurities to distinct clusters, but does not learn their correct structure since the poses inferred from aligning to the LSU template structure are incorrect.

3.2.3 *Ab initio* heterogeneous reconstruction

We test the ability of cryoDRGN to perform *ab initio* heterogeneous reconstruction from datasets with different latent structure. We generate four datasets (each 50k projections, D=64) from an atomic model of a protein complex, containing either a 1D continuous motion, 2D continuous motion, 1D continuous circular motion, or a mixture of 10 discrete conformations (Figure 3-8). We train cryoDRGN with a 1D

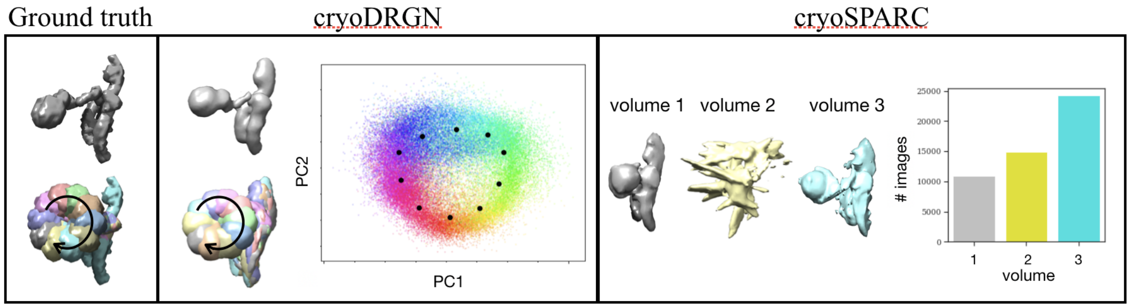


Figure 3-5: *Left:* Ground truth volume containing a continuous circular 1D motion. *Middle:* Reconstructed structures from cryoDRGN match the ground truth volumes with the correct continuous deformation. We visualize 10 structures (superimposed) sampled at the depicted points in the latent space. The distribution of images in the latent space (visualized in 2D with PCA) matches the topology of the true data manifold. *Right:* Reconstructed volumes from discrete 3-class reconstruction in cryoSPARC and the distribution of images over the three reconstructed volumes.

Dataset	**cryoDRGN**	**cryoDRGN+tilt**	**cryoSPARC**
Linear 1D motion	2.50(0.62)	2.35(0.36)	3.60(2.27)
Linear 2D motion	4.44(2.50)	2.93(1.02)	6.90(3.77)
Circular 1D motion	4.05(2.40)	2.63(0.74)	4.87(2.17)
Discrete 10 class	4.95(3.16)	2.58(1.00)	5.69(5.15)

Table 3.2: Reconstruction accuracy quantified by an FSC=0.5 resolution metric between the reconstructed volumes corresponding to each image and its ground truth volume. We report the average and standard deviation across 100 images in the dataset (lower is better; best possible is 2 pixels).

latent variable for the linear 1D dataset and a 10D latent variable for the other 3 datasets. As a baseline, we perform multiclass reconstruction in cryoSPARC sweeping K=2-5 classes. We compare against K=3, which had the best qualitative results.

We also propose a modification to cryoDRGN in order to train on *tilt series pairs* datasets. Tilt series pairs is a variant of cryo-EM in which, for each image X_i, a corresponding image X_i' is acquired after tilting the imaging stage by a known angle. This technique was originally employed to identify the chirality of molecules [2], which is lost in the projection from 3D to 2D. We propose using tilt series pairs to encourage invariance of q_ξ with respect to pose transformations for a given $\hat{V}_{\mathbf{z}}$ (and incidentally to identify the chirality of $\hat{V}_{\mathbf{z}}$). We make minor modifications to the architecture as described in Section 3.4.

In Figure 3-5, we show that cryoDRGN reconstructed volumes for the circular 1D dataset qualitatively match the ground truth structures. Note that while we only visualize 10 structures sampled along the latent space, the volume decoder can reconstruct the full continuum of states. In contrast, cryoSPARC multiclass reconstruction, a discrete mixture model of independent structures, is only able to reconstruct 2 (originally unaligned) structures which resemble the ground truth. Volumes contain blurring artifacts from clustering images from different conformations into the assumed-homogeneous clusters in the mixture model. Results for the remaining datasets are given in Figures 3-11-3-14.

We quantitatively measure performance on this task with a “per-image FSC” resolution metric computed between the MAP volume for each image $V_{z_i|\hat{X}_i}$ and the ground truth volume which generated each image, averaged across images in the dataset (Table 3.2). We find that cryoDRGN reconstruction accuracy is much higher than state-of-the-art discrete multiclass reconstruction in cryoSPARC, with further improvement achieved by training on tilt series pairs.

3.3 Conclusions

We present a novel neural network-based reconstruction method for single particle cryo-EM that learns continuous variation in protein structure. We applied cryoDRGN on a real dataset of highly heterogeneous ribosome assembly intermediates and demonstrate automatic partitioning of structural states. In the presence of simulated continuous heterogeneity, we show that cryoDRGN learns a continuous representation of structure along the true reaction coordinate, effectively disentangling imaging orientation from intrinsic structural heterogeneity. The techniques described here may also have broader applicability to image and volume generative modelling in other domains of computer vision and 3D shape reconstruction.

3.4 Appendix - Methods

3.4.1 Branch and bound implementation details

We perform a global search over $SO(3) \times \mathbb{R}^2$ for the maximum-likelihood pose for each image given the current decoder MLP. Two techniques are used to improve the efficiency of the search over poses: (1) discretizing the search space on a uniform grid and sub-dividing grid points after pruning candidate poses with *branch and bound*, and (2) band pass limiting the objective to low frequency components and incrementally increasing the k-space limit at each iteration (*frequency marching*).

Our branch and bound algorithm for pose optimization is given in Algorithm 1. Briefly, we discretize $SO(3)$ uniformly using the Hopf fibration [12] at a predefined base resolution of the grid and incrementally increase the grid resolution by sub-dividing grid points. At each resolution of the grid, the set of candidate poses is pruned using a branch and bound (BNB) optimization scheme, which alternates between a computationally inexpensive lower bound on the objective function evaluated at all grid points and an upper bound consisting of the true objective evaluated on the best lower-bound candidate. Grid points whose lower bound is higher than this value are excluded for subsequent iterations. In our case, the loss is evaluated on low-frequency components of the image; specifically, Fourier components with $|\mathbf{k}| < k_{max}$ is an effective lower bound, as it is both inexpensive to compute and captures most of the power (and thus the error). This bound encodes the intuition that low-frequency components dominate pose estimation. We concomitantly increase k_{max} at each iteration of grid subdivision.

At each iteration, some poses are excluded by BNB, and the remaining poses are further discretized. Although BNB is risk-free in the sense that the optimal pose at a given resolution will not be pruned, our application of it is not risk-free as a candidate pose with high loss at a given resolution doesn't guarantee that its neighbor in the next iteration will not have a lower loss. Irrespective, in practice, we find that at a sufficiently fine base resolution, we obtain good results on a tractable timescale (hours on a single GPU).[2]

We reimplement the uniform multiresolution grids on $SO(3)$ based on [12], using the Healpix [4] grid for the sphere and the Hopf fibration to uniformly lift the grid to $SO(3)$. The base grid on $SO(3)$ contains 576 orientations. We use the ordinary grid for translations containing 7^2 points with an extent of 20 pixels for D=128 datasets. We subdivide the grid 5 times for a final resolution of 0.92 degrees for the orientation and 0.08 pixels for the translation. For D=64 datasets, we use a translational grid with extent of 10 pixels.

Algorithm 1 CryoDRGN branch and bound with frequency marching

```
procedure OPTPHI($\hat{X}, \hat{V}_{\mathbf{z}}$)    ▷ Find the optimal image pose given the current decoder
    $k_{min} \leftarrow 12,\ k_{max} \leftarrow D/2,\ N_{iter} \leftarrow 5$
    $\Phi \leftarrow SO(3) \times \mathbb{R}^2$ grid at base resolution
    $k \leftarrow k_{min}$
    for $iter = 1 \ \dots \ N_{iter}$ do
        for $\phi_i \in \Phi$ do    ▷ Compute lower bound at all grid points
            $lb(\phi_i) \leftarrow$ loss between $\hat{X}$ and SLICE($\hat{V}_{\mathbf{z}}, \phi_i$) at $\mathbf{k} < k$
        $\phi^* \leftarrow \arg\min(lb)$
        $ub \leftarrow$ loss between $\hat{X}$ and SLICE($\hat{V}_{\mathbf{z}}, \phi^*$) at $\mathbf{k} < k_{max}$    ▷ Compute upper bound
        $\Phi_{new} \leftarrow \{\}$
        for $\phi_i \in \Phi$ do    ▷ Subdivide grid points below the upper bound
            if $lb(\phi_i) < ub$ then
                $\Phi_{new} \leftarrow \Phi_{new} \cup$ SUBDIVIDE($\phi_i$)
        $\Phi \leftarrow \Phi_{new}$
        $k \leftarrow k + (k_{max} - k_{min})/(N_{iter} - 1)$    ▷ Increase frequency band limit
    return $\phi^*$
```

[2] The difference in loss between nearby poses could be incorporated into the BNB lower bound, but this would require assumptions about the smoothness of the loss with respect to pose. We leave this detail for future work.

3.4.2 Training details

Given an imaging dataset, $\hat{X}_1, ...\hat{X}_N$, we summarize three training paradigms of cryoDRGN. 1) For homogeneous reconstruction, we only train the volume decoder p_θ and perform BNB pose inference for the unknown ϕ_i's for each image. 2) As an intermediate task, we can perform heterogeneous reconstruction training the image encoder q_ξ and the volume decoder p_θ with known ϕ_i's to skip BNB pose inference. 3) For *ab initio* heterogeneous reconstruction, we jointly train q_ξ and p_θ to learn a continuous latent representation, performing BNB pose inference for the unknown pose of each image.

Unless otherwise specified, the encoder and decoder networks are both MLPs containing 10 hidden layers of dimension 128 with ReLU activations. A fully connected architecture is used instead of a convolutional architecture because the images are not represented in real space.

Instead of representing both the real and imaginary components of each image, we use the closely-related Hartley space representation [5]. The Hartley transform of real-valued functions is equivalent to the real minus imaginary component of the FT, and thus is real valued. The Fourier slice theorem still holds and the error model is equivalent.

In this work, we simplify the image generation model to Gaussian white noise. Therefore, for a given image, the negative log likelihood for a reconstructed slice from the decoder corresponds to the mean squared error between the phase-shifted image and the oriented slice from the volume decoder. We leave the implementation of a colored noise model to future work.

We use the Adam optimizer [6] with learning rate of 5e-4 for experiments involving noiseless, homogeneous datasets, and 1e-4 for all other experiments. All models are implemented in PyTorch [7].

3.5 Homogeneous reconstruction

3.5.1 Dataset preparation

Simulated datasets: From a ground truth 3D volume, we simulated datasets following the cryo-EM image formation model by 1) rotating the 3D volume in real space by R, where $R \in SO(3)$ is sampled uniformly, 2) projecting (integrating) the volume along the z-axis, 3) shifting the resulting 2D image by t, where t is sampled uniformly from $[-10, 10]^2$ pixels, and 4) optionally adding noise to an SNR of 0.1, a typical value for cryo-EM data [1]. Following convention in the cryo-EM field, we define SNR as the ratio of the variance of the signal to the variance of the noise. We define the noise-free signal images to be the entire DxD image. 50k projections were generated for each dataset with image size of D=128.

Real dataset: To generate the real cryo-EM dataset for homogeneous reconstruction, images from EMPIAR-10028 [11] were downsampled by a factor of 4 by clipping in Fourier space. The images were then 'phase flipped' in Fourier space by their contrast transfer function, a given real-valued function with range $[-1, 1]$ determined by the microscopy conditions, i.e. the Fourier components are negated where the CTF is negative.

3.5.2 Training

For each dataset, we train the volume decoder (10 hidden layers of dimension 128) in minibatches of 10 images with random orientations for the first epoch to learn a volume with roughly correct spatial extent, followed by 4 epochs with branch and bound (BNB) pose inference (30 min/epoch noiseless, 80 min/noisy datasets). Since BNB pose inference is the bottleneck during training, we employ a multiscale training protocol, where after 4 epochs with BNB pose inference, the latent pose is fixed, and we train a separate, larger volume decoder (10 hidden layers of dimension 500) for 15 epochs with fixed poses to "refine" the structure to high resolution (20 min/epoch). Training times are reported for 50k, D=128 image datasets trained on a Nvidia Titan

V GPU.

3.5.3 Supplementary results

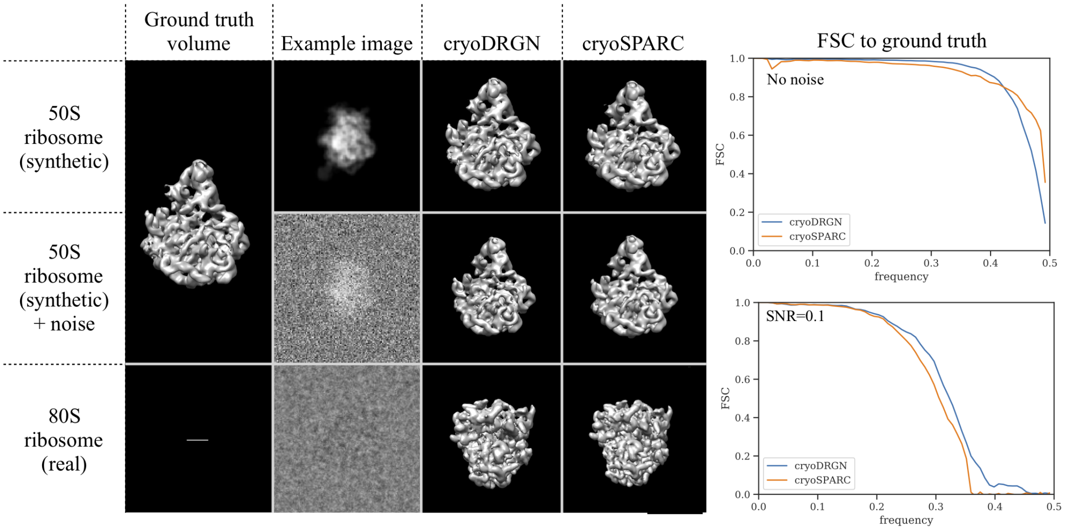


Figure 3-6: *Left:* CryoDRGN *ab initio* homogeneous reconstruction on 2 synthetic datasets and 1 real cryo-EM dataset matches the state of the art. *Right:* Fourier shell correlation (FSC) curves between the reconstructed volume and the ground truth volume for the synthetic ribosome datasets.

3.6 Heterogeneous ribosome reconstruction with pose supervision

Dataset preparation: We used the dataset from EMPIAR-10076 which contains 131,899 images of the *E. coli* large ribosomal subunit (LSU) in various stages of assembly [3]. Images were downsampled to D=128 by clipping in Fourier space. Poses were determined by aligning the images to a mature LSU structure obtained from a homogeneous reconstruction of the full resolution dataset in cryoSPARC, i.e. “a consensus reconstruction”.

Baseline: In the original analysis of this dataset, multiple rounds of multiclass refinement in sweeps of varying number of classes followed by expert manual alignment and clustering of similar volumes were used to identify 6 classes, labeled A-F consisting

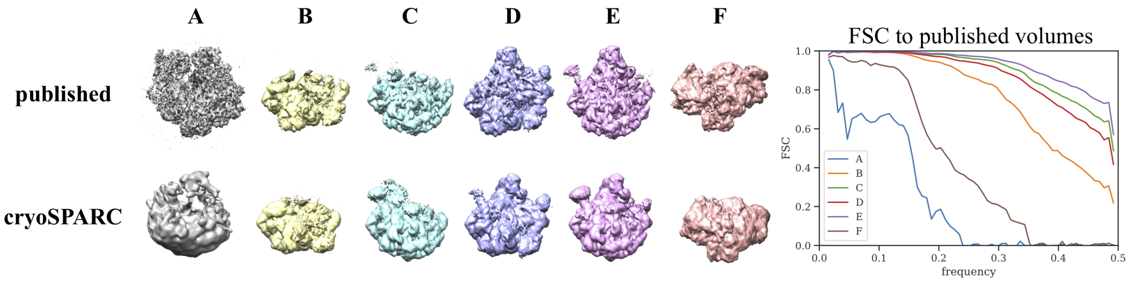


Figure 3-7: Reconstructed volumes from cryoSPARC multiclass refinement using the published structures of the 6 major states, low pass filtered to 25 Å as initial models. Right: FSC curves between the cryoSPARC reconstructed and published volumes.

of 4 major structural states of the LSU (classes B-E) and 2 additional structures of the 70S and 30S ribosome, class A and F, respectively.

Since the published dataset did not contain the corresponding image cluster assignments, we perform multiclass refinement in cryoSPARC using the published structures of the 6 major states, low pass filtered to 25 Å as initial models, to reproduce the results and obtain image cluster assignments. Aside from class A and F (low population impurities in the sample), the remaining structures correlate well with the published volumes (Figure 3-7).

cryoDRGN training: We train cryoDRGN with a 1-D latent variable in minibatches of 10 images for 200 epochs, treating image pose as fixed (11 min/epoch on a Nvidia Titan V GPU). To simplify representation learning for q_ξ, we center and phase flip images before inputting to the encoder. We encode and decode a circle of pixels with diameter D=128 instead of the full 128x128 image.

3.6.1 Supplementary results

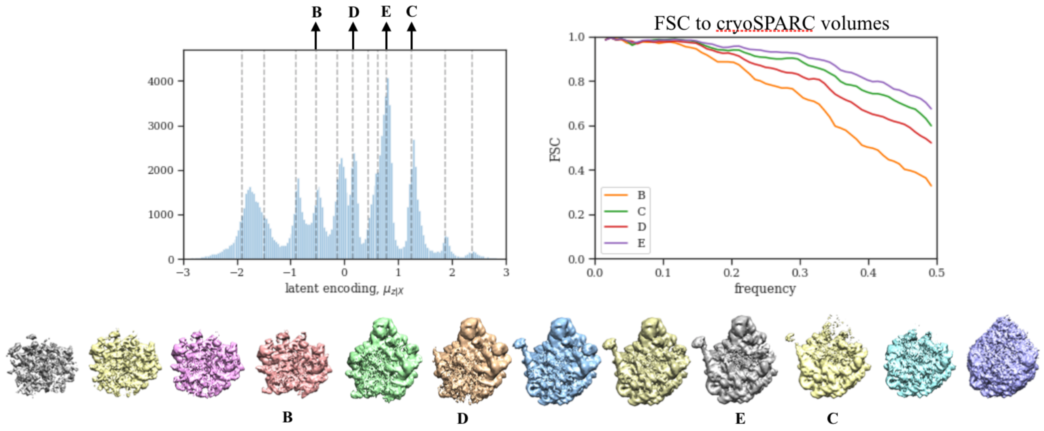


Figure 3-8: *Left:* Latent encoding for each image of the dataset from EMPIAR-10076. *Bottom:* Volumes from 12 sampled values along the latent space (dashed lines). *Right:* Fourier shell correlation (FSC) curves for 4 structures against the published volumes for classes B-E from corresponding to structural states of the large ribosomal subunit during assembly [3].

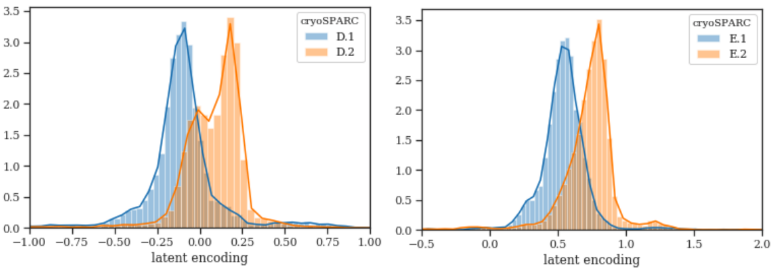


Figure 3-9: The latent encoding aligns with cluster assignments from a successive round of multiclass refinement in cryoSPARC on the subset of images from class D and E.

3.7 *Ab initio* heterogeneous reconstruction

3.7.1 Dataset preparation

Linear 1D motion: We generated a dataset containing one continuous degree of freedom as follows: From an atomic model of a protein complex, a single bond in the atomic model was rotated while keeping the remaining structure fixed, and 50 atomic models were sampled along this reaction coordinate. 1000 projections with random rotations and in-plane translations were generated for each model, yielding a total of 50k images, approximating a uniform distribution along a continuous reaction coordinate.

Linear 2D motion: We extended the linear 1D motion dataset by introducing a second degree of freedom from rotating a bond in the atomic model that connected a different protein in the complex. Similar to the 1D motion dataset, from a starting configuration, the original bond was rotated $+/-N$ degrees, and 50 models were sampled along this reaction coordinate. Then from the starting conformation, the second bond was rotated $+/-90$ degrees, and 50 additional models were sampled along the second reaction coordination. 500 projections were generated from each model, yielding a total of 50k images.

Circular 1D motion: For this dataset, we rotated a bond a full 360 degrees and sample 100 models along this circular reaction coordinate. 500 projections were generated from each model, yielding a total of 50k images.

Discrete 10 class: For this dataset, we sampled 10 random configurations for the proteins in the complex. 5000 projection images were generated from each model, yielding a dataset containing a mixture of 10 discrete states.

For all four datasets, random rotations were generated uniformly from $SO(3)$, and translations were sampled uniformly from $[-5, 5]$ pixels. The image size was D=64 with absolute spatial extent of 720 Å and Nyquist limit of 22.5 Å. A schematic of the simulated motions are shown in Figure 3-10.

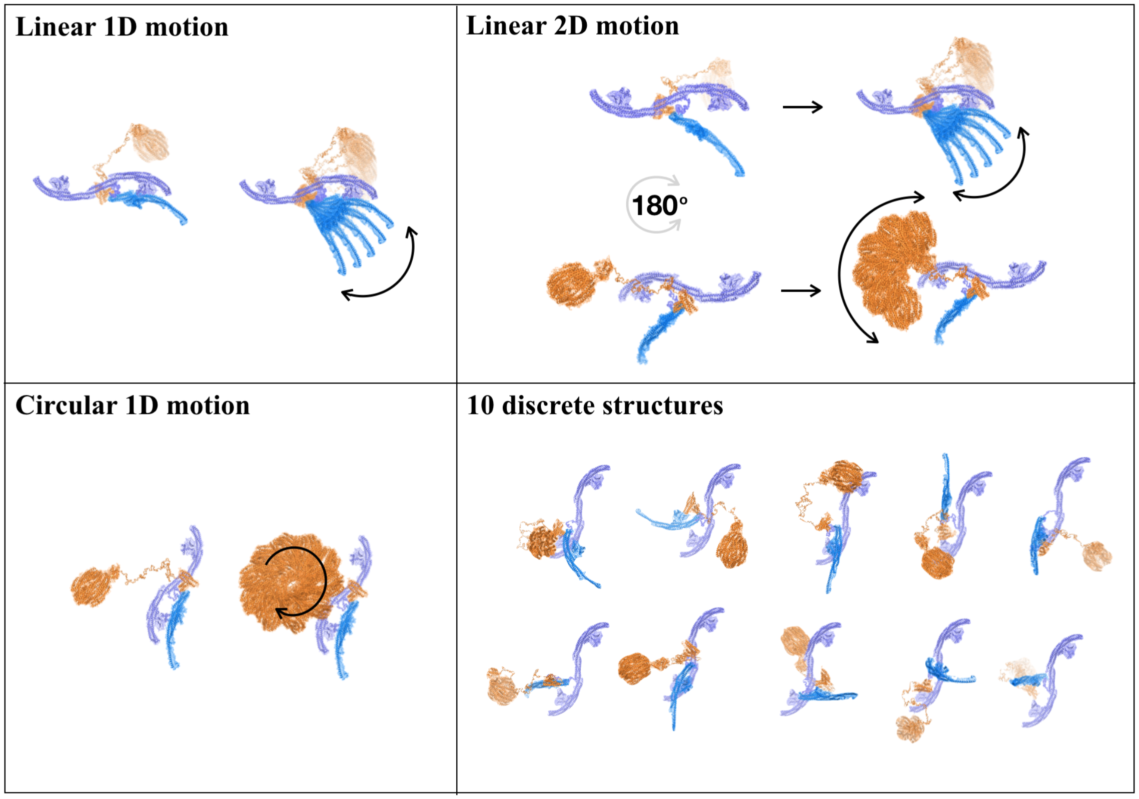


Figure 3-10: The ground truth atomic model and the heterogeneity introduced for synthetic datasets.

3.7.2 Tilt series pairs

Tilt series pairs is a variant of cryo-EM in which, for each image X_i, a corresponding image X_i' is acquired after tilting the imaging stage by a known angle. This technique was originally employed to identify the chirality of molecules [2], which is lost in the projection from 3D to 2D and therefore cannot be inferred from standard cryo-EM. Inferential procedures such as expectation maximization converge to one handedness or the other depending on their initialization. In multiclass reconstruction, different classes are not guaranteed to possess the same handedness even if there is a high relatedness between structures. We remark on this experimental technique as we propose using tilt series pairs to encourage invariance of q_ξ with respect to pose transformations for a given $\hat{V}_{\mathbf{z}}$ (and incidentally also to identify the chirality of $\hat{V}_{\mathbf{z}}$). To train on tilt series pairs, the encoder is split into two MLPs, the first learning an

	cryoSPARC			
Dataset	K=2	K=3	K=4	K=5
Linear 1D motion	5.11(3.82)	3.60(2.27)	7.40(4.16)	7.59(4.58)
Linear 2D motion	6.89(2.21)	6.90(3.77)	5.98(2.10)	6.76(4.47)
Circular 1D motion	5.16(2.70)	4.87(2.17)	7.50(3.32)	4.62(1.93)

Table 3.3: CryoSPARC reconstruction accuracy quantified by per-image FSC for different numbers of classes K. We report the average and standard deviation across 100 images in the dataset (lower is better; best possible is 2 pixels).

intermediate encoding of each image, and the second mapping the concatenation of the two encodings to the latent space. We use an 8 layer MLP with output dimension 128 for the former and a 2 layer MLP with input dimension 256 for the latter. All hidden layers have dimension 128. For branch and bound, the combined loss over both images is evaluated for each grid point of $SO(3) \times \mathbb{R}^2$. To generate the image $X_{tilt,i}$ associated with X_i, prior to rotating the volume by R_i, we rotate the volume by a constant 45 degrees around the x-axis.

3.7.3 Training

We trained cryoDRGN in minibatches of 5 images for 40 epochs without tilt series pairs and 20 epochs with tilt series pairs. We trained a 1-D latent variable for the linear 1D motion dataset, and 10-D latent variables for the remaining datasets. Random angles were used for the first epoch of training to learn roughly the correct spatial extent of the volume and BNB pose inference was used for the remaining epochs. The runtime was 120 min/epoch vs 2 min/epoch with and without BNB pose inference, respectively, on a Nvidia Titan V GPU.

3.7.4 Supplementary results

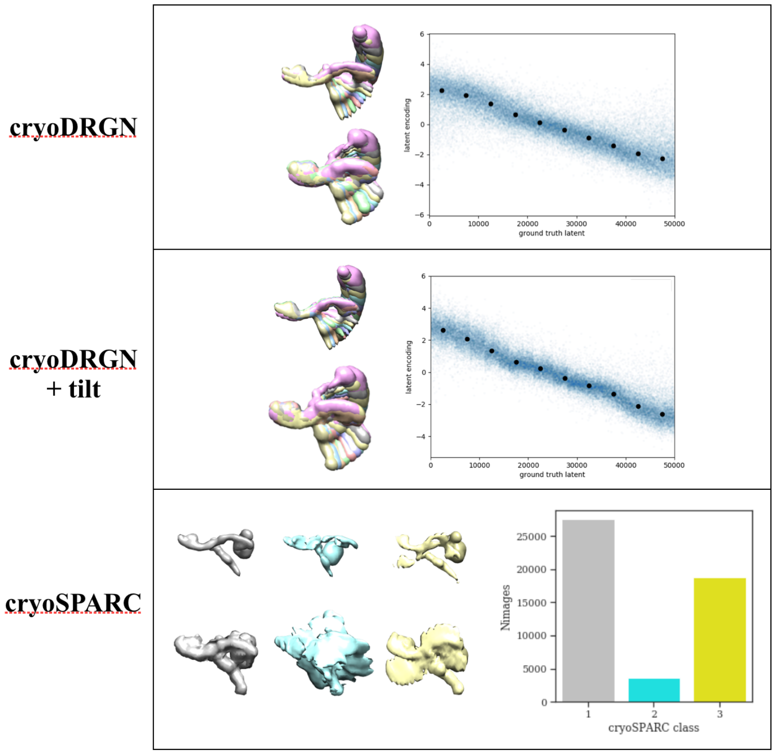


Figure 3-11: Reconstruction results for the linear 1D dataset by cryoDRGN and by discrete multiclass reconstruction in cryoSPARC. *Top:* Reconstructed structures from cryoDRGN sampled along the latent space (at depicted points) matches the ground truth variation. The predicted latent encoding correlates with the ground truth latent degree of freedom. *Middle:* CryoDRGN results with tilt series *Bottom:* Reconstructed volumes and the distribution of images over clusters from discrete multiclass reconstruction in cryoSPARC. Volumes are visualized at high and low isosurface, showing artifacts in the cryoSPARC structures.

	cryoDRGN			**cryoDRGN+tilt**		
Dataset	z-D=1	z-D=2	z-D=10	z-D=1	z-D=2	z-D=10
Linear 1D motion	2.50(0.62)	2.34(0.12)	–	2.35(0.36)	2.43(0.26)	–
Linear 2D motion	7.16(4.69)	4.38(3.15)	4.44(2.50)	3.38(1.18)	2.97(1.24)	2.93(1.02)
Circular 1D motion	5.61(4.36)	4.95(2.91)	4.05(2.40)	3.12(0.96)	2.65(0.67)	2.63(0.74)

Table 3.4: CryoDRGN reconstruction accuracy quantified by per-image FSC with different dimensions for z. We report the average and standard deviation across 100 images in the dataset (lower is better; best possible is 2 pixels).

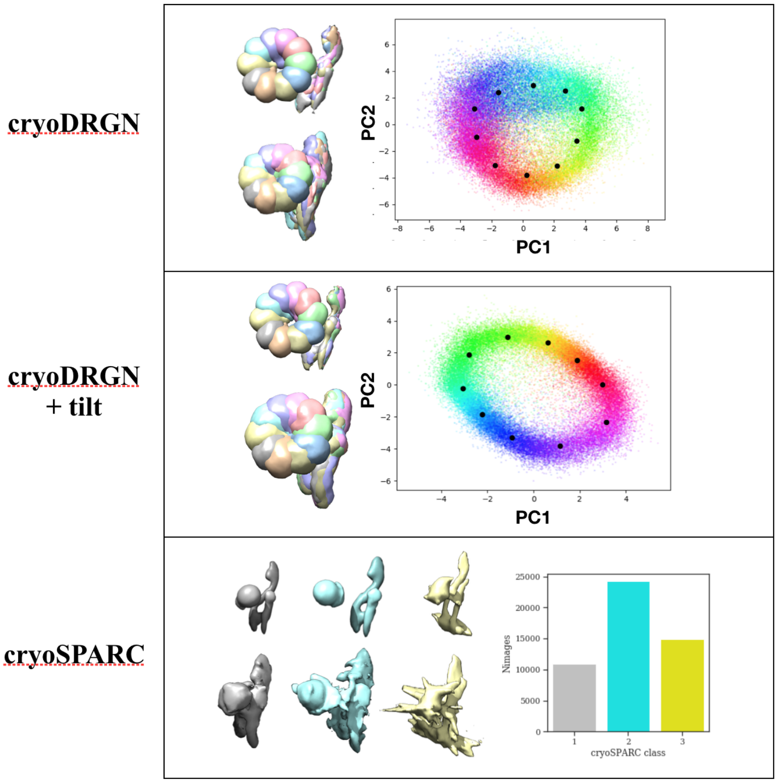


Figure 3-12: Reconstruction results for the circular 1D dataset by cryoDRGN and by discrete multiclass reconstruction in cryoSPARC. *Top:* Reconstructed structures from cryoDRGN sampled along the latent space (at depicted points) matches the ground truth variation. The distribution of images in the latent space matches the circular topology of the true data manifold. *Middle:* CryoDRGN results with tilt series *Bottom:* Reconstructed volumes and the distribution of images over clusters from discrete multiclass reconstruction in cryoSPARC. Volumes are visualized at high and low isosurface, showing artifacts in the cryoSPARC structures.

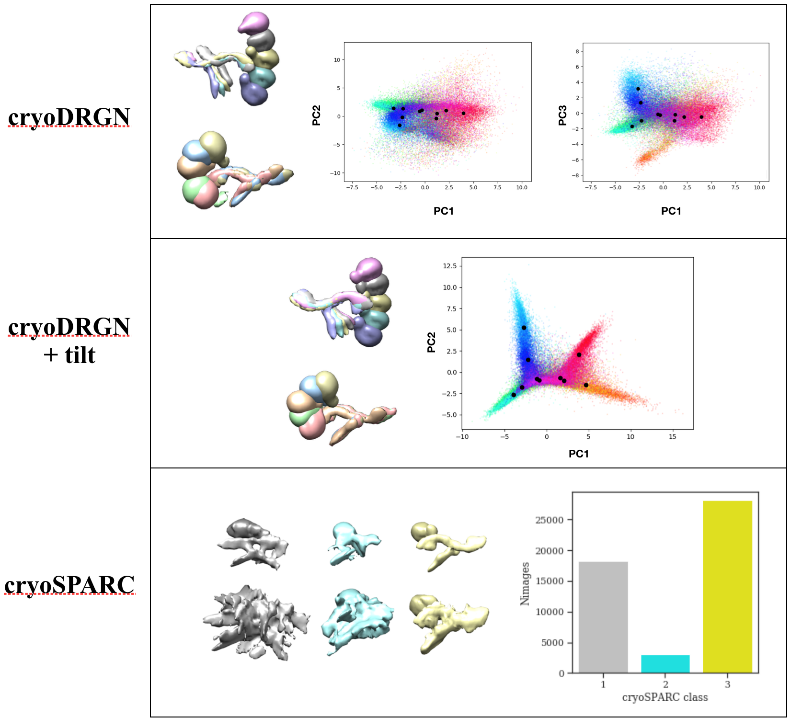


Figure 3-13: Reconstruction results for the linear 2D dataset by cryoDRGN and by discrete multiclass reconstruction in cryoSPARC. *Top:* Reconstructed structures from cryoDRGN sampled along the latent space (at depicted points) roughly matches the ground truth variation, however the distribution of images in the latent space does not recapitulate the true data manifold well. *Middle:* CryoDRGN results with tilt series reconstructs the true structural variation and the distribution of images in the latent space matches the topology of the true data manifold. *Bottom:* Reconstructed volumes and the distribution of images over clusters from discrete multiclass reconstruction in cryoSPARC. CryoSPARC volumes are visualized at high and low isosurface, showing artifacts at low isosurface

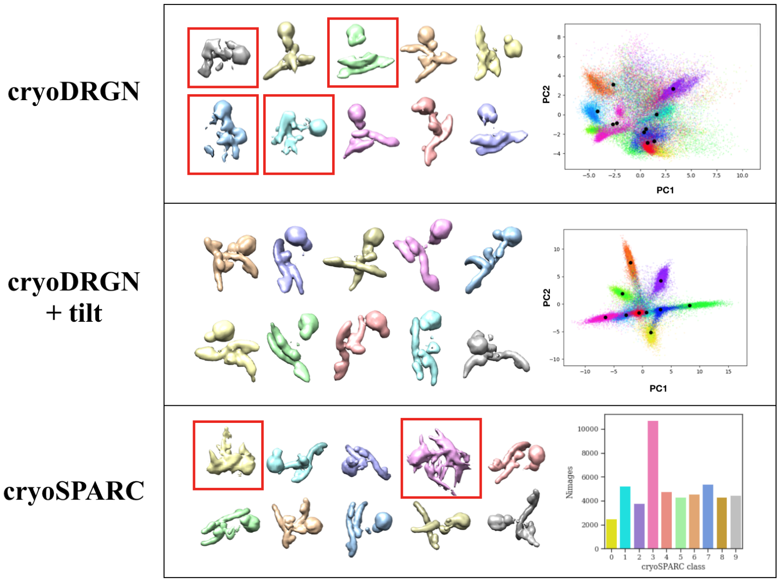


Figure 3-14: Reconstruction results for the dataset containing 10 discrete structures by cryoDRGN and by discrete multiclass reconstruction in cryoSPARC. *Top:* The majority of reconstructed structures from cryoDRGN sampled along the latent space (at depicted points) matches the ground truth structures, however some are incorrect (red boxes), and the learned data manifold is not well separated into clusters. *Middle:* CryoDRGN results with tilt series reconstructs the 10 structures and clusters the images in the latent space accordingly. *Bottom:* Reconstructed volumes from discrete multiclass reconstruction in cryoSPARC and the distribution of images over clusters. CryoSPARC learns 8 out of 10 structures correctly.

Bibliography

[1] William T Baxter, Robert A Grassucci, Haixiao Gao, and Joachim Frank. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *Journal of structural biology*, 166(2):126–132, May 2009.

[2] D M Belnap, N H Olson, and T S Baker. A method for establishing the handedness of biological macromolecules. *Journal of structural biology*, 120(1):44–51, October 1997.

[3] 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.

[4] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

[5] Ralph VL Hartley. A more symmetrical fourier analysis applied to transmission problems. *Proceedings of the IRE*, 30(3):144–150, 1942.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[8] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature methods*, 14(3):290–296, March 2017.

[9] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. June 2020.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Wilson Wong, Xiao-Chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H W Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e01963, June 2014.

[12] Anna Yershova, Swati Jain, Steven M LaValle, and Julie C Mitchell. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *The International Journal of Robotics Research*, 29(7):801–812, May 2010.

[13] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

Chapter 4

CryoDRGN: Applications and software

This chapter explores the application of cryoDRGN to real cryo-EM datasets. We first show that our neural network representation of structure can model single density maps at high resolution, before demonstrating the full cryoDRGN framework for unsupervised heterogeneous reconstruction.

We find that cryoDRGN is a powerful and general approach for analyzing structural heterogeneity in macromolecular complexes of varying size and expected sources of heterogeneity. We show cryoDRGN can uncover residual heterogeneity in "homogeneous" datasets of the RAG1-RAG2 complex and the 80S ribosome, model large compositional changes of the assembling 50S ribosome, and continuous conformational changes of the precatalytic spliceosome. Remarkably, cryoDRGN's unsupervised approach for representation learning can readily identify and filter impurities in the dataset and can identify rare structural states containing as few as 1,000 particles. CryoDRGN is distributed as an open-source tool that can be easily integrated in existing pipelines and is freely available at cryodrgn.csail.mit.edu.

This chapter presents work described in [48] performed jointly with Tristan Bepler, Bonnie Berger, and Joey Davis.

4.1 Introduction

Proteins and their complexes are dynamic macromolecular machines that carry out the essential biological processes responsible for life. Although the mechanism of these macromolecular machines is often deduced from a static three-dimensional (3D) structure, a more complete understanding could be achieved if one could analyze the full distribution of conformations relevant to function.

Single particle cryo-electron microscopy (cryo-EM) is a rapidly maturing method for high-resolution structure determination of large macromolecular complexes [28, 9]. Major advances in hardware [5, 41, 21] and software [41, 21, 47, 8, 39, 6] have streamlined the collection and analysis of cryo-EM datasets such that structures of rigid macromolecules can routinely be solved at near-atomic resolution [4, 46]. Increasingly, cryo-EM has been applied to study heterogeneous complexes as the experimental procedure is less sensitive to sample heterogeneity than other methods for structure determination [11, 15]. Additionally, because single particle cryo-EM can capture millions of snapshots of the molecule of interest, each carrying a unique molecule in its own conformational state [40], cryo-EM holds promise in revealing the conformational landscape of dynamic macromolecular complexes. However, reconstructing ensembles of 3D volumes from such snapshots remains a major computational challenge.

Existing tools for heterogeneous reconstruction make often-limiting assumptions on the observed structural heterogeneity. Most commonly, heterogeneity is modeled as though it originates from a small number of independent, discrete states, implemented as "3D classification" or "heterogeneous refinement" in many cryo-EM software packages [39, 33, 23, 14]. However, these discrete classification approaches require specifying initial models for refinement, and because the number and nature of the underlying structural states is unknown *a priori*, this approach is error-prone and often results in the omission of potentially relevant structures. More critically, such discrete approaches are ill-suited for reconstructing structures undergoing continuous conformational changes.

Advanced methods for heterogeneous reconstruction seek to more closely model the continuous nature of flexible molecules. Multi-body refinement, available in RELION, models the structure as the sum of user-defined rigid bodies that are allowed to rotate relative to one another, placing structural assumptions on the observed heterogeneity [27]. Continuous heterogeneity has also been described using principle component analysis (PCA)-based approaches [22, 30, 43], including the recent 3D Variability Analysis (3DVA) algorithm available in cryoSPARC [35]. Although the linear subspace model of these approaches can provide a summary of the overall variability within the dataset, the visualized heterogeneity contains artifacts when a molecule's conformational deformations are poorly approximated by linear interpolations along basis volumes. In the manifold embedding approach proposed in Dashti et al. [10, 12], heterogeneous structures are recovered by binning particles along the data manifold followed by traditional homogeneous reconstruction. Additional algorithms for continuous heterogeneous reconstruction have been shown on synthetic datasets [20, 25].

Here, we present cryoDRGN (Deep Reconstructing Generative Networks), a method for heterogeneous cryo-EM reconstruction based on deep neural networks. We hypothesized that neural networks, which are known for their ability to model complex, nonlinear functions [16], could learn heterogeneous ensembles of cryo-EM density maps. We first show that our neural network representation of structure can model single density maps at high resolution, before demonstrating the full cryoDRGN framework for unsupervised heterogeneous reconstruction.

We find that cryoDRGN is a powerful and general approach for analyzing structural heterogeneity in macromolecular complexes of varying size and expected sources of heterogeneity. We show that the cryoDRGN approach can uncover residual heterogeneity in "homogeneous" datasets of the RAG1-RAG2 complex and the 80S ribosome, model large compositional changes of the assembling 50S ribosome, and continuous conformational changes of the precatalytic spliceosome. Remarkably, cryoDRGN's unsupervised approach for representation learning can readily identify and filter impurities in the dataset and can identify rare structural states containing as few as 1,000 particles. CryoDRGN is distributed as an open-source tool that can be easily integrated in existing pipelines and is freely available at cryodrgn.csail.mit.edu.

4.2 Results

4.2.1 The cryoDRGN method

CryoDRGN performs heterogeneous reconstruction by learning a deep generative model of 3D structure from single particle cryo-EM images. The method consists of a specialized image encoder – volume decoder architecture, which learns an encoding of 2D particle images into a continuous vector space described by the latent variable $z \in \mathbb{R}^n$ (i.e. the latent space), and the concomitant reconstruction of 3D cryo-EM density maps from this latent space representation (Fig 4-1A). This choice of model assumes that the heterogeneous structures can be embedded within a continuous, low-dimensional manifold in the latent space, where the dimensionality of the latent space is defined by the user. The model is specified in the Fourier domain in order to relate 2D images as planar slices of the 3D volume [7], whose orientation is previously determined from a consensus reconstruction. The neural networks are jointly trained from random initialization using stochastic gradient descent on an objective function that seeks to maximize (a variational lower bound on) the data likelihood as in standard Variational Autoencoders (VAEs) [19]. Additional architectural and training details of cryoDRGN are provided in the **Methods (Section 4.4)**.

After training, the output of cryoDRGN analysis includes: 1) per-particle latent encodings, z_i, describing the dataset's heterogeneity and 2) a neural network model of 3D density maps that can directly reconstruct a density map given z_i. Specifically, the encoder network encodes particle images into the continuous latent space, which allows for visualization and inspection of the particle distribution (**Fig. 4-1B, center**). The trained decoder network can then generate 3D density maps given arbitrary values of the latent variable. For example, representative structures can be generated from regions of latent space with high particle density, and continuous conformational trajectories can be reconstructed by sampling points along a trajectory through

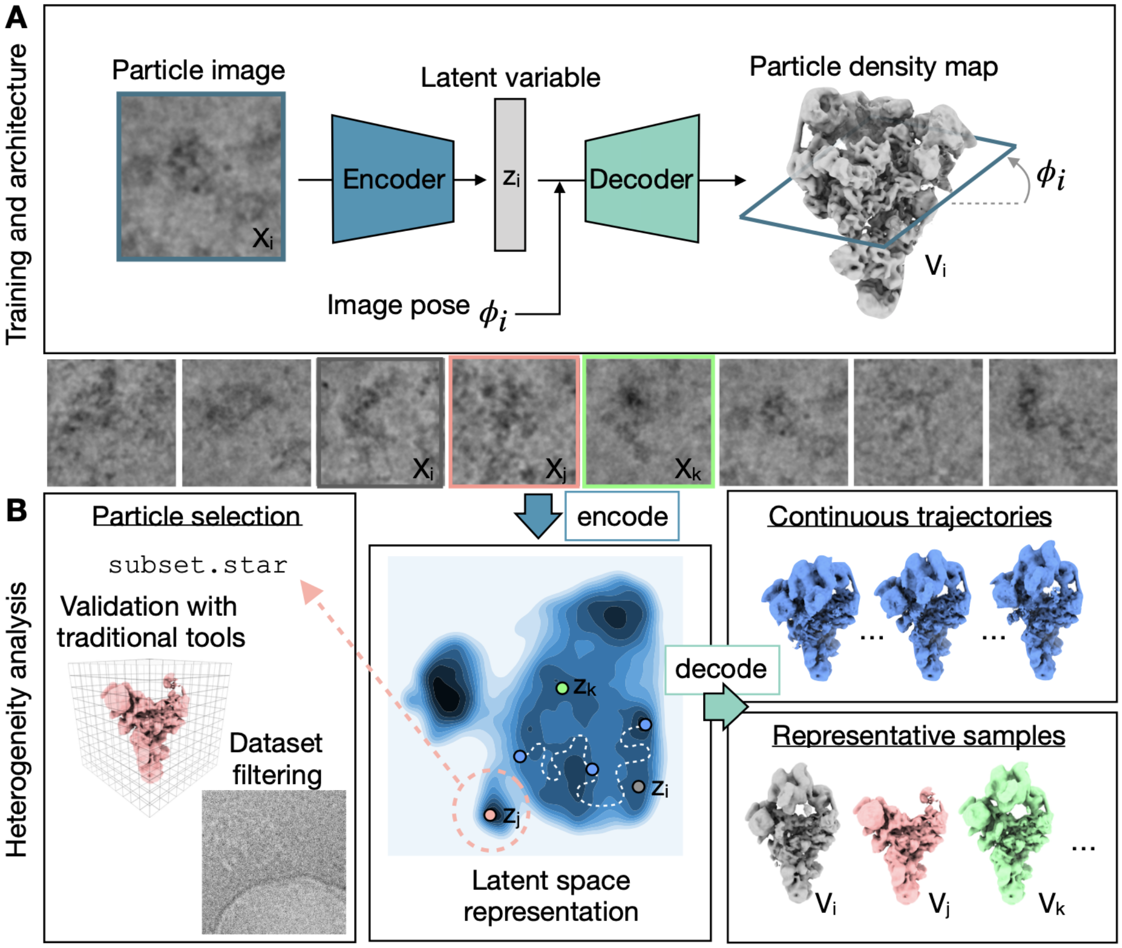


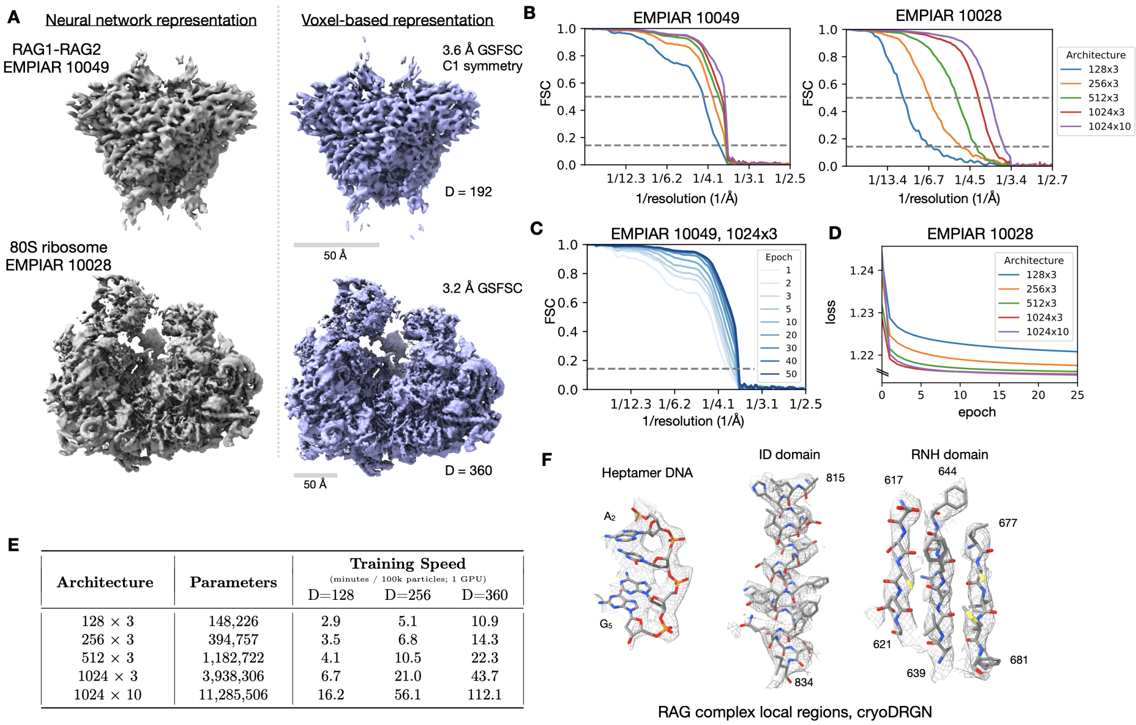
Figure 4-1: **A**, The cryoDRGN model consists of two neural networks structured in an image-encoder–volume-decoder architecture with a continuous latent variable representation of heterogeneity. During training, each particle image is encoded into the low-dimensional latent space and then reconstructed as its corresponding model slice based on the Fourier slice theorem. Image and volume data are depicted in real space for visual clarity. **B**, Once a cryoDRGN model is trained, the full dataset of particle images is encoded into the latent space, which is visualized here as a contour map with darker regions corresponding to higher particle density (center). The decoder, which represents an ensemble of 3D density maps, can directly generate density maps from arbitrary values of the latent variable (right). The particle stack may also be filtered using the latent space representation for validation of specific structures with traditional tools or to remove impurities from the dataset (left). Example images are from EMPIAR-10180 [32]

latent space (**Fig. 4-1B, right**). Notably, any cryoDRGN-generated volume can be orthogonally validated by traditional reconstruction approaches [39] using nearby particles in the latent space (**Fig. 4-1B, left**). Lastly, any regions of the latent space that are enriched in impurities or imaging artifacts may be selected, and the encompassed particles filtered from subsequent analysis (**Fig. 4-1B, left**).

4.2.2 Neural networks can represent cryo-EM density maps

We first evaluated the cryoDRGN volume decoder in representing high resolution cryo-EM density maps. To learn the homogeneous structure of the RAG1-RAG2 signal end complex (RAG – 369 kDa) [37] and the Plasmodium falciparum 80S ribosome (Pf80S – 4.2 MDa) [45], we trained the volume decoder network with no latent variable input, using image poses obtained from C1 homogeneous refinements in cryoSPARC [33] (**Methods, Section 4.4**). Trained on full-resolution images, the cryoDRGN decoder produced structures that correlated with the traditional, voxel-based reconstruction (**Fig. 4-2A**) at resolutions up to 3.6 Å for RAG and 3.9 Å for Pf80S at an FSC = 0.5 threshold (**Fig. 4-2B**), demonstrating the efficacy of this neural-network based representation of 3D structure.

As neural networks have a fixed capacity for representation that is constrained by their architecture, we next compared decoder architectures of different sizes to evaluate the tradeoff between representation power and training speed. We found that larger architectures, which have more trainable parameters, result in density maps that correlate with the traditionally reconstructed map at higher resolutions (**Fig. 4-2B**). The networks are trained through multiple passes through the dataset (i.e. epochs) (b) with lower values of the objective function (**Fig. 4-2D**) as training progressed. Notably, while the resolution of the learned structure increased with neural network size, we found that larger models were slower to train (**Fig. 4-2E**). These tradeoffs suggest that the architecture and image size should be tuned to suit the desired balance of speed and achievable resolution. Lastly, we found that the cryoDRGN architecture was capable of learning density maps at sufficiently high resolution to visualize structural features such as bulky side-chains that are consistent



Architecture	Parameters	Training Speed (minutes / 100k particles; 1 GPU) D=128	 D=256	 D=360
128 × 3	148,226	2.9	5.1	10.9
256 × 3	394,757	3.5	6.8	14.3
512 × 3	1,182,722	4.1	10.5	22.3
1024 × 3	3,938,306	6.7	21.0	43.7
1024 × 10	11,285,506	16.2	56.1	112.1

Figure 4-2: A, Density maps of the RAG1–RAG2 complex (EMPIAR-10049) [37] and of the eukaryotic Pf80S ribosome (EMPIAR-10028) [45] reconstructed by cryoDRGN's decoder neural network (left) and a traditional, voxel-based reconstruction in cryoSPARC (right). The cryoDRGN volumes were generated from decoder networks with three hidden layers and 1,024 nodes per hidden layer (denoted as $1,024 \times 3$) trained for 25 epochs. b, FSC curves between density maps produced by the cryoDRGN decoder with varying architectures and the traditional reconstruction in a. c,d, Evolution of the FSC curve in b and the training curve over multiple epochs of cryoDRGN model training. e, Training speed in min per 105 images for cryoDRGN decoder networks of different architectures on different image sizes (D, in pixels) on a single Nvidia V100 graphics processing unit (GPU). The number of trainable parameters is specified for decoder networks trained on $D = 256$ images. f, Representative regions (insertion domain (ID), RNase H-like domain (RNH)) of the RAG1–RAG2 density map from cryoDRGN in a superimposed with the published atomic model (Protein Data Bank (PDB) 3JBX).

with our FSC-based resolution estimates (**Fig. 4-2F**).

4.2.3 CryoDRGN models both discrete and continuous structural heterogeneity

We next sought to evaluate the complete cryoDRGN framework for heterogeneous reconstruction using simulated datasets (Fig 4-3A,B). Datasets modelling continuous heterogeneity were produced by rotating a single dihedral angle of a hypothetical protein complex to simulate a conformational transition along a 1-dimensional reaction coordinate. Single particle cryo-EM images were then simulated either: uniformly along this reaction coordinate (*Uniform*); with bias towards particular conformations exemplary of cooperative transitions (*Cooperative*); or with strong bias leading to unobserved transition states (*Noncontiguous*). A dataset simulating discrete compositional heterogeneity was produced by mixing particles of the bacterial 30S, 50S, and 70S ribosome (*Compositional*). We then provided each of these four simulated datasets and their corresponding poses to cryoDRGN and trained a 1-D latent variable model ($|z| = 1$) (**Methods**).

We found that cryoDRGN was capable of reconstructing both continuous and discrete heterogeneous ensembles (**Fig. 4-3C-H**). On the *Uniform* conformational heterogeneity dataset, cryoDRGN reconstructed density maps that reproduced the ground truth continuous motion of the complex (**Fig. 4-3C**). When trained on the *Compositional* dataset, cryoDRGN reconstructed density maps of the 30S, 50S, and 70S ribosomes at distinct values of the latent variable (**Fig. 4-3D**).

In addition to reconstructing heterogeneous density maps, cryoDRGN produces a latent encoding for each particle that can be compared to the ground truth reaction coordinate (**Fig. 4-3E-H**). For the datasets with continuous conformational changes, the latent encoding of each image correlated with position along the reaction coordinate given by the dihedral angle of the underlying model (Spearman r = -0.996, 0.992, and 0.988 for *Uniform*, *Cooperative*, and *Noncontiguous*, respectively) (**Fig. 4-3E**). We observed that the qualitative features of the distribution of latent encodings

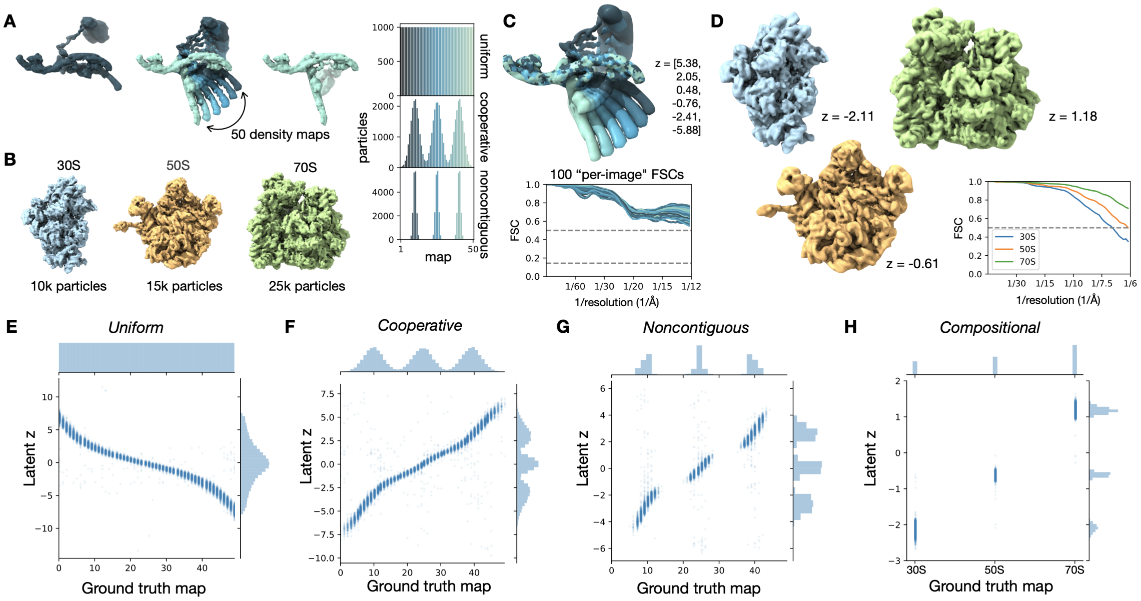


Figure 4-3: **A**, Ground-truth density maps simulating continuous heterogeneity generated by sampling conformations along a 1D conformational transition from the leftmost to the rightmost structure (left). Particles along this conformation transition were sampled uniformly (top) or from a mixture of Gaussian distributions of varying widths (middle, bottom) to simulate various degrees of cooperative transitions between three states. **B**, Compositional heterogeneity simulated by mixing particles of the 30S, 50S and 70S bacterial ribosomal complexes. **C**, Density maps reconstructed by cryoDRGN trained on the uniformly sampled dataset in **A**. Six structures were sampled from the specified values of the latent variable (top). "Per-image FSC" curves are shown, where for 100 images equally spaced along the reaction coordinate, we computed the FSC between a map generated by cryoDRGN at the predicted latent encoding for each image and the ground-truth density map for that image (bottom). See **Methods (Section 4.4)** for description of the "per-image FSC" approach. **D**, Density maps reconstructed by cryoDRGN from the compositional dataset in **B** and their FSCs to the corresponding ground-truth density map. **E–H**, Predicted latent space encoding for each particle image of different simulated datasets versus the ground-truth reaction coordinate describing the motion (**E–G**) or the ground-truth class assignment (**H**). All cryoDRGN reconstructions use a 1D latent variable model.

matched the ground truth, with three modes in the latent encoding distribution for the Cooperative dataset (**Fig. 4-3F**) and distinct clusters for the *Noncontiguous* and *Compositional* datasets (**Fig. 4-3G, H**). We note that, in general, the parameterization of a reaction coordinate is non-unique (e.g. when described by the learned latent variable or by dihedral angle, leading to different marginal distributions in **Fig. 4-3E**). To quantitatively assess that cryoDRGN has learned the correct distribution of structures, we compute a "per-image FSC", which compares reconstructed density maps with the ground-truth on images across the reaction coordinate (**Methods, Section 4.4**), and found that the reconstructed structures of all four datasets correlated well with the ground truth distribution (**Fig. 4-3C, Supplementary Fig. 4-7**).

4.2.4 CryoDRGN uncovers residual heterogeneity from "homogeneous" cryo-EM datasets

We next evaluated cryoDRGN's ability to perform heterogeneous reconstruction on real cryo-EM datasets of the RAG complex [37] and the Pf80S ribosome [45] from above. Ru et al. reported RAG complex structures from two distinct datasets – the "signal end complex", which failed to resolve the distal ends of the 12-RSS and 23-RSS DNA elements or the nonamer binding domain (NBD) of RAG-1; and the "paired complex", which resolved these elements at sufficient resolution for atomic model building (**Fig. 4-4A**). To test whether cryoDRGN could newly uncover heterogeneity of these distal elements in the "signal end complex", we trained a cryoDRGN 10-D latent variable model on the deposited particle images (EMPIAR-10049) [37]. We found that cryoDRGN revealed significant heterogeneity of the 12-RSS, 23-RSS, and NBD (**Fig. 4-4B**). In addition to maps that only resolve the symmetric core (light grey in **Fig. 4-4B**), cryoDRGN revealed structures with RSS positioning that aligns with the canonical conformation found in the "paired complex" atomic model [37] (dark blue), tilting of the RSS strands (light blue), linear 23-RSS DNA (purple), as well as the presence (yellow) and absence (coral) of the NBD. These representative maps were selected out of a large ensemble of generated structures (**Methods, Section**

4.4) from different regions of the latent space (**Fig. 4-4C**). A trajectory sampling from the continuous distribution modeled by cryoDRGN is shown in Supplementary Video 1 of Zhong et al. [48]. To validate the presence of these heterogeneous states in the dataset, we performed heterogeneous 3D refinement in cryoSPARC [33] using the cryoDRGN density maps as initial models, which reproduced the heterogeneity of the RSS elements (**Supplementary Fig. 4-8**). Subsequent work by Ru et al. suggested that the conformational dynamics and asymmetric positioning of the 12- and 23-RSS by NBD in the pre-cleavage form are fundamental to the structural mechanism underlying the 12-23 rule of V(D)J recombination [38]. Our results newly demonstrate that such heterogeneity is also present in the post-cleavage "signal end" RAG complex.

When analyzing a homogeneous reconstruction of the Pf80S ribosome, Wong et al. observed flexibility in the small subunit head region and missing density for peripheral rRNA expansion segment elements [45]. To explore if this unresolved density resulted from residual heterogeneity, we trained a cryoDRGN 10-D latent variable model on their deposited dataset (EMPIAR-10028) [45] and reconstructed an ensemble of density maps that not only contained structures consistent with the homogeneous reconstruction, but also revealed rotation of the 40S small subunit (SSU) (**Fig. 4-4D**), heterogeneity within the SSU head (**Supplementary Fig. 4-9**), and motion of many peripheral rRNA expansion segments (See Supplementary Video 2 of Zhong et al. [48]). By visualizing a representative 40S rotated and unrotated density map, we found that cryoDRGN was able to simultaneously capture the large-scale inter-subunit rotation and coordinated smaller-scale structural rearrangements, including motion of the L1 stalk, disappearance of an rRNA helix, and the disappearance of the inter-subunit bridge formed by the C-terminal helix of eL8, which is consistent with Sun et al.'s characterization of Pf80S dynamics [42] (**Fig. 4-4D**).

We then visualized the 10-D latent space representation of the Pf80S particles with PCA (**Fig. 4-4E**) and with Uniform Manifold Approximate and Projection (UMAP) [24] (**Supplementary Fig. 4-10**). The 40S rotated density map originated from a region of the particle distribution separated along the first PC of the latent space. To validate the presence of this state, we extracted 4,889 particles constituting the

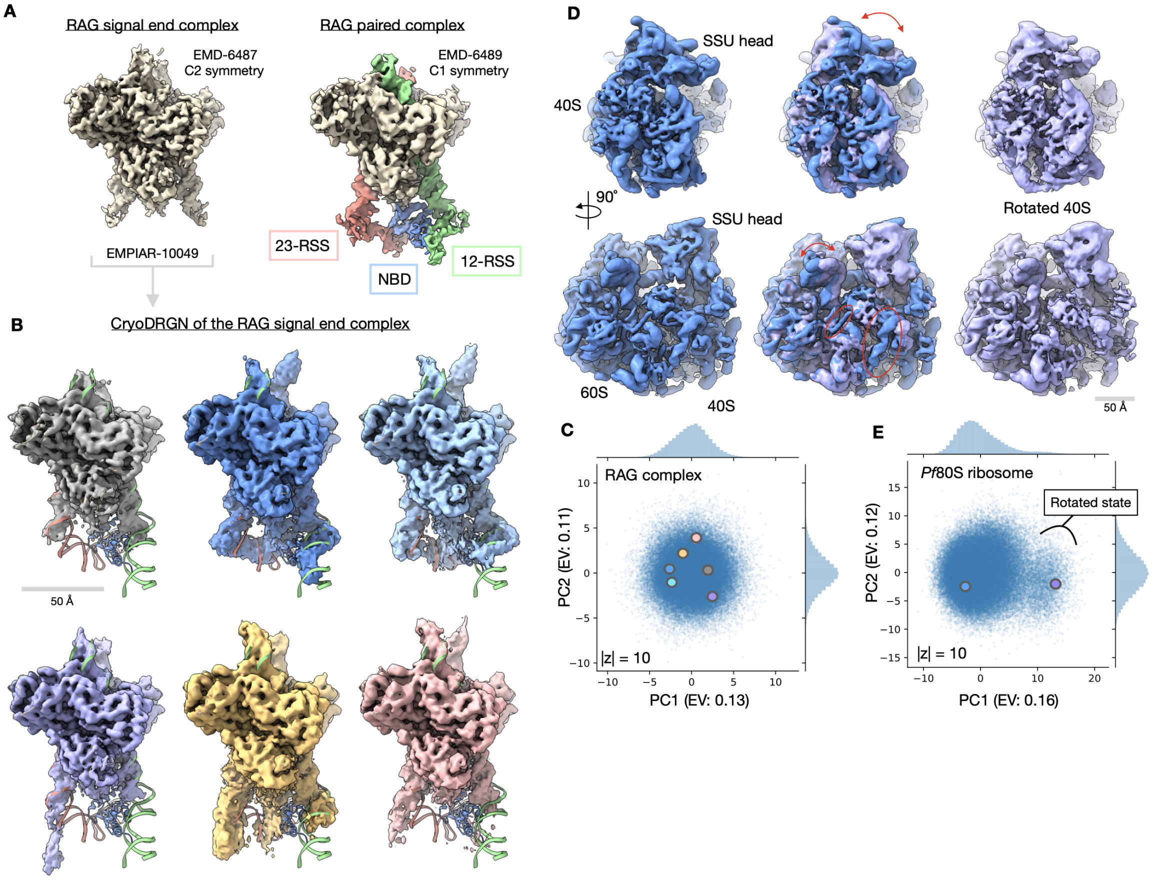


Figure 4-4: **A**, Published density maps of the 369-kDa RAG1–RAG2 complex. The signal-end complex (left) shows the C2 symmetric core, and the paired complex (right) resolves additional asymmetric 12- and 23-RSS DNA elements and the RAG1 NBD that extend below the core. **B**, Representative density maps of the RAG signal-end complex (EMPIAR-10049) [37] reconstructed by cryoDRGN. Density maps resolve variable conformations of the 12- and 23-RSS DNA elements and the NBD, which are missing from the homogeneous refinement. The docked atomic model (PDB 3JBW) of the RAG paired complex includes an asymmetric conformation of the RSS and NBD elements that extend from the core RAG complex. **C**, Latent space representation of particle images from the EMPIAR-10049 dataset [37], visualized using PCA with explained variance (EV) noted. Structures from b are marked with the corresponding color. **D**, Density map of the 4.2-MDa Pf80S ribosome (EMPIAR-10028) [45] in an unrotated (blue) and rotated (purple) state reconstructed by cryoDRGN. Arrows indicate rotation of the 40S subunit relative to the 60S subunit (top) and motion of the L1 stalk (bottom). Circles indicate differential occupancy of the C-terminal helix of eL8 and an rRNA helix between the two states. **E**, Latent space representation of particle images from the EMPIAR-10028 dataset [45], visualized using PCA with explained variance noted. Structures from **D** are marked with the corresponding color. A cluster of particles separated along PC1 of **D** that corresponds to the rotated state of the Pf80S ribosome is noted. Additional density maps from these datasets are shown in **Supplementary Data Fig. 4-9**.

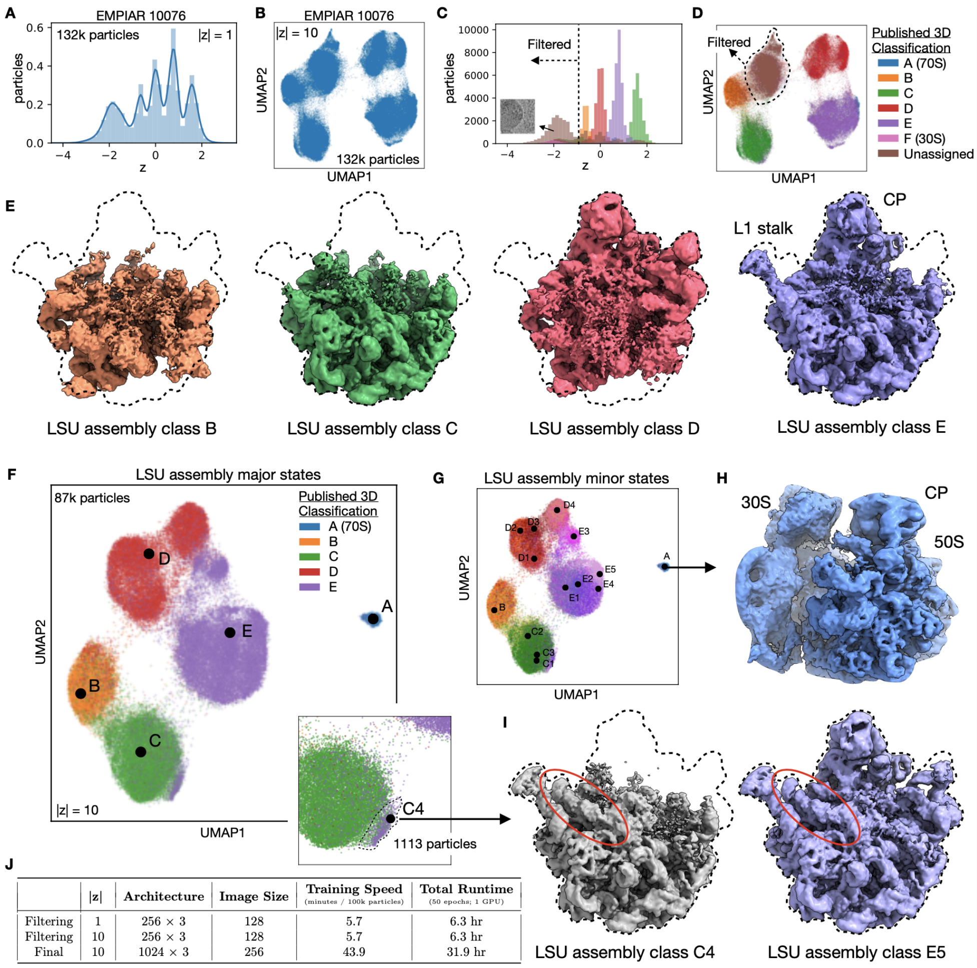
outlying cluster (**Methods, Section 4.4**). Traditional homogeneous reconstruction of these particles in cryoSPARC produced a 6.4 Å reconstruction of the rotated 40S state consistent with the cryoDRGN structure (**Supplementary Fig. 4-10**). Additionally, by sampling many density maps from the latent space, we observed that structures with density missing from the SSU head group were located within a subregion of the main cluster of the UMAP visualization (**Supplementary Fig. 4-9**). We hypothesize that the 40S rotated state appears as a visually distinct cluster because more mass changes to rotate the entire 40S subunit, as opposed to the missing SSU head group state, which involves changes in a smaller region of the 40S subunit.

4.2.5 CryoDRGN automatically partitions assembly states of the bacterial ribosome

Next, we sought to evaluate cryoDRGN on a highly heterogeneous cryo-EM dataset of the *E. coli* large ribosomal subunit (LSU) undergoing assembly (EMPIAR-10076) [11]. This dataset is known to contain substantial compositional and conformational heterogeneity; In the original analysis, multiple expert-guided rounds of hierarchical 3D classification resulted in 13 discrete structures that were grouped into 4 major assembly states. Here, we aimed to assess if cryoDRGN could automatically reveal these heterogeneous states without user-guided 3D classification.

As initial pilot experiments, we first trained 1-D and 10-D latent variable models on downsampled images of the dataset (**Methods, Section 4.4**). The dataset's latent space representation exhibited distinct peaks in the 1-D case or clusters in the 10-D case when visualized with UMAP [24] (**Fig. 4-5A,B**) that correspond to the major assembly states when grouped by the published 3D classification labels (**Fig. 4-5C,D**). As the particles were obtained by crudely fractionating a lysate in order to capture the full ensemble of cellular assembly intermediates, a substantial fraction of the published particle stack corresponds to 30S or non-ribosomal impurities. These unassigned particles were outliers in the latent representation **(Fig. 4-5C, Supplementary Fig. 4-11**), and neither 2D class averages nor a traditional 3D reconstruction of these particles produced structures consistent with assembling LSU ribosomes (**Supplementary Fig. 4-5**). As we did not wish to devote representation capacity of the cryoDRGN neural networks in modelling these impurities, we used the latent representation to filter the dataset before further analysis, taking the intersection of the particle stack after filtering based on the 1-D and 10-D latent variable model (**Methods, Section 4.4**).

To explore the heterogeneity within the LSU assembly states, we trained a cryoDRGN 10-D latent variable model on the remaining images at higher resolution (**Methods, Section 4.4**). The decoder network reconstructed density maps matching the reported major (**Fig. 4-5E**) and minor (**Supplementary Fig. 4-12**) assembly states of the LSU. We visualized the encodings of particle images in the 10-D latent space with UMAP and observed clusters corresponding to the major (**Fig. 4-5F**) and minor states (**Fig. 4-5G, Supplemental Fig. 4-12**) of LSU assembly after coloring by the published 3D classification. From the latent representation, we also noted a clearly separated cluster of particles assigned to class A, and structures sampled from this region of latent space reconstructed the 70S ribosome, an impurity in the dataset (**Fig. 4-5H**). Finally, we identified a small cluster of 1,100 particles adjacent to the class C cluster whose particles were originally classified into class E (**Fig. 4-5F, inset**). The density map reconstructed by the decoder from this region revealed a previously unreported assembly intermediate that we newly define as class C4 (**Fig. 4-5I**). Like the other class C structures, class C4 lacked the central protuberance, but possessed clearly resolved density for rRNA helix 68, which was only present in the mature E4 and E5 classes from Davis et al. [11]. Traditional homogeneous reconstruction of the particle images constituting this cluster reproduced a similar, albeit lower-resolution structure, which confirmed the existence of this structural state in the original dataset (**Supplementary Fig. 4-13**). We found that the cryoDRGN latent representation is highly reproducible across replicates (**Supplementary Fig. 4-14**). CryoDRGN experiments and runtimes are summarized in **Fig. 4-5J**. In addition to illustrating cryoDRGN's ability to model extremely heterogeneous datasets without user-driver classification, this analysis further demonstrated that cryoDRGN can identify novel



	|z|	Architecture	Image Size	Training Speed (minutes / 100k particles)	Total Runtime (50 epochs; 1 GPU)
Filtering	1	256 × 3	128	5.7	6.3 hr
Filtering	10	256 × 3	128	5.7	6.3 hr
Final	10	1024 × 3	256	43.9	31.9 hr

Figure 4-5: **A,B,** Latent space representation of particle images of the assembling LSU (EMPIAR-10076) [11] as a normalized histogram or UMAP embeddings after training a cryoDRGN 1D or 10D latent variable model, respectively. **C,D,** Latent space representation of particles colored by major LSU assembly state assigned from the 3D classification in Davis et al. [11]. Impurities in the dataset were assigned and subsequently filtered based on a cutoff of $z = -1$ in the 1D case (dotted line) and cluster assignment from a five-component Gaussian mixture model (GMM) in the 10D case. The dashed line in **D** indicates a rough outline of cluster assignment, shown in **Supplementary Fig. 4-11**. **E**, Density maps of the four major assembly states of the LSU reconstructed by cryoDRGN after training on the filtered dataset. Dashed lines indicate outlines of the fully mature 50S ribosome, with the central protuberance (CP) noted. **F,G,** Latent space representation of the filtered dataset, colored by major and minor assembly states assigned from the 3D classification in Davis et al. [11]. Points denote cluster centers for the corresponding assembly state. Major assembly state labels correspond to the structures from **E**. Inset shows a magnified view of the state C cluster and a population of particles originally misclassified into state E. **H,I,** CryoDRGN reconstruction of additional density maps, showing the 70S ribosome, an impurity during purification and LSU minor states C4 and E5. The newly identified C4 state resembles major state C in maturation but contains rRNA helix 68, previously present only in mature assembly states E4 and E5. **J,** Hyperparameters and runtime of the initial pilot experiments for particle filtering (**A**–**D**) and the final cryoDRGN model (**E**–**I**) trained on the assembling LSU dataset. Additional density maps are shown in **Supplementary Fig. 4-12**.

and rare (1 % of all particles) structural classes that would likely be overlooked by traditional hierarchical classification.

4.2.6 CryoDRGN reveals dynamic continuous motions in the pre-catalytic spliceosome

Finally, to assess cryoDRGN's ability to model large continuous conformational changes, we reanalyzed a dataset of the pre-catalytic spliceosome (EMPIAR-10180) [32]. Using extensive, expert-guided focused classifications, Plaschka et al. reconstructed a composite map for this complex and suggested that the complex sampled a continuum of conformations with large motions of the SF3b subcomplex [32]. In our analysis, we first trained a 10-D latent variable model on the downsampled images using image poses derived from a consensus reconstruction (**Methods, Section 4.4**). Multiple clusters were observed in the latent space encodings of the dataset's particle images (**Fig. 4-6A**). In sampling structures from the latent space, the generated density maps revealed expected spliceosome conformations from the largest cluster, poorly resolved structures likely due to imaging artifacts from the leftmost cluster, structures lacking density for the SF3b subcomplex from a third cluster, and extra density of the U2 core, which is thought to be highly dynamic [15], from the uppermost cluster (**Fig. 4-6B**). To focus our analysis on bona-fide pre-catalytic spliceosome particles, we leveraged the latent space representation to eliminate any particles that mapped to the undesired clusters from two replicate runs (**Methods, Section 4.4**).

With the filtered particle stack, we trained a 10-D model on higher resolution images and visualized the dataset's latent encodings in 2-D using PCA (**Fig. 4-6C**). The visualized data manifold was unfeatured, consistent with a molecule undergoing non-cooperative conformational changes. By generating structures along the first principal component of the latent space encodings, we reconstructed a trajectory of the SF3b and helicase subcomplexes in motion, smoothly transitioning from an elongated state to one compressed against the body of the spliceosome (**Fig. 4-6D**). This large scale-motion is consistent with motions derived from the first principal

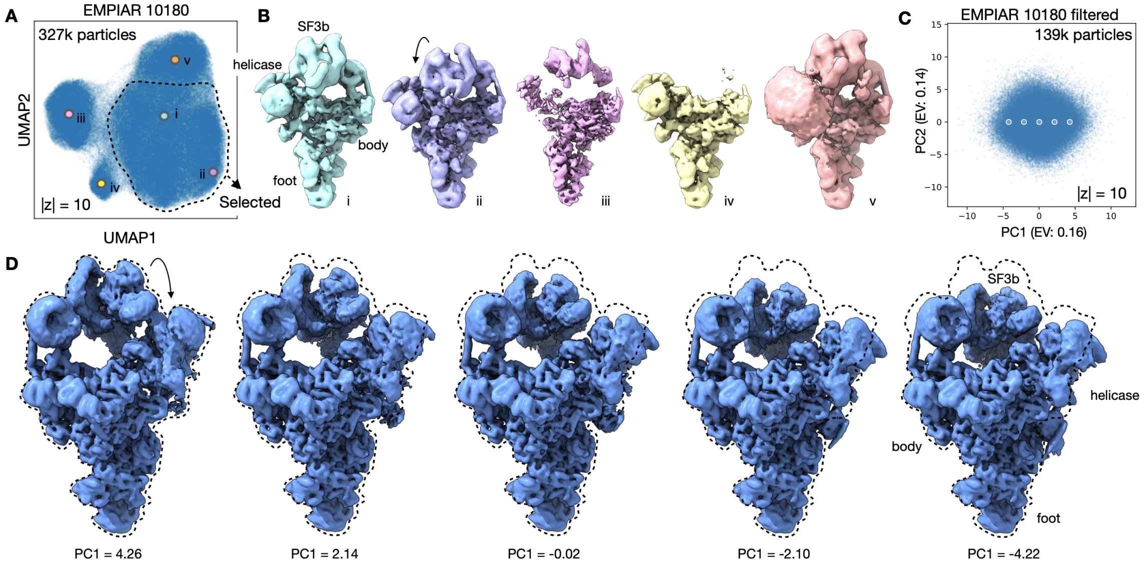


Figure 4-6: **A,** UMAP visualization of the latent space representation of particle images of the pre-catalytic spliceosome (EMPIAR-10180) [32] after training a 10D latent variable model with cryoDRGN. **B,** Representative structures generated at points shown in a that depict the expected structures of the pre-catalytic spliceosome (i,ii), structures likely corrupted by imaging artifacts (iii), the complex lacking the SF3b subcomplex (iv) and the complex with the U2 core (v). Density maps are shown at identical isosurface levels except for (v), which required a lower value to highlight the U2 core. **C,** PCA projection of latent space encodings after training a 10D latent variable model on the dataset filtered for the selected region in a. **D,** Structures generated by traversing along PC1 of the latent space representation at the points shown in **C**. Additional density maps are shown in **Supplementary Fig. 4-17**.

component of rigid body orientations from multi-body analysis (**Supplementary Fig. 4-15**) and in the first principal component of 3DVA's linear subspace model (**Supplementary Fig 4-16**). A similar traversal along the second PC produced a continuous trajectory of the SF3b and helicase subcomplexes moving in opposition (**Supplementary Fig. 4-17**). The anticorrelated motion of the SF3b and helicase subcomplexes in PC2, together with their correlated motion in PC1, suggests that the two domains move independently in the imaged ensemble. Finally, although trajectories along latent space PCs provide a summary of the extent of variability in the structure, cryoDRGN can also generate structures at arbitrary points from the latent space. By traversing along the nearest neighbor graph of the latent encodings and generating structures at the visited nodes, cryoDRGN generated a plausible trajectory of the conformations adopted by the pre-catalytic spliceosome (Shown in Supplemental Video 4 of Zhong et al [48]), highlighting the potential of single particle cryo-EM to uncover the conformational dynamics of molecular machines.

4.3 Discussion

This work introduces cryoDRGN, a method using neural networks to reconstruct 3D density maps from heterogeneous single particle cryo-EM datasets. The power of this approach lies in its ability to represent heterogeneous structures without simplifying assumptions on the type of heterogeneity. In principle, cryoDRGN is able to represent any distribution of structures that can be approximated by a deep neural network, a broad class of function approximators for continuous, nonlinear functions [16]. This flexibility contrasts with existing methods that impose limiting assumptions on the types of structural heterogeneity present in the sample. For example, 3D classification assumes a mixture of discrete structural classes; multibody refinement assumes conformational changes are composed of user-defined rigid-body motions; and 3DVA assumes that heterogeneity is generated from linear combinations of density maps. Although these approaches have proven useful, their model for heterogeneity is often mismatched with the true structural heterogeneity in many systems, and thus can introduce bias into reconstructions. In contrast, we empirically show that the cryoDRGN architecture can model both discrete compositional heterogeneity and continuous conformational changes without the aforementioned structural assumptions. For example, we discovered heterogeneous states of the RAG complex and Pf80S ribosome that were originally averaged out of the homogeneous reconstruction. When analyzing the assembling *E. Coli* LSU dataset, cryoDRGN learned an ensemble of LSU assembly states without *a priori* specification of the number of states or initial models as is required for 3D classification. Finally, when analyzing the pre-catalytic spliceosome, we found that the continuous conformational changes reconstructed by cryoDRGN lacked the rigid-body boundary artifacts from multibody refinement's mask-based approach [27] (**Supplementary Fig. 4-15**) or linear interpolation artifacts from 3DVA's linear subspace model [35] (**Supplementary Fig. 4-16**).

4.3.1 Interpretation of the latent space

A key feature of cryoDRGN is its ability to provide a low-dimensional representation of the dataset's heterogeneity given by each particle's latent encoding. Subject to optimization, cryoDRGN organizes the latent space such that structurally related particles are in close proximity. In simulated and real datasets, we find that continuous motions are embedded along a continuum in latent space (**Fig. 4-3E-G, 4-6C**) and that compositionally distinct states manifest as clusters (**Fig. 4-3H, 4-5F**). These empirical results demonstrate that visualization of the distribution of latent encodings can be informative in exploring the structural heterogeneity within the imaged ensemble, and even suggest a possible interpretation of the latent space as a pseudo-conformational landscape. However, we note that cryoDRGN's objective function aims only to reproduce the distribution of structures and does not guarantee that the latent space layout (or its 2D visualization) will produce interpretable features of the underlying energy landscape. Furthermore, structures reconstructed from unoccupied regions of the latent space will not in general correspond to true physical structures, as cryoDRGN optimizes the likelihood of the observed data and these structures are not observed. Finally, in real datasets, there may exist images that do not originate from the standard single particle image formation model, for example, false positives encountered during particle picking. We demonstrated the utility of the latent space representation in identifying such impurities, ice artifacts, and other out-of-distribution particle images that may be filtered out in subsequent analyses (**Fig. 4-5A-D, 4-6**). We emphasize that different datasets have diverse sources of heterogeneity, and thus the interpretation of the cryoDRGN latent space is highly dataset-dependent. We provide interactive analysis tools in the cryoDRGN software for exploring the learned latent space.

4.3.2 Visualizing structural trajectories

In addition to encoding particles in an unsupervised manner, cryoDRGN can reconstruct 3D density maps from user-defined positions in latent space. Because cryoDRGN learns a generative model for structure, an unlimited number of structures can be generated and analyzed, thus enabling visualization of structural trajectories. By leveraging the latent encodings of the particle images, users can directly traverse the data manifold and only sample structures from regions of latent space with significant particle occupancy. Indeed, we applied a well-established graph-traversal algorithm [1] to visualize data-supported motions in the RAG complex, the Pf80S ribosome, bL17-independent assembly of the bacterial ribosome, and the pre-catalytic spliceosome (Shown as Supplemental Videos in Zhong et al. [48]). We note that while this approach is useful in visualizing potential structural changes linking one state to another, they do not necessarily reveal the kinetically preferred path.

4.3.3 Practical considerations in choosing training hyperparameters

Although this method emphasizes an unsupervised approach to analyzing structural heterogeneity, cryoDRGN does require that the user define the dimensionality of the latent space and the architecture of both the encoder and decoder networks. We find that in practice, training a smaller architecture on downsampled images is effective at distinguishing bona-fide particles from contaminants and imaging artifacts (**Fig. 4-5A-D, Fig. 4-6A**), and we recommend users initially employ such pilot experiments to filter their dataset. Additionally, we find that in our tested datasets, a 10-D latent space provides sufficient representation capacity to effectively model structural heterogeneity, and that this 10-D space can be readily visualized with PCA or UMAP. Notably, we recommend the use of such a 10-D latent space instead of lower dimensional space as we have found that 10-D spaces result in much more rapid overall training, which is consistent with similar observations of related overparameterized neural network architectures [3]. Finally, users must specify the number of nodes and layers in the neural networks, hyperparameters that limit the complexity of the learned function. Here, we find an inverse relationship between neural network size and the achievable resolution of a given structure (**Fig. 4-2B**). Training larger networks on larger images is significantly slower (**Fig. 4-2E**), and we recommend that users perform an initial assessment using down-sampled images and relatively small networks before proceeding to high-resolution reconstructions. We note that use of excessively complex models (i.e. large architectures or latent variable dimensions) can lead to overfitting, which may be alleviated by standard neural network regularization techniques such as early stopping or using a simpler model [3]. We provide recommended training settings in the cryoDRGN software.

4.3.4 Discovering new states using cryoDRGN

CryoDRGN can be used to identify novel clusters of structurally-related particles, which can then be visualized by generating density maps from that region of latent space. Indeed, in analyzing the LSU assembly dataset, we noted a new structural state, C4, that was completely missed in traditional hierarchical classification. C4 provides structural evidence that a functionally critical intersubunit helix (h68) can dock in a native conformation in the absence of the central protuberance (**Fig. 4-5I**). Notably, we could validate the existence of this state by performing traditional homogeneous refinement using 1,000 particles from this cluster in the cryoDRGN latent space (**Supplementary Fig. 4-13**). Although we were able to readily identify this state from a distinct cluster present in the UMAP visualization (**Fig. 4-5G**), in general, the definition of distinct “states” may not be as readily apparent (e.g. the “missing SSU head” state in **Supplementary Fig. 4-9**), and we view the unsupervised identification of states from the cryoDRGN structural ensemble as an exciting area to pursue.

In future work, we envision using cryoDRGN to reveal the number of discrete classes, their constituent particles, and to produce initial 3D models that could be used as inputs for a traditional 3D reconstruction. Given the mature state of such tools [49, 34], this data-driven classification approach followed by traditional homogeneous reconstruction, particle polishing, and higher order image aberration correction, has the potential to produce high-resolution structures of the full spectrum of discrete structural states.

4.3.5 *De novo* pose estimation

As implemented, cryoDRGN uses pose estimates resulting from a traditional consensus 3D reconstruction. In analyzing four publicly available datasets, we found that such consensus pose estimates were sufficiently accurate to generate meaningful latent space encodings and to produce interpretable density maps of distinct structures. It is clear, however, that this approach will fail if the degree of structural heterogeneity in the dataset results in inaccurate pose estimates. For example, a mixture of structurally unrelated complexes will align poorly to a consensus structure, and thus produce poor pose estimates. Notably, our framework is differentiable with respect to pose variables, which, in principle, should allow for on-the-fly pose-refinement or *de novo* pose estimation. Future work will explore the efficacy of incorporating such features to enable fully unsupervised reconstruction of heterogeneous distributions of protein structure from cryo-EM images.

4.4 Methods

4.4.1 Datasets

Simulated compositionally heterogeneous dataset generation

To generate the compositionally heterogeneous dataset, the 30S, 50S and 70S subunits of the *E. coli* ribosome were extracted from PDB 4YBB in PyMOL [2]. A density map of each subunit was generated from the atomic model using the molmap [44] command in Chimera [31] at a grid spacing of 1.5 Å/pix and a resolution of 4.5 Å. The resulting volume was padded to a box size of $D = 256$, where D is the width in pixels along one dimension. Simulated particle images were generated with a custom Python script available in the cryoDRGN software by rotating the density map with a random rotation sampled uniformly from SO(3), projecting along the z-axis, and shifting the image with an in-plane translation sampled uniformly from $[-20, 20]^2$ pixels. Images were then downsampled to $D = 128$ by Fourier clipping using a custom Python script, corresponding to a Nyquist limit of 6 Å. Projection images were multiplied with the CTF in Fourier space, where the CTF was computed from defocus values randomly sampled from those given in EMPIAR-10028 [45], no astigmatism, an accelerating voltage of 300 kV, a spherical aberration of 2 mm, and an amplitude contrast ratio of 0.1. An envelope function with a B-factor of 100 Å^2 was applied. Noise was added with a signal to noise ratio (SNR) of 0.1 where the noise-free signal images were defined as the entire $D \times D$ image. After performing this procedure for each subunit, 10k, 15k, and 25k simulated particles of the 30S, 50S, and 70S ribosome, respectively, were combined.

Simulated conformationally heterogeneous dataset generation

To simulate continuous conformational heterogeneity, 50 density maps were generated along a one-dimensional reaction coordinate defined by rotating a dihedral angle in an atomic model of a hypothetical protein complex. Each model was generated at 0.03 radian increments of the bond rotation, leading to a total range of 1.5 radians.

Density maps were generated using the molmap [44] command in Chimera [31] at a grid spacing of 6 Å/pix and resolution of 12 Å, and padded to a box size of $D = 128$. For the *Uniform* dataset, 1000 projection images were generated for each density map at random orientations and in-plane translations sampled from $[-10, 10]^2$ pixels. For the nonuniform datasets, particles were generated along the reaction according to a 3-component Gaussian mixture model with means at the 10th, 25th, and 40th density map and standard deviation of 0.09 and 0.03 radians for the *Cooperative* and *Noncontiguous* datasets, respectively. Sampled reaction coordinate values were binned to convert into a particle distribution among the 50 generated density maps, and clipped at values of the reaction coordinate beyond the 50 maps. A total of 50k particles were generated for each dataset. CTF and noise at an SNR=0.1 were added to all datasets using the same procedure described above with CTF defocus values randomly sampled from EMPIAR-10028 [45].

Real cryo-EM datasets

Picked particles and the star file containing CTF parameters were downloaded from the Electron Microscopy Public Image Archive (EMPIAR) [17] for datasets EMPIAR-10049, EMPIAR-10028, EMPIAR-10076, and EMPIAR-10180. Particle images were downsized to the image size used in training by clipping in Fourier space with a custom Python script available in the cryoDRGN software.

4.4.2 Consensus reconstructions

Homogeneous 3D reconstruction of the Pf80S ribosome (EMPIAR-10028) was performed in cryoSPARC v2.4 [33] using the *ab initio* reconstruction job followed by the homogeneous refinement job with default parameters. The final reconstruction reported a GSFSC 0.143 [36] resolution of 3.1 Å with a tight mask and 4.1 Å unmasked.

Homogeneous 3D reconstruction of the bL17-depleted ribosome assembly intermediates (EMPIAR-10076) was performed as above, leading to a final structure with a GSFSC 0.143 resolution of 3.2 Å with a tight mask and 4.8 Å unmasked.

Homogeneous 3D reconstruction of the RAG complex (EMPIAR-10049) was performed as a "Homogeneous Refinement (NEW!)" job in cryoSPARC v2.15 with all default settings, including C1 symmetry. The asymmetric PC map of the RAG complex was used as an initial model (EMDB 6489), low pass filtered by 30 Å. The final structure had GSFSC 0.143 resolution of 3.6 Å with a tight mask and 4.6 Å unmasked.

Poses from a consensus reconstruction of the pre-catalytic spliceosome were obtained from the star file deposited in EMPIAR-10180.

4.4.3 CryoDRGN homogeneous reconstruction

CryoDRGN decoder networks with no input latent variable were trained for 50 epochs on full-resolution images of the RAG complex (D=192, 1.23 Å/pix) and the Pf80S ribosome (D=360, 1.34 Å/pix). The tested architectures were MLPs with ReLU activations, where the network size was either 3 hidden layers of width 128 (denoted 128 × 3), 256 × 3, 512 × 3, 1024 × 3, or 1024 × 10. Image poses were set to poses obtained from a consensus reconstruction in cryoSPARC [33], described above. Networks were trained on minibatches of 8 images using the Adam optimizer [18] with a learning rate of 0.0001. Once training completed, the decoder network was evaluated on the 3D coordinates of a $D \times D \times D$ voxel array spanning $[-0.5, 0.5]^3$, where D is the image size in pixels along one dimension. For visualization in **Figure 4-2**, the RAG complex density maps were sharpened by -54 Å^2 and -127.4 Å^2 for the cryoDRGN and cryoSPARC map, respectively, based on Guinier analysis [36] performed in a custom Python script; both the cryoSPARC and cryoDRGN density maps of the Pf80S ribosome were sharpened using the published B-factor of -80.1 Å^2.

4.4.4 Map-to-map FSC

Fourier shell correlation curves were computed between the cryoSPARC density maps and cryoDRGN density maps using a custom Python script available in the cryoDRGN software. Real space masks were defined by first thresholding the cryoDRGN volume at half of the 99.99th percentile density value. The mask was then dilated by 25 Å from the original boundary, and a soft cosine edge was used to taper the mask to 0 at 15 Å from the dilated boundary.

4.4.5 CryoDRGN heterogeneous reconstruction

Model training

A summary of the datasets, hyperparameters, and runtimes for all cryoDRGN heterogeneous reconstruction experiments is given in Supplementary Table 4.1. CryoDRGN encoder-decoder networks were trained from their randomly initialized values for each single particle cryo-EM dataset. Image poses used for training were either the ground truth poses for simulated datasets or poses obtained from a consensus reconstruction as described above. All networks were trained on minibatches of 8 images using the Adam optimizer with a learning rate of 0.0001. After training, the dataset images were evaluated through the encoder to obtain the latent encoding for each image. We define the latent encoding as the *maximum a posteriori* value of $q_\xi(z|x)$ predicted by the encoder.

Latent space visualization

For latent spaces with dimension greater than 2, the distribution of latent encodings were visualized with standard dimensionality reduction techniques such as PCA and UMAP35. PCA projections of latent space particle distributions were computed using the implementation provided by scikit-learn [29]. Two-dimensional UMAP [24] embeddings were computed using version 0.4.1 of the Python implementation (https://github.com/lmcinnes/umap) with the default settings of k=15 for the k-nearest neighbors graph and a minimum distance parameter of 0.1. Automated tools to analyze and visualize the latent space given the outputs of model training are provided in the cryoDRGN software.

Density map generation

Density maps were generated for a given value of the latent variable z by evaluating the trained decoder on z and the 3D coordinates of a $D \times D \times D$ voxel array spanning $[-0.5, 0.5]^3$. For higher dimensional latent spaces ($|z| > 1$), to generate representative samples from different regions of the latent space, we perform k-means clustering of the dataset's latent encodings to partition the latent space into k regions. A representative density map for each region is generated at the "on-data" cluster center; We define the latent encoding closest in Euclidean distance to the k-means cluster center as the "on-data" cluster center. Automated tools to generate k representative density maps following this procedure are provided in the cryoDRGN software.

4.4.6 Heterogeneous reconstruction of simulated datasets

For each simulated heterogeneous dataset, a 1-D latent variable model was trained for 100 epochs. The encoder architecture was 256 × 3 and the decoder architecture was 512 × 5. The image poses used for training were the ground truth image poses. After training on the Uniform simulated dataset, structures shown in **Figure 4-3C** were generated at the 5th, 23rd, 41st, 59th, 77th, and 95th percentile values of the latent encodings, and sharpened by a B-factor of -100 $Å^2$. After training on the *Compositional* simulated dataset, structures shown in **Figure 4-3D** were generated at the k-means cluster centers after performing k-means clustering with k=3 on the latent encodings and sharpened by a B-factor of -100 $Å^2$. Spearman correlation was computed using the implementation provided in the scipy version 1.5.2 Python package (https://www.scipy.org).

4.4.7 Per-image FSC

For simulated datasets where the ground-truth distribution of structures is known, "per-image FSC" curves can be computed between cryoDRGN-reconstructed density maps and the ground-truth density maps to quantitatively evaluate the reconstructed ensemble. To compute a per-image FSC, an FSC curve is computed between the density map generated by the cryoDRGN decoder at the value of the latent variable predicted for a given particle image and the ground-truth density map used to generate the image. 100 images randomly sampled according to the ground truth distribution of structures were used in the assessment of each of the simulated datasets. No real-space mask was used in computing the FSC.

4.4.8 Heterogeneous reconstruction of the RAG complex (EMPIAR-10049)

A 10-D latent variable model was trained on full-resolution particle images from EMPIAR-10049 (D=192, 1.23 Å/pix) and their consensus reconstruction poses for 25 epochs. The encoder and decoder architectures were 1024 × 3.

Density map generation: After training, k-means clustering with k=100 was performed on the predicted latent encodings for the dataset, and volumes were generated at the “on-data” cluster centers using the decoder network. Six structurally diverse representative structures were manually selected for visualization in **Figure 4-4A**.

Traditional heterogeneous refinement: To validate the heterogeneous RSS and NBD conformations observed in cryoDRGN, we use the 6 selected density maps, low pass filtered by 20 Å, as initial models to a heterogeneous refinement job in cryoSPARC v2.15.

4.4.9 Heterogeneous reconstruction of the 80S ribosome (EMPIAR-10028)

Pilot experiments: A 10-D latent variable model was trained on downsampled images (D=128, 3.78 Å/pix) from EMPIAR-10028 and their consensus reconstruction poses for 50 epochs. The encoder and decoder architectures were 256 × 3.

Particle filtering: After training, k-means clustering with k=20 was performed on the predicted latent encodings for the dataset. One cluster contained 860 particles that were outliers when viewing the projected encodings along the first and second principal component. This observation was reproducible, and the particles belonging to the outlier cluster from either of two replicates (960 particles in total) were removed from the dataset.

High-resolution training: After particle filtering, a 10-D latent variable model was trained on a random 90% the remaining 104,280 images (D=256, 1.88 Å/pix) for 25 epochs. The encoder and decoder architectures were 1024 × 3.

Density map generation: After training, k-means clustering with k=50 was performed on the predicted latent encodings for the dataset, and volumes were generated at the "on-data" cluster centers using the decoder network. A representative structure of the rotated state and the unrotated state were manually selected for visualization in **Figure 4-4B**. A representative structure of the missing head group state was manually selected for visualization in **Supplementary Figure 4-3**. The numbered k-means cluster centers shown in **Supplementary Figure 4-9A**, originally arbitrarily ordered, were reordered based on hierarchical clustering of the latent encodings with Euclidean distance metric and average linkage.

Validation with traditional reconstruction: To validate the 40S rotated state, we selected 4,889 particles as the cluster from k-means clustering with k=20 that was separated along PC1 (**Supplementary Fig. 4-10**). These particles were then input to a homogeneous refinement job in cryoSPARC v2.15. The cryoDRGN density map, low pass filtered by 30 Å, was used as the initial model.

4.4.10 Heterogeneous reconstruction of the assembling 50S ribosome (EMPIAR-10076)

Pilot experiments: A 1-D and a 10-D latent variable model were trained on downsampled images (D=128, 3.3 Å/pix) from EMPIAR-10076 with poses from a consensus reconstruction for 50 epochs. The encoder and decoder architectures were 256 × 3.

Particle filtering: From the 1-D experiment, particles with $z < -1$ were removed from subsequent analysis. From the 10-D experiment, a 5-component, full-covariance Gaussian mixture model (GMM) was fit to the latent encodings using scikit-learn [29], and particles from the outlier cluster were removed. The outlier cluster was identified by visualizing the magnitude of the latent encodings (**Supplementary Fig. 4-10**). The intersection of both filtered particles stacks was used for subsequent analysis. 2D classification of the kept and removed particles was performed in cryoSPARC v2.4 [33] using all default options except for the number of 2D classes, which was set to 20. *Ab initio* reconstruction of the kept and removed particles was performed in cryoSPARC v2.4 [33] using all default options.

High-resolution training: A 10-D latent variable model was trained on a random 90% of the remaining 97,031 images (D=256, 1.7 Å/pix) for 50 epochs. The encoder and decoder architectures were 1024 × 3. Two additional replicates were run, one with the exact settings from a different random initialization and a second with latent variable dimension $|z| = 8$.

Density map generation: After training, the dataset's latent encodings were viewed in 2D with UMAP [24]. Density maps corresponding to the major and minor assembly states were generated at the "on-data" mean latent encoding for each class, i.e. $\hat{z}_M = \frac{1}{|M|} \sum_{i \in M} z_i$, where M is the set of particles assigned to a given class in the published 3D classification.

Map-to-map FSC: The map-to-map FSC was computed between the cryoDRGN and published density map for each minor class. Density maps were aligned in Chimera and a loose real-space mask (obtained as described above) was applied before computing an FSC curve.

Reproducibility analysis: For each replicate, a 5-component, full-covariance GMM was fit to the UMAP embeddings using scikit-learn [29]. UMAP axes were negated to facilitate visual comparison. Label assignments were permuted to ensure consistent assignments between replicates. Clustering consistency was computed as the percentage of particles with identical GMM labels.

New assembly state C4: Particles corresponding to the new assembly state were manually selected from the UMAP embeddings with an interactive lasso tool in a custom visualization script available in the cryoDRGN software, whose outline is shown in the **Figure 4-5F, inset**. The mean latent encoding of the resulting 1,113 selected particles was used to generate the structure representative for this new assembly state.

Validation of C4 with traditional refinement: The particles associated with class C4 were then input to a homogeneous refinement job in cryoSPARC v2.15. The cryoDRGN density map, low pass filtered to 30 Å, was used as the initial model.

4.4.11 Heterogeneous reconstruction of the pre-catalytic spliceosome (EMPIAR-10180)

Pilot experiments: A 10-D latent variable model was trained on downsampled images (D=128, 4.25 Å/pix) from EMPIAR-10180 for 50 epochs. The encoder and decoder architectures were 256 × 3. Poses were obtained from the consensus reconstruction values given in the consensus_data.star deposited to EMPIAR-10180.

Particle filtering: The UMAP embeddings showed multiple clusters where the largest cluster corresponded to fully formed pre-catalytic spliceosomes. Particles corresponding to other clusters were removed from subsequent analyses by first performing k-means clustering with k=20 on the latent encodings, and removing k-means clusters whose structure did not resemble the fully formed pre-catalytic spliceosome (11 out of 20 k-means clusters in one replicate, and 10 out of 20 in a second replicate).

High-resolution training: A 10-D latent variable model was trained on a random 90% of the remaining 155,247 images (D=256, 2.1 Å/pix) for 50 epochs. The encoder and decoder architectures were 1024 × 3.

Density map generation: After training, the dataset's latent encoding was viewed in 2-D with UMAP and PCA. Density maps in **Figure 4-6D** were generated at the latent encoding values along the PC1 axis at five equally spaced points between the 5th and 95th percentile of PC1 values. Density maps in **Supplementary Figure 4-17** were generated at the latent encoding values along the PC2 axis at five equally spaced points between the 5th and 95th percentile of PC2 values. Density map generation along PC axes is implemented in a custom script in the cryoDRGN software.

4.4.12 Latent space graph traversal for generating trajectories

Trajectories were generated by first creating a nearest-neighbors graph from the latent encodings of the images, where a neighbor was defined if the Euclidean distance was below a threshold computed from the statistics of all pairwise distances. We choose a value for each dataset such that the average number of neighbors across all nodes is 5. Edges were then pruned such that a given node does not have more than 10 neighbors. Then, Djikstra's algorithm [1] was used to find the shortest path along the graph connecting a series of anchor points, and density maps were generated at the z value of the visited nodes. Anchor points were either defined manually or set to be the "on-data" cluster centers after performing k-means clustering of the latent encodings. CryoDRGN's graph traversal algorithm is provided in the cryoDRGN software.

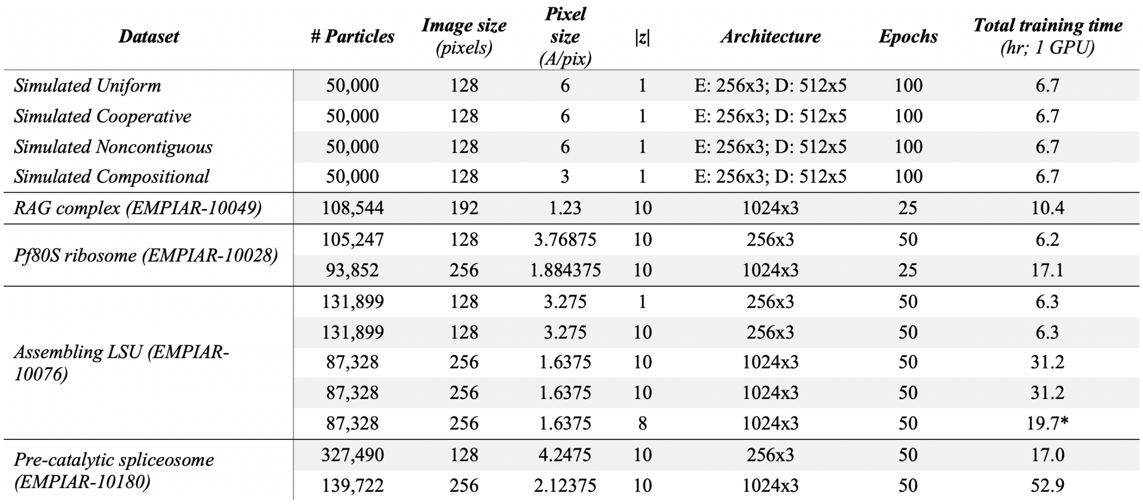
4.4.13 Density map visualization

All density map figures and trajectories were prepared with ChimeraX [13] and are viewed at identical isosurface levels for a given model unless otherwise specified.

4.4.14 3DVA

3D Variability Analysis [35] was performed in cryoSPARC v2.15 on the 139,722 particles and their consensus poses comprising the filtered EMPIAR-10180 dataset used in cryoDRGN analysis. Three variability modes were solved with all default options and the low-pass filter resolution set to 7 Å. 3DVA per-particle latent encodings were extracted from the cryoSPARC metadata file. Spearman correlation was computed using the implementation provided in the scipy version 1.5.2 Python package (https://www.scipy.org). To visualize the component 1 trajectory in **Supplementary Figure 4-16D**, the consensus density map was combined with the component 1 eigen-volume at 5 equally spaced points between the 1st and 99th percentile value of the 3DVA component 1 latent encoding distribution.

4.5 Supplementary Tables



Dataset	*# Particles*	*Image size (pixels)*	*Pixel size (A/pix)*	*|z|*	*Architecture*	*Epochs*	*Total training time (hr; 1 GPU)*
Simulated Uniform	50,000	128	6	1	E: 256x3; D: 512x5	100	6.7
Simulated Cooperative	50,000	128	6	1	E: 256x3; D: 512x5	100	6.7
Simulated Noncontiguous	50,000	128	6	1	E: 256x3; D: 512x5	100	6.7
Simulated Compositional	50,000	128	3	1	E: 256x3; D: 512x5	100	6.7
RAG complex (EMPIAR-10049)	108,544	192	1.23	10	1024x3	25	10.4
Pf80S ribosome (EMPIAR-10028)	105,247	128	3.76875	10	256x3	50	6.2
	93,852	256	1.884375	10	1024x3	25	17.1
Assembling LSU (EMPIAR-10076)	131,899	128	3.275	1	256x3	50	6.3
	131,899	128	3.275	10	256x3	50	6.3
	87,328	256	1.6375	10	1024x3	50	31.2
	87,328	256	1.6375	10	1024x3	50	31.2
	87,328	256	1.6375	8	1024x3	50	19.7*
Pre-catalytic spliceosome (EMPIAR-10180)	327,490	128	4.2475	10	256x3	50	17.0
	139,722	256	2.12375	10	1024x3	50	52.9

Table 4.1: Summary of dataset statistics, training hyperparameters, and runtimes for cryoDRGN heterogeneous reconstruction experiments. The neural network architecture is denoted as $d \times l$, where d indicates the number of nodes per layer and l is the number of hidden layers. The architecture corresponds to both the encoder (E) and decoder (D) MLPs unless otherwise specified. Total training times were recorded from training on a single Nvidia Tesla V100 32GB memory GPU card on either an Intel Xeon Gold 6130 CPU (2.10GHz, 791GB of RAM) or an IBM Power9 node with 1.2 TB of RAM. The reported training times may be overestimated since the presence of any concurrently running programs was not controlled for. (*) The training time of the third replicate of EMPIAR-10076 is substantially faster as using $|z| = 8$ better satisfies tensor shape constraints for Nvidia Tensor Core hardware acceleration.

4.6 Supplementary Figures

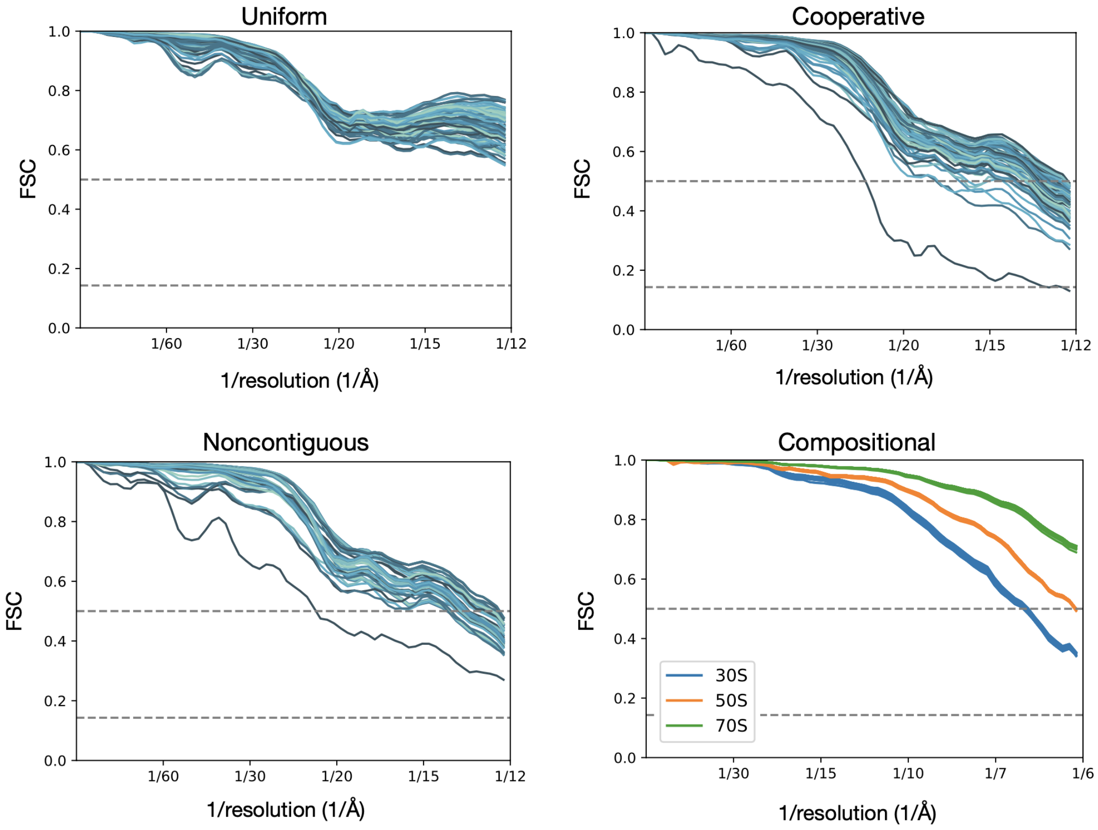


Figure 4-7: Per-image FSC curves between ground-truth maps and density maps from cryoDRGN trained on simulated heterogeneous datasets. For each dataset, we compute 100 "per-image FSC" curves between generated and ground-truth density maps (Section 4.4.7). Images are sampled at equally spaced percentiles along the reaction coordinate for the *Uniform*, *Cooperative*, and *Noncontiguous* datasets. For the *Compositional* dataset, the per-image FSC for 20, 30, and 50 randomly sampled images of the 30S, 50S, and 70S ribosome, respectively, are shown. No mask is used in computing the FSC.

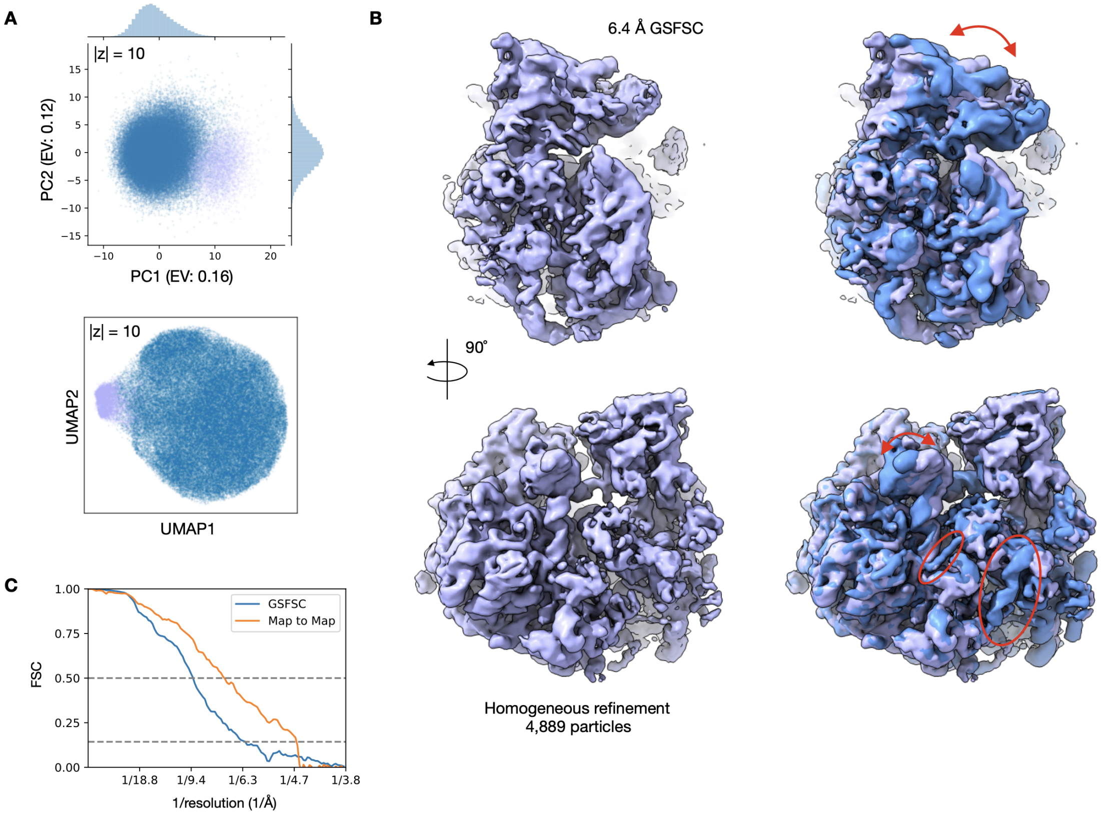


Figure 4-8: RAG complex density maps reconstructed by cryoDRGN and by heterogeneous refinement in cryoSPARC [33]. **A)** Front (top) and back (bottom) view of the six cryoDRGN density maps of the RAG complex from **Figure 4-4B**. **B)** Density maps from 3D classification in cryoSPARC using the cryoDRGN density maps in (**A**) as initial models. Gold-standard FSC resolution and number of particles used in reconstruction are noted. **C)** Two side views of the density maps from 3D classification in (**B**), focusing on the RSS and NBD.

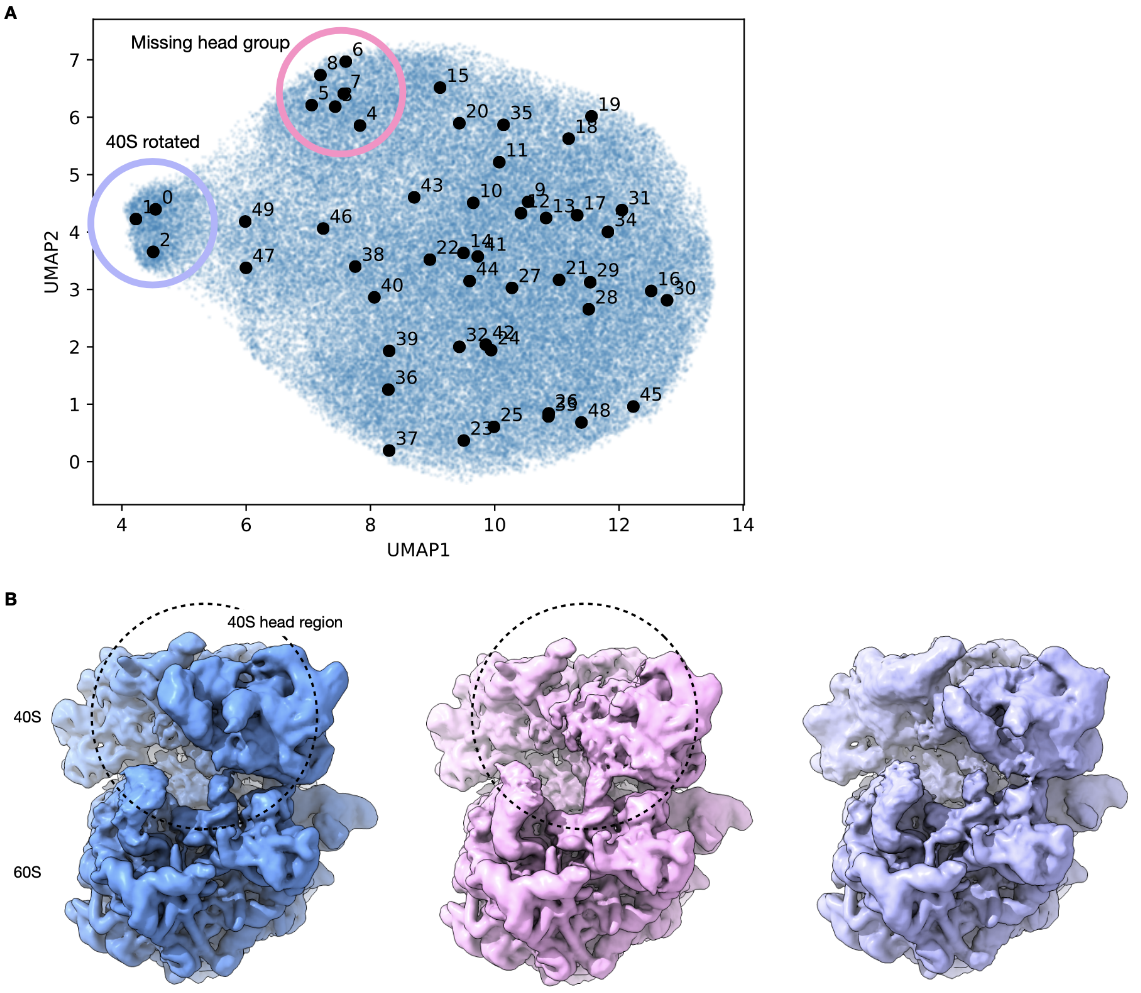


Figure 4-9: Structural heterogeneity in the small subunit of the Pf80S ribosome. **A)** UMAP visualization of latent space encodings of EMPIAR-10028 particles with 50 sampled points shown in black. Sampled points are ordered according to distances in latent space (Section 4.4.9). Visual inspection of the 50 volumes generated at the depicted points reveals 3 volumes with the 40S in a rotated state (purple) and 6 volumes with portions of the 40S head region missing (pink). **B)** Density map of the 80S ribosome with the missing head group reconstructed by cryoDRGN (pink) compared with the density maps from **Figure 4-4C** showing the canonical (blue) and 40S-rotated (purple) forms of the 80S ribosome. The density maps are generated from points 32, 4, and 1 in panel A from left to right.

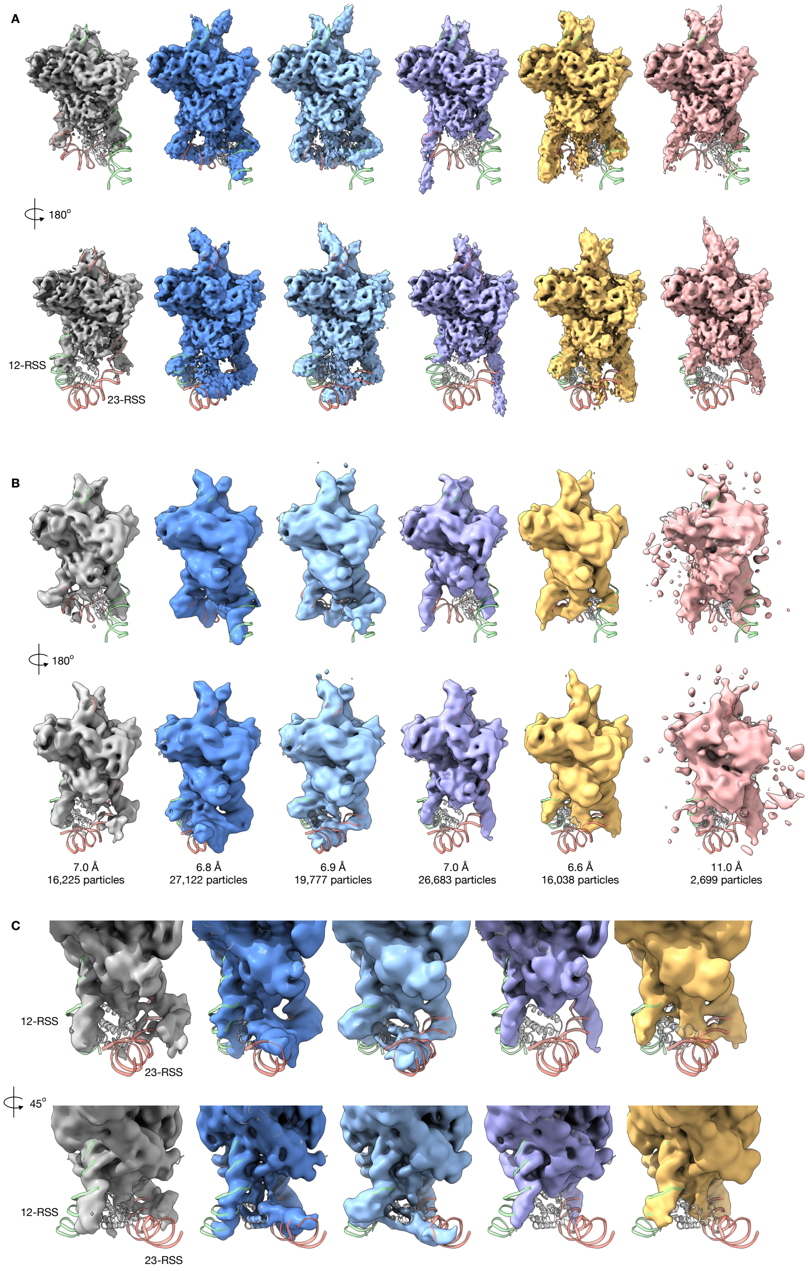


Figure 4-10: Validation of Pf80S rotated state with cryoSPARC. **A)** PCA and UMAP visualization of the cryoDRGN latent space representation of Pf80S particle images with 4,889 particles separated along PC1, selected with k-means clustering, colored in purple (Section 4.4.9). **B)** Density map from cryoSPARC homogeneous refinement (purple) using the 4,889 particles selected in (**A**). The density map is also shown superimposed with the cryoDRGN unrotated state (blue) and annotated as in **Figure 4-4C**. **C)** Gold standard FSC (GSFSC) curve between independent half-maps of the cryoSPARC refinement of the Pf80S rotated state and map-to-map FSC between the cryoDRGN and cryoSPARC density map of the Pf80S rotated state. Dotted lines indicate 0.5 and 0.143 cutoffs.

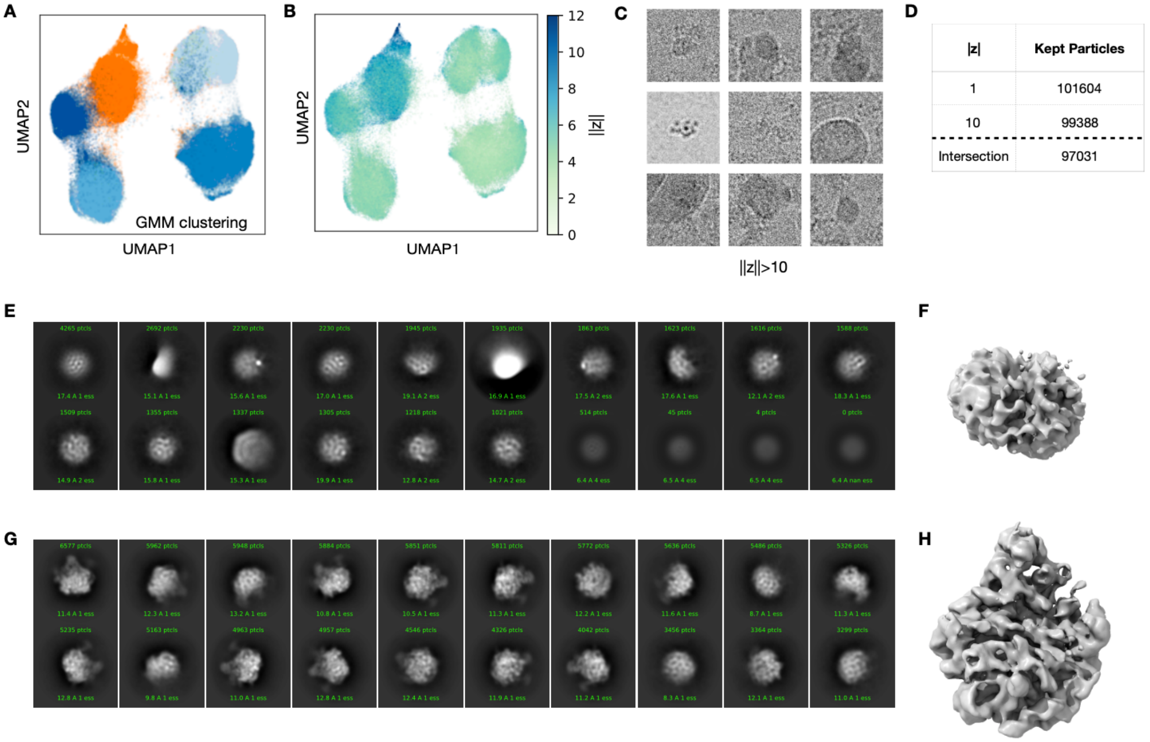


Figure 4-11: Filtering of particles from the assembling ribosome dataset. **A)** UMAP visualization of the 10-D latent encodings from cryoDRGN as in **Figure 4-5B**, colored by cluster after fitting a 5-component Gaussian mixture model. The cluster that was removed from subsequent analysis is colored orange. **B)** UMAP visualization of (**A**), colored by the magnitude of the latent encodings, $\|z\|$. **C)** Nine randomly sampled particle images from EMPIAR-10076 with $\|z\| > 10$ as predicted from cryoDRGN training in (**A,B**). Each image is 419.2 Åalong each side. **D)** Table summarizing dataset filtering. **E,F)** 2D classification and *ab initio* reconstruction of the 34,868 removed particles. **G,H)** 2D classification and *ab initio* reconstruction of the 97,031 kept particles.

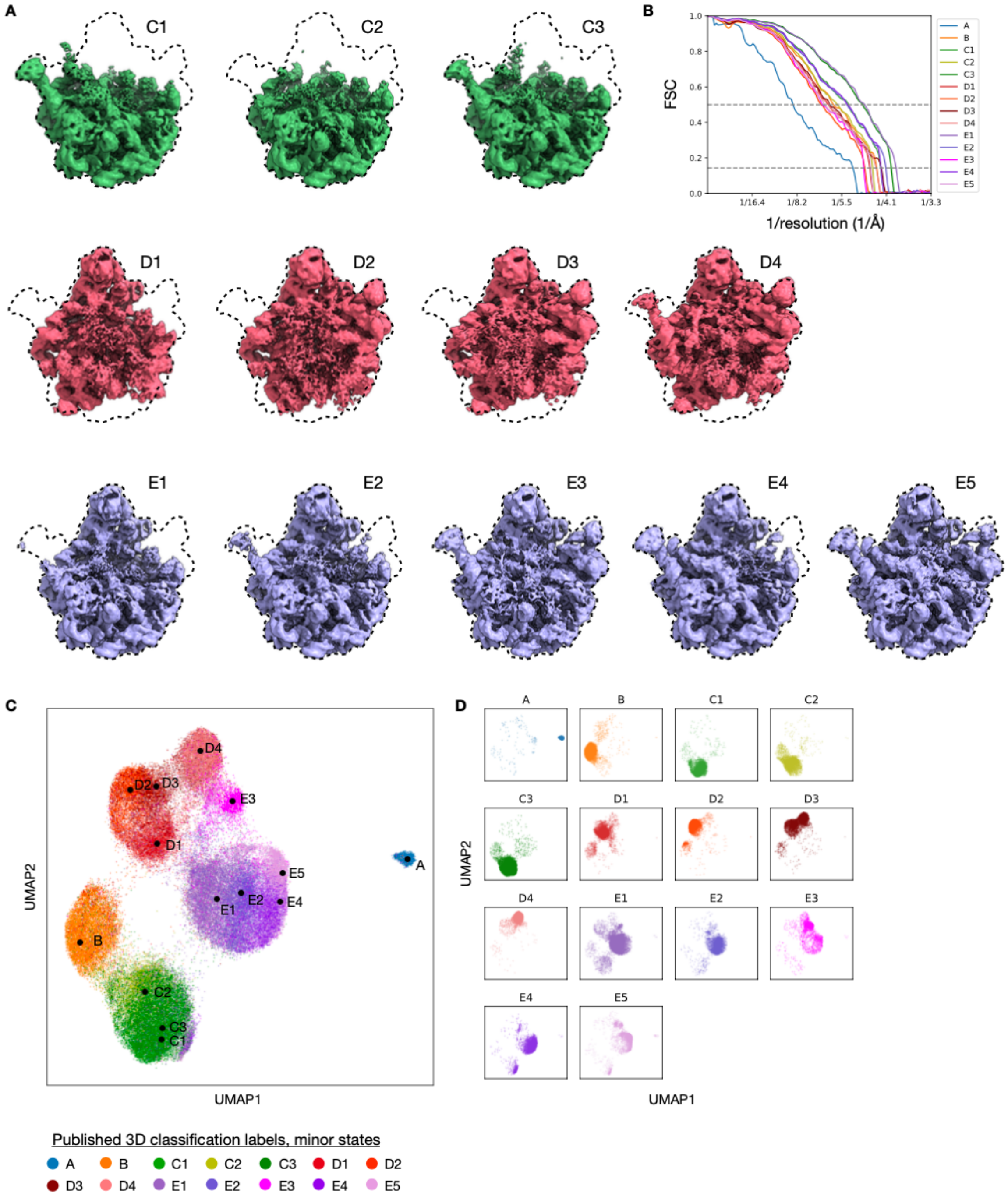


Figure 4-12: Minor LSU assembly states reconstructed by cryoDRGN. **A)** Density maps of the LSU minor assembly states reconstructed by cryoDRGN. Each cryoDRGN structure is generated at mean of the latent encoding of particles with the corresponding class assignment from Davis et al. [11]. **B)** Map-to-map FSC curves between the generated cryoDRGN density maps and the published density map from Davis et al. [11]. Published resolutions for assembly states B-E ranged between 4-5 Å. Dotted lines indicate 0.5 and 0.143 cutoffs. **C,D)** Reproduction of the cryoDRGN latent space shown in **Figure 4-5G**, colored by minor assembly state (**C**), or viewed in separate panels (**D**).

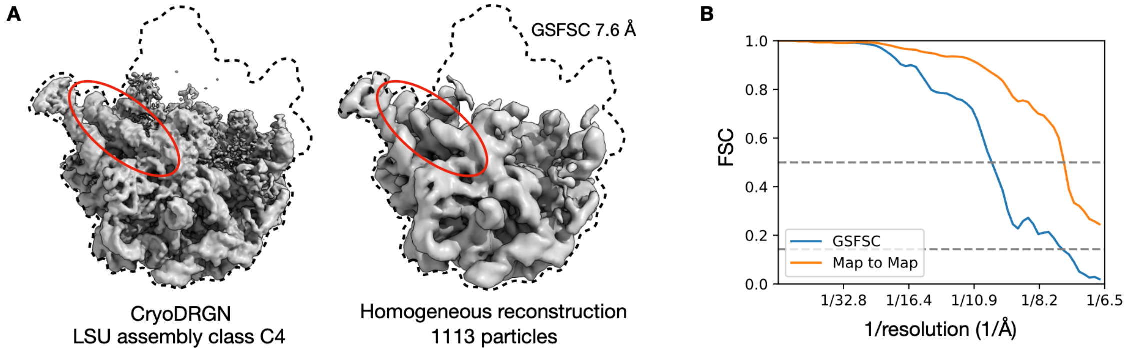
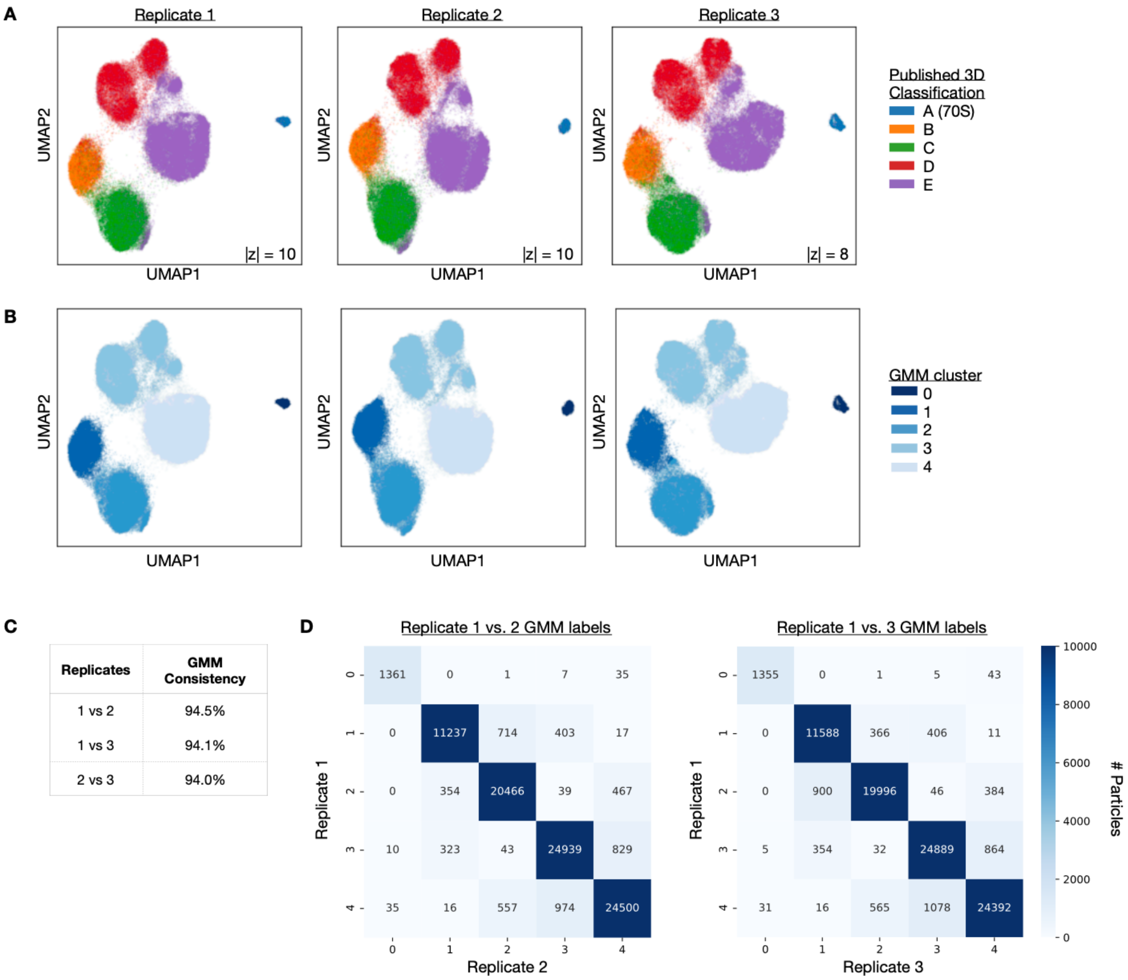


Figure 4-13: Validation of LSU class C4 with cryoSPARC. **A)** Density map from cryoSPARC homogeneous refinement of the 1,113 particles selected from the cryoDRGN latent representation that constitute class C4 (right), compared with the density map generated by cryoDRGN (left) from **Figure 4-5I**. rRNA helix 68 is circled in red. **B)** Gold standard FSC (GSFSC) curve between independent half-maps of the cryoSPARC reconstruction and map-to-map FSC between the cryoDRGN and cryoSPARC maps shown in (**A**). Dotted lines indicate 0.5 and 0.143 cutoffs.



Replicates	GMM Consistency
1 vs 2	94.5%
1 vs 3	94.1%
2 vs 3	94.0%

Figure 4-14: Reproducibility of cryoDRGN's latent space representation of the assembling ribosome. **A)** UMAP visualization of the latent encodings from replicate runs of cryoDRGN trained on the filtered particles of EMPIAR-10076. Particle embeddings are colored by major assembly state assigned from 3D classification in Davis et al. [11]. **B)** UMAP visualization of (**A**), colored by cluster after fitting a 5-component Gaussian mixture model on the UMAP embeddings. **C, D)** Consistency of the GMM labeling between replicates reported as the percentage of particles with identical labels (**C**) and the confusion matrix of GMM cluster assignments (**D**).

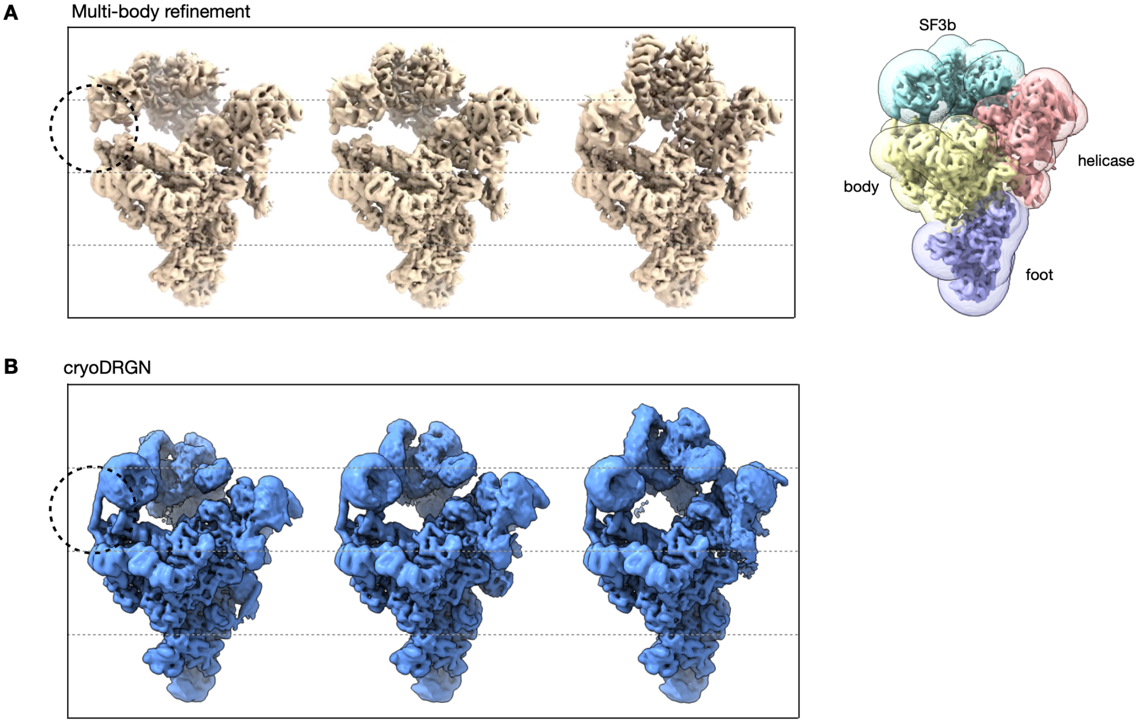


Figure 4-15: Comparison of multi-body refinement [26] and cryoDRGN of the pre-catalytic spliceosome. **A)** Visualization of a rigid-body trajectory from multibody refinement of the pre-catalytic spliceosome. Snapshots are extracted from the trajectory along PC1 of rigid-body orientations, showing a large-scale motion of the SF3b subcomplex. The masks that define the rigid-body decomposition of the complex are shown on the right. The circle highlights a helix that breaks at the boundary between bodies where the rigid-body assumption no longer holds. Adapted from Video 3 of Nakane et al. [27] and density maps and masks deposited in EMPIAR-10180. **B)** Alternate view of cryoDRGN's PC1 traversal in **Figure 4-6**. CryoDRGN learns the same overall motion of the SF3b subcomplex, however its neural network representation lacks the helix-breaking artifact.

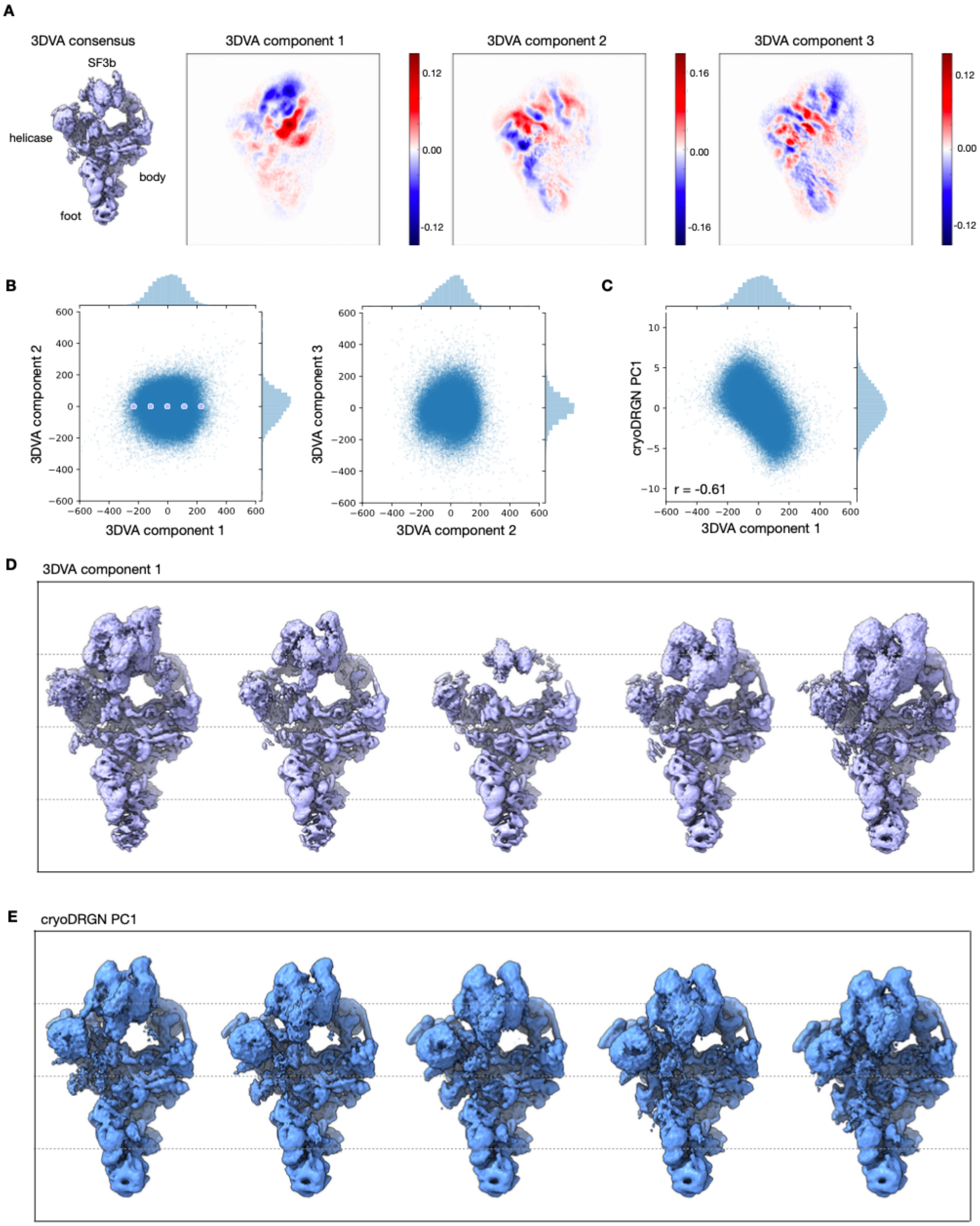


Figure 4-16: Comparison of cryoSPARC's 3D variability analysis (3DVA) [35] and cryoDRGN. **A)** Density map of the consensus reconstruction and 2D projections of the top three 3DVA variability components (i.e. eigen-volumes) that form a linear basis describing structural heterogeneity of the pre-catalytic spliceosome. **B)** 3DVA latent encodings of particles from the filtered EMPIAR-10180 dataset. **C)** Comparison of 3DVA component 1 latent encodings and PC1 of the cryoDRGN 10-D latent encodings from **Figure 4-6C**. Correlation indicates Spearman correlation. **D)** 3DVA component 1 trajectory at the depicted points in (**B**). **E)** Alternate view of the density maps from the cryoDRGN PC1 trajectory in **Figure 4-6D**.

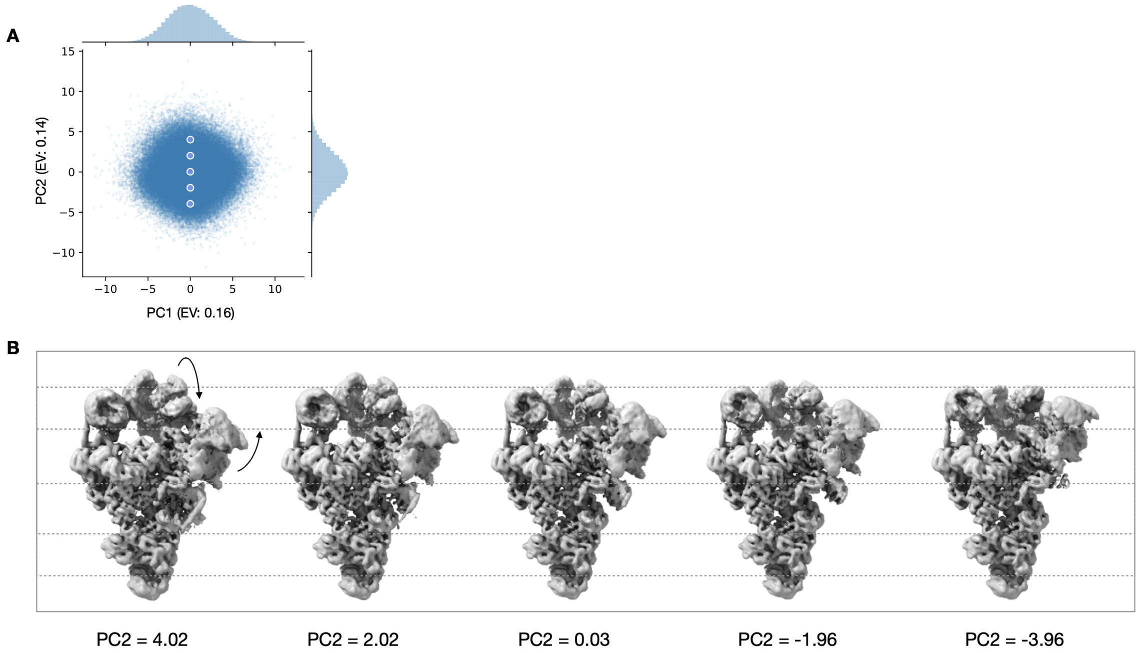


Figure 4-17: Additional structures of the pre-catalytic spliceosome reconstructed by cryoDRGN. **A)** PCA projection of the 10-D latent encodings from cryoDRGN as in **Figure 4-6C** with 5 points along PC2. **B)** Structures generated by traversing along PC2 of the latent space representation at points shown in (**A**).

Bibliography

[1] Cormen, t. h., leiserson, c. e., rivest, r. l. & stein, c. introduction to algorithms. 595–601 (MIT press and McGraw-Hill, 2009).

[2] The PyMOL molecular graphics system, version 2.3 (schrodinger, 2019).

[3] Zhang, c., bengio, s., hardt, m., recht, b. & vinyals, o. understanding deep learning requires rethinking generalization. in international conference on learning representations (ICLR, 2017).

[4] T Ahmed, Z Yin, and S Bhushan. Cryo-EM structure of the large subunit of the spinach chloroplast ribosome. *Sci. Rep.*, 6, 2016.

[5] B E Bammes, R H Rochat, J Jakana, D H Chen, and W Chiu. Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 nyquist frequency. *J. Struct. Biol.*, 177, 2012.

[6] T Bepler. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods*, 16, 2019.

[7] R N Bracewell. Strip integration in radio astronomy. *Aust. J. Phys.*, 9, 1956.

[8] Marcus A Brubaker, Ali Punjani, and David J Fleet. Building proteins in a day: Efficient 3D molecular reconstruction. April 2015.

[9] Y Cheng. Single-particle cryo-EM—how did it get here and where will it go. *Science*, 361, 2018.

[10] A Dashti. Trajectories of the ribosome as a brownian nanomachine. *Proc. Natl. Acad. Sci. U. S. A.*, 111, 2014.

[11] 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.

[12] J Frank and A Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods*, 100, 2016.

[13] T D Goddard. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.*, 27, 2018.

[14] T Grant, A Rohou, and N Grigorieff. cisTEM, user-friendly software for single-particle image processing. *Elife*, 7, 2018.

[15] David Haselbach, Ilya Komarov, Dmitry E Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and conformational dynamics of the human spliceosomal bact complex. *Cell*, 172(3):454–464.e11, January 2018.

[16] K Hornik, M B Stinchcombe, and H White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2, 1989.

[17] A Iudin, P K Korir, J Salavert-Torres, G J Kleywegt, and A Patwardhan. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, 13, 2016.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations (ICLR)*, 2013.

[20] Roy R Lederman and Amit Singer. Continuously heterogeneous hyper-objects in cryo-EM and 3-D movies of many temporal dimensions. *arXiv.org*, April 2017.

[21] X Li. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods*, 10, 2013.

[22] W Liu and J Frank. Estimation of variance distribution in three-dimensional reconstruction. i. theory. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 12(12):2615–2627, December 1995.

[23] D Lyumkis, A F Brilot, D L Theobald, and N Grigorieff. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.*, 183, 2013.

[24] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. February 2018.

[25] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *arXiv.org*, July 2019.

[26] T Nakane, D Kimanius, E Lindahl, and S H Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *Elife*, 7, 2018.

[27] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[28] E Nogales. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods*, 13, 2015.

[29] F Pedregosa. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, 12, 2011.

[30] P A Penczek, M Kimmel, and C M T Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure*, 19, 2011.

[31] E F Pettersen. UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25, 2004.

[32] C Plaschka, P C Lin, and K Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546, 2017.

[33] A Punjani, J L Rubinstein, D J Fleet, and M Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods*, 14, 2017.

[34] A Punjani, H Zhang, and D J Fleet. Non-uniform refinement: adaptive regularization improves single particle cryo-EM reconstruction. *Nat. Methods*, 17, 2020.

[35] Ali Punjani and David J Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.*, 213(2):107702, June 2021.

[36] P B Rosenthal and R Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.*, 333, 2003.

[37] H Ru. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell*, 163, 2015.

[38] Heng Ru, Pengfei Zhang, and Hao Wu. Structural gymnastics of RAG-mediated DNA cleavage in V(D)J recombination. *Curr. Opin. Struct. Biol.*, 53:178–186, December 2018.

[39] S H W Scheres. RELION: implementation of a bayesian approach to cryo-EM structure determination. *J. Struct. Biol.*, 180, 2012.

[40] F J Sigworth. Principles of cryo-EM single-particle image processing. *Microscopy*, 65, 2016.

[41] C Suloway. Automated molecular microscopy: the new leginon system. *J. Struct. Biol.*, 151, 2005.

[42] Ming Sun, Wen Li, Karin Blomqvist, Sanchaita Das, Yaser Hashem, Jeffrey D Dvorin, and Joachim Frank. Dynamical features of the plasmodium falciparum ribosome during translation. *Nucleic Acids Res.*, 43(21):10515–10524, December 2015.

[43] H D Tagare, A Kucukelbir, F J Sigworth, H Wang, and M Rao. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *J. Struct. Biol.*, 191, 2015.

[44] G Tang. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.*, 157, 2007.

[45] W Wong. Cryo-EM structure of the plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife*, 3, 2014.

[46] D Wrapp. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367, 2020.

[47] K Zhang. Gctf: real-time CTF determination and correction. *J. Struct. Biol.*, 193, 2016.

[48] 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, 2021.

[49] J Zivanov, T Nakane, and S H W Scheres. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ*, 7, 2020.

Chapter 5

CryoDRGN2: *Ab initio* neural reconstruction

Protein structure determination from cryo-EM data requires reconstructing a 3D volume (or distribution of volumes) from many noisy and randomly oriented 2D projection images. While the standard homogeneous cryo-EM reconstruction task aims to recover a single static structure, recently-proposed neural and non-neural cryo-EM reconstruction methods can reconstruct distributions of structures, thereby enabling the study of protein complexes that possess intrinsic structural (compositional or conformational) heterogeneity. These *heterogeneous reconstruction* methods, however, require fixed image poses, which are typically estimated from an upstream homogeneous reconstruction and are not guaranteed to be accurate under highly heterogeneous conditions.

In this chapter, we describe cryoDRGN2, an extension of cryoDRGN for *ab initio* heterogeneous reconstruction. CryoDRGN2 jointly estimates image poses and learns a neural model of a distribution of 3D structures without any prior pose or volumetric information. To achieve this, we adapt search algorithms from the traditional cryo-EM literature, and describe the optimizations and design choices required to make such a search procedure computationally tractable in the neural model setting. We show that cryoDRGN2 is robust to the high noise levels of real cryo-EM images, trains faster than earlier neural methods, and achieves state-of-the-art performance on real heterogeneous cryo-EM datasets.

This chapter presents work described in [57] performed jointly with Adam Lerer, Joey Davis, and Bonnie Berger and presented at the 2021 International Conference on Computer Vision.

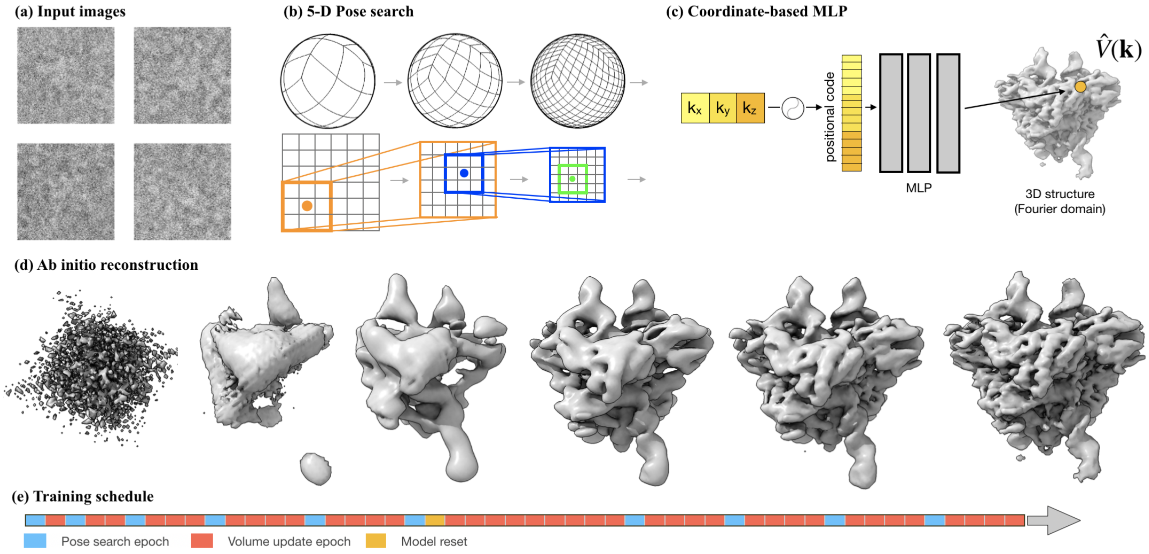


Figure 5-1: Overview of the cryoDRGN2 approach for *ab initio* reconstruction. (a) Example cryo-EM images of the RAG1-RAG2 complex [EMPIAR-10049]. (b) A multi-resolution 5-D pose search procedure doubles the resolution of the search grid at each iteration. (c) Coordinate-based MLP representation for volume (d) Volumes during *ab initio* training on the RAG dataset (e) A hypothetical training schedule interleaves pose search (expensive) and volume update (cheap) epochs. The model may also be reset after initial iterations to avoid vanishing gradients in training the neural volume.

5.1 Introduction

The last decade has seen explosive growth in the development and application of single particle cryo-electron microscopy (cryo-EM) for 3D structure determination of proteins and other biomolecules. Driven by parallel developments in improved hardware and image processing algorithms, many challenging structures not amenable to crystallographic approaches have now been solved at atomic or near-atomic resolution with cryo-EM [19, 29, 36].

Central to structure determination with cryo-EM is the computational reconstruction of the target molecule's 3D electron scattering potential (i.e. volume) from an experimentally-derived dataset of microscopy images. In a cryo-EM experiment, a purified solution of the molecule of interest is frozen in a thin layer of vitreous ice and imaged at cryogenic temperatures using a transmission electron microscope. After initial pre-processing of the raw micrographs, the resulting imaging dataset contains thousands to millions of noisy and randomly oriented 2D projection images (Fig. 5-1a). The goal of the cryo-EM reconstruction task is to infer the underlying 3D structure or structures present in the recorded images.

The 3D reconstruction task is a challenging inverse problem primarily due to the unknown image poses and the high amount of noise in the images. It is further complicated by the potential for each molecule to adopt variable conformations. A major opportunity thus exists to use cryo-EM to visualize and study complex distributions of dynamic protein structures, and numerous algorithms have been proposed to extract multiple structures from the imaging dataset, termed *heterogeneous reconstruction* [44].

Recent neural methods have shown promise in instantiating expressive continuous latent variable models for structural variability in cryo-EM data. In particular, cryoDRGN performs heterogeneous reconstruction by learning a deep generative model for 3D cryo-EM volumes. The first instantiation of cryoDRGN performed joint optimization of pose and heterogeneity with a branch and bound (BNB) algorithm for pose search (referred to as cryoDRGN-BNB here) [56]; however, this version scaled poorly and could not produce high-quality reconstructions of real cryo-EM datasets. In follow-up work, high-quality reconstructions on real data were achieved by modifying cryoDRGN to take previously estimated poses from a homogeneous reconstruction as input, hence eliding the difficult pose search procedure [55]. By estimating the intrinsic structural heterogeneity separately from the extrinsic pose variables, these methods are limited to mildly heterogeneous conditions where pose inference remains accurate.

In this chapter, we revisit the problem of joint optimization of image pose and volumes in cryoDRGN. In particular, we consider 5-D camera pose optimization in the context of a feed-forward MLP representation of volume, and propose search techniques to address the high render time of MLPs relative to voxel-based representations. We further identify a pathological case of vanishing gradients during training that we hypothesize originates from distributional shifts of the objective function during joint optimization. With these techniques, we improve upon both the speed and accuracy of cryoDRGN-BNB and demonstrate for the first time that neural models can achieve state-of-the-art accuracy for fully unsupervised *ab initio* reconstruction on both homogeneous and heterogeneous real cryo-EM datasets.

5.2 Background and Related Work

The standard cryo-EM reconstruction task involves reconstructing a single volume $V : \mathbb{R}^3 \to \mathbb{R}$ from many noisy and randomly oriented 2D projection images of V. As cryo-EM images are orthographic integral projections of the volume, 2D images can be related to the 3D volume by the Fourier slice theorem [3], which states that the Fourier transform of a 2D projection is a central slice from the 3D Fourier transform of the volume. The generative process for image $\hat{X}$ in the Fourier domain is thus written:

$$\hat{X}(k_x, k_y) = \hat{g} S(t) \hat{V}(R^T (k_x, k_y, 0)^T) + \epsilon \tag{5.1}$$

where $\hat{V} : \mathbb{R}^3 \to \mathbb{R}$ is the electron scattering potential (*volume*), $R \in SO(3)$, the 3D rotation group, is an unknown orientation of the volume, and $S(t)$ is a phase shift operator corresponding to in-plane translation in real space by $t \in \mathbb{R}^2$, which models imperfect centering of the volume within the image. The image signal is multiplied by $\hat{g}$, the contrast transfer function (CTF) for the microscope before being corrupted with frequency-dependent Gaussian noise and registered on a discrete grid of size $D \times D$, where D is the size of the image along one dimension.

Under this model, the probability of observing an image $\hat{X}$ with pose $\phi = (R, t)$ from volume $\hat{V}$ is:

$$p(\hat{X}|R,t,\hat{V}) = \frac{1}{Z}\exp\left(\sum_{l}\frac{-1}{2\sigma_l^2}\left|\hat{g}_l A_l(R)\hat{V} - S_l(t)\hat{X}_l\right|^2\right) \tag{5.2}$$

where $A(R)\hat{V} = \hat{V}(R^T(\cdot,\cdot,0^T)$ is a linear slice operator corresponding to rotation by R and linear projection along the z-axis in real space, l is a two-component index over Fourier coefficients for the image, σ_l is the width of the Gaussian noise expected at each frequency, and Z is a normalization constant. We refer the reader to [44] for a review of cryo-EM image formation and reconstruction methods.

Reconstruction algorithms are formulated as optimization of this statistical model, typically done in an iterative fashion with expectation maximization (E-M) or gradient descent-based approaches [44]. In the E-M approach, starting from an initial model, images are aligned with the model (E-step). Then aligned images are "backprojected" to yield an updated estimate of V (M-step). Many software tools exist for 3D refinement[41, 47, 12, 34, 23]. Scheres [40] first proposed a Bayesian framework for *maximum a posteriori* estimation of $\hat{V}$, marginalizing over the posterior distribution of ϕ_i's.

While full marginalization can address uncertainty in pose variables, it is computationally demanding and many algorithms instead use a single maximum likelihood estimate for pose [12, 23, 47, 34]. In these iterative approaches, convergence of E-M to the correct structure depends strongly on the initialization, which is commonly obtained from other data sources, e.g. negative stain EM or approximations from previously-solved, related structures. In Brubaker et al. [4], stochastic gradient descent was proposed for data-driven, *ab initio* reconstruction of a low-resolution initial model, which was implemented in the cryoSPARC software package [34].

Heterogeneous reconstruction: Structural heterogeneity of the imaged protein complex is unknown *a priori* and can present in many forms (e.g. continuous motions vs. discrete compositional changes). Early approaches modeled structures as generated from a discrete mixture model of a small number of volumes [43, 42, 23], and the modelling of continuous motions was seen as a major challenge for the field. Advanced methods for heterogeneity analysis have since been proposed that learn continuous models of molecular variation [27, 33, 9, 26, 21, 20, 55, 5, 32, 58]. These methods all model structural heterogeneity with previously-posed images (e.g. from a homogeneous reconstruction), which limits the scope of heterogeneity analysis to structures where the consensus reconstruction is accurate.

Neural cryo-EM reconstruction: Until recently, all existing cryo-EM reconstruction methods used 3D voxel arrays to parameterize volume(s). Zhong *et al.* proposed cryoDRGN [56], a coordinate-based neural architecture to directly approximate the continuous 3D density function. Preliminary work describing cryoDRGN proposed the joint optimization of pose and heterogeneity with a branch and bound (BNB) algorithm for pose search (referred to as cryoDRGN-BNB in this work) [56]. Later work extended the cryoDRGN approach to real datasets by using image poses from a consensus (homogeneous) reconstruction [55], and has been successfully applied to identify novel structures [55, 13]. CryoGAN presented an alternate paradigm at the proof-of-concept stage for homogeneous reconstruction, which obviates the need for inference of image pose via distribution matching [14]. Recently, learning-based approaches for reconstruction attempt to infer pose by optimizing a parametric function to approximate the posterior over pose variables [37, 28]. These approaches have only been shown on synthetic datasets, and it remains to be seen how robust the optimization of this function is for real cryo-EM data.

Related work in computer vision: CryoDRGN models a continuous volume representation for protein structure that is related to the 3D representation used in other domains of computer vision [22, 30, 46, 45, 25]. Most similar is the neural radiance fields (NeRF) model used for novel view synthesis (NVS) of 3D natural scenes [25]. Unlike the natural image data used to train NeRF and related models however, the cryo-electron microscope produces noisy integral projections and thus the cryoDRGN coordinate-based neural model is specified in the Fourier domain with image generation modeled as central slices instead rendering with ray tracing.

In the standard setup of optimizing NeRF models for NVS [25], camera poses are treated as known. iNeRF inverts this process, and estimates camera poses via gradient descent to backpropagate the loss on a pretrained NeRF model directly into pose parameters [52]. NeRF— performs joint optimization of camera pose and the 3D scene/shape using gradient descent, updating poses from their initial, random values [50]. Here, we show that gradient descent on the cryo-EM reconstruction objective for image pose optimization fails. Instead, we propose an exhaustive pose search procedure performed concurrently with the optimization of the volume representation to achieve state-of-the-art performance on real cryo-EM datasets.

5.3 Method

In this section, we briefly overview the cryoDRGN architecture. Then, we describe the pose search algorithm in cryoDRGN2 and a series of strategies we use to speed up pose search. We then describe our overall training schedule, which alleviates a potential pathology when optimizing neural models under a nonstationary objective.

5.3.1 Overview of cryoDRGN

CryoDRGN parameterizes cryo-EM volumes using a coordinate-based MLP with parameters θ to directly approximate the continuous density function, $\hat{V}_\theta : \mathbb{R}^3 \to \mathbb{R}$ (Figure 5-1c). The model is specified in Hartley space [15] (which is closely related to Fourier space as the real minus imaginary Fourier components for real-valued signal). Thus, input Cartesian coordinates represent Hartley transform coefficients, and cryo-EM images (i.e. integral projections) are 2D central slices of the model whose orientation is determined from the image pose (Section 5.2). In heterogeneous reconstruction, the volume representation is augmented with a latent variable that is learned using amortized variational inference in the framework of variational autoencoders (VAEs) (An overview of the architecture is shown in Figure 5-8). Image poses, $\phi \in SO(3) \times \mathbb{R}^2$, are treated explicitly as geometric operations on a Cartesian coordinate lattice spanning $[-0.5, 0.5]^2$ that are input to the model. Training a cryoDRGN network involves optimizing neural network weights θ, and image poses ϕ_i to maximize the likelihood of the experimental data under the image formation model (Equation 5.1). For more details, see Zhong et al. 2020 [56].

5.3.2 5-D pose search

In neural reconstruction, each model evaluation $\hat{V}_\theta(k)$ of coordinate k (which corresponds to a Fourier coefficient of an image's pixel) requires an expensive MLP evaluation. This is in contrast to voxel-based reconstruction, where image pixel values are computed by linear interpolation. In this work, we rethink the search procedure to minimize the number of neural network evaluations.

In cryoDRGN2, the pose ϕ_i for a given image $\hat{X}_i$ is estimated using a hierarchical search procedure over multi-resolution grids on the space of rotations and in-plane translations. We begin with an exhaustive search in the 5-D space of rotations and in-plane translations at some base resolution, γ_0, followed by an iterative refinement of the K most likely candidate poses by binary search at successively higher resolution grids, $\gamma_1 = \gamma_0/2, ..., \gamma_M$, (Fig. 5-1b). We also employ *frequency marching* [2], where we band-limit the signal to low frequency components of the image, and successively increase the frequency band-limit from k_{min} to k_{max} through the M iterations of pose refinement; this both decreases computational cost and prevents over-fitting on high-frequency noise while the grid is too coarse to align the high-frequency features. Finally, we note that the choice of the base grid resolution γ_0 has a significant impact on the accuracy of pose search, and that the base grid resolution used in state-of-the-art tools was not computationally tractable in cryoDRGN-BNB. In the next sections, we discuss various speedups of the pose search procedure in cryoDRGN2 to enable fast and accurate pose search comparable with traditional state-of-the-art tools.

Speeding up exhaustive search by interpolation

Consider the cost of the exhaustive search procedure. Using a base resolution of 15° and a translation base grid of 14×14 (our defaults) leads to 903,168 pose evaluations for a single image. Each pose evaluation consists of a squared error evaluation between the model $\hat{V}$ and the D^2 pixels of a central slice.

To minimize neural network evaluations, we combine the interpolation ideas from voxel-based reconstruction with our neural model. Instead of evaluating the MLP for each pixel in each pose, for the exhaustive search one can compute a 3D lattice within the frequency cutoff and compute pixel estimates by interpolation. In practice, we use interpolation only for the in-plane rotations, which reduces model evaluations by a factor of 24 (for the 15° resolution grid), taking the exhaustive search step off the critical path.

Interpolation is only accurate if the underlying function is smooth. Smoothness in Fourier space corresponds to the function being flat at large $\mathbf{r}$ in real space, which is satisfied as long as the model output is centered and smaller than the box size. Importantly, it's *not* satisfied for the images, which have a high degree of noise throughout the image; therefore, we found it crucial to interpolate the model output rather than the image.

Leveraging a cheap translation operator

Search over in-plane translations does not require extra model evaluations because translations in real space map to multiplication by an exponential function in Fourier space, which can be computed exactly without additional model evaluations.

For efficiency, we apply translations to the (single) image rather than the 4,608 model estimates at different poses. Computing the optimal pose now consists of finding the minimum mean squared error (MSE between approximately 10^5 model estimates with about 10^2 translated images. Taking advantage of the identity $(A - B)^2 = |A|^2 + |B|^2 - 2A \cdot B$, all the MSEs can be computed as a single matrix multiply between the rotated estimates and the translated images (plus some norms), which is both memory-efficient and very fast on modern CPU and GPU architectures.

The fact that evaluating translations is essentially free leads us to a new approach for pose refinement, which effectively factorizes the search over $SO(3) \times \mathbb{R}^2$ to independent searches over $SO(3)$ and $\mathbb{R}^2$ under some (standard) assumptions. In earlier work, the top $K \leq 24$ most likely candidate poses were selected for refinement; a grid of 2^{3+2} new poses was evaluated for each candidate [56]. In practice, these candidates often corresponded to multiple translations of the *same rotation*, while other promising rotations were discarded. In this work, we instead pick the $K \approx 8$ most likely candidate *rotations* and the single most-likely translation for each of these rotations, t^*; at the next resolution of pose refinement, we search a grid of 2^3 new rotations at the higher resolution but check a large grid of candidate translation grid points centered around t^* at 2x the resolution with .5x the translational grid extent (see Fig. 5-1b)[1]. This allows us to pursue a larger number of candidate poses, and makes the algorithm less sensitive to the choice of translation resolution.

5.3.3 Training Schedule and Model Re-initialization

Traditional cryo-EM reconstruction consists of an expectation-maximization procedure of alternating pose inference (E-step) and volume estimation (M-step). In neural reconstruction, the volume estimation consists of gradient descent on a reconstruction loss for the maximum-likelihood poses [56]. However, we observe that the representation quality of the coordinate-based MLP is limited by the number of gradient updates, and the computational cost of each update is dominated by the pose search procedure (about 10x slower with pose search than without). Thus, in cryoDRGN2 we increase the number of gradient updates by reusing each computed pose for N gradient updates. Specifically, we alternate training epochs that perform pose search with epochs that reuse the latest computed poses (Fig. 5-1e). For simplicity we set a constant pose search frequency (e.g. N=5), however, further speedups can likely be achieved with different (e.g. exponential) training schedules.

Vanishing gradients

We observed a pathology in neural network training when performing *ab initio* reconstruction on a particularly challenging dataset. Due to the alternating updates of pose and volume, the neural network training objective changes during the course of training: in early epochs, the pose estimates are less accurate leading to an in-

[1] Since we only check a local grid of translations around the minimum for each rotation candidate, a key assumption for this approach is that the loss surface with respect to translation is unimodal given the rotation. This is satisfied for biological datasets with respect to translation. We note that it's not satisfied for rotations, e.g. molecular complexes with symmetries will have a local minimum at each symmetry operator offset from the global minimum, so it is crucial that multiple candidate rotations be refined.

ability to resolve features for large $|k|$; as a result, SGD minimizes the L2 loss by predicting a constant function at these high frequencies (Fig. 5-4a). Later in training when high-frequency features can be resolved, the gradient of these high-frequency predictions is 0 with respect to the input k and model parameters, leading to an inability to update the volume approximation given the new poses (Fig. 5-4c). We validated that this is a vanishing gradients issues rather than e.g. a local minimum of the loss, by explicitly computing the gradient $d\hat{V}_k/dw_j$ and $d\hat{V}_k/dk$ for the coordinate k highlighted in Figure 5-4c) and observing that they are zero.

Training pathologies due to sparse or vanishing gradients to the parameters is well documented and a variety of solutions have been proposed [6, 7]. However, these analyses typically focus on supervised learning, whereas we conjecture that it is precisely the non-stationarity of the objective that leads to this pathology. We found that proposed solutions like a Leaky ReLU activation [24] or residual corrections [17] did not fully solve the problem.

We found that resetting the coordinate MLP model and optimizer state intermittently during training (while retaining the image poses inferred from the old model) resolved the vanishing gradient problem, as shown in Fig. 5-4. The training schedule including model reset is illustrated in Fig. 5-1e. We leave a further analysis of this vanishing gradient problem as well as alternative methods for warm-starting training from an old model [1] to future work.

5.3.4 Hyperparameters

A listing of the cryoDRGN2 pose search algorithm and its hyperparameters is given in Section 5.6.1. To choose reasonable defaults for the base resolution γ_0, number of grid subdivision M, kept poses per subdivision K, and frequency marching bounds k_{min} and k_{max}, we perform a hyperparameter sweep of possible values, evaluated by aligning a subset of images to a pre-trained cryoDRGN model (Section 5.6.1).

As training speed depends on many external factors, we do not perform an ablation of each of these techniques to assess computational speedups. Instead, we verify that our overall pose search algorithm is accurate and compare the overall training time of

Method	**Grid setting**	**Time**	**Accuracy**
cDRGN-BNB	$30^\circ, 2.8$ pix	0:01:32	0.691
cryoDRGN2	$30^\circ, 2.8$ pix	0:00:23	0.643
cryoDRGN2	$15^\circ, 1.4$ **pix**	**0:00:52**	**0.004**

Table 5.1: Comparison of pose search algorithm and hyperparameter choices. Timing and accuracy (mean rotation error) are measured for the alignment of 1000 images from the 80S dataset on a pretrained cryoDRGN model.

our reconstruction method relative to prior work (Table 5.1).

We note that existing traditional reconstruction methods achieved high pose accuracy with a $\gamma_0 = 15^\circ$ or 7.5° grid on $SO(3)$ (which corresponds to 4,608 or 36,864 rotations, respectively)[59], but cryoDRGN-BNB [56] was restricted to $\gamma_0 = 30^\circ$ (576 rotations) due to computational limitations. Depending on the smoothness of the underlying objective, using too coarse of a search resolution can lead to missing the global minimum. With the techniques described above, we are able to use a base resolution of 15° or even 7.5°, leading to much higher pose accuracy (Table 5.1).

5.4 Results

We qualitatively and quantitatively evaluate cryoDRGN2 for *ab initio* reconstruction in both homogeneous and heterogeneous settings. We first validate our pose search algorithm on synthetic homogeneous datasets (*hand, spike*) and compare to baseline methods. Next, we perform homogeneous reconstruction on three real cryo-EM datasets of variable difficulty (*80S, RAG12, spliceosome*). We highlight a particularly challenging test case of the RAG1-RAG2 complex. Lastly, we show heterogeneous reconstruction on synthetic and real heterogeneous cryo-EM data (*Linear1d, spliceosome*).

5.4.1 Homogeneous reconstruction of synthetic datasets

Data and setup: We create two synthetic *homogeneous* datasets from a ground truth volume of a hand and of the SARS-CoV-2 spike protein (PDB: 6VYB) [49] by

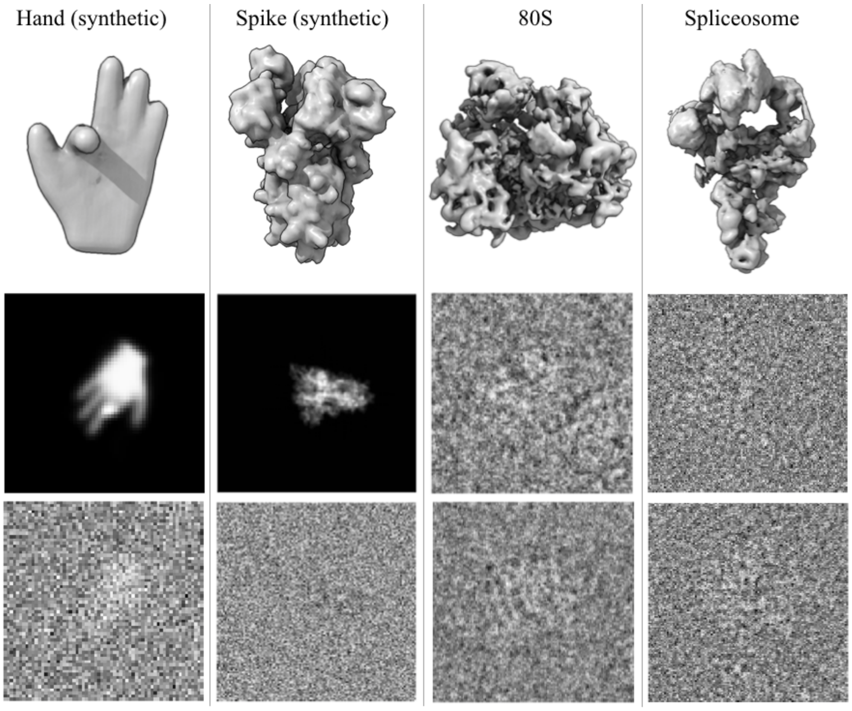


Figure 5-2: Ground truth (synthetic) or reference (real) volumes with corresponding example cryo-EM images below. Synthetic datasets show a noiseless and a corresponding noisy image (SNR=0.1).

following the standard image formation model (50k images, D=64/128 for *hand*/*Spike*, see Section 5.6.2 for more details). We test on both the noiseless version and a noisy (SNR=0.1), realistic version of the dataset. We compare cryoDRGN2 against two methods that use a branch and bound algorithm for pose search: prior work cryoDRGN-BNB [56] and cryoSPARC [34], a state-of-the-art, traditional (voxel-array-based) software for cryo-EM reconstruction. Results with cryoSPARC are obtained from *ab initio* reconstruction followed by homogeneous refinement in cryoSPARC v2.15. We additionally compare the performance of cryoDRGN2 to two other paradigms for pose estimation: a learning-based method for pose inference (pose-VAE) where we use a variational encoder to predict 3D pose variables and the direct gradient-based optimization of pose variables (pose-GD). For pose-GD, we randomly initialize 3D pose variables, and initialize the volume from a pre-trained model. Additional experimental details are in Section 5.6.3.

We find that cryoDRGN2 obtains high accuracy on our synthetic datasets similar to other pose search algorithms (cryoDRGN-BNB and cryoSPARC). Similar to Ullrich

Method	**Hand**		**Spike**	
	SNR = inf	*SNR = 0.1*	*SNR = inf*	*SNR = 0.1*
Pose VAE	5.99/6.66	5.97/6.64	5.98/6.67	5.98/6.65
Pose GD	5.97/6.61	5.97/6.65	5.96/6.63	5.97/6.66
cryoSPARC	0.012/0.002	0.692/0.071	0.0007/0.0003	0.065/**0.002**
cryoDRGN-BNB	0.39/0.007	**0.06**/0.25	0.10/0.0006	0.066/0.012
cryoDRGN2	**0.002/0.0003**	0.086/**0.027**	**0.0004/0.0001**	**0.057**/0.011

Table 5.2: Rotation pose accuracy for homogeneous reconstruction of noiseless and noisy synthetic cryo-EM datasets quantified by mean/median error between predicted rotations $\{\hat{R}_i\}$ to the ground truth rotations $\{R_i\}$. The rotation error for image i is defined as the square Frobenius norm $||R_i - \hat{R}_i||_F^2$ after a rigid 6-D global alignment of the set of images.

Method	**80S**		**RAG12**		**Spliceosome**	
	Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
cryoSPARC	**0.0186**	**0.0001**	3.7806	0.3084	**0.0853**	**0.0015**
cryoDRGN-BNB	0.6151	0.0020	4.1621	4.6371	2.2187	0.1854
cryoDRGN2	0.0578	0.0008	3.4254	0.0386	0.1958	0.0046
cryoDRGN2+r	0.0590	0.0008	**3.3730**	**0.0226**	0.1947	0.0044

Table 5.3: Rotation pose accuracy for homogeneous reconstruction of real cryo-EM datasets quantified by mean/median error between predicted rotations $\{\hat{R}_i\}$ to the reference rotations $\{R_i\}$. The rotation error for image i is defined as the square Frobenius norm $||R_i - \hat{R}_i||_F^2$ after a rigid 6-D global alignment of the set of images.

et al. [48] we find that the gradient-based approaches perform poorly, likely due to the non-convexity of the objective with respect to pose. Pose errors to ground truth poses are given in Table 5.2. Visualizations of the reconstructed volumes of the *Hand* are shown in Figure 5-9.

5.4.2 Homogeneous reconstruction of real datasets

Data and setup: We use three experimental cryo-EM datasets publicly available on the EMPIAR database: the 80S ribosome (EMPIAR-10028) [51], the RAG1-RAG2 complex (EMPIAR-10049) [39], and the pre-catalytic spliceosome (EMPIAR-10180) [31, 27]. Images were downsampled to D=128 for all experiments. Real datasets have varying degree of difficulty due to differences in contrast (i.e. signal) for different

Method	**80S**		**RAG12**		**Spliceosome**	
	Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
cryoSPARC	**0.0008**	**0.0006**	**0.0096**	**0.0033**	0.1674	0.1663
cDRGN-BNB	0.0071	0.0030	0.0745	0.0775	0.0418	0.0172
cryoDRGN2	0.0022	0.0015	0.0324	0.0265	0.0094	0.0036
cryoDRGN2+r	0.0022	0.0015	0.0338	0.0311	**0.0093**	**0.0035**

Table 5.4: Translation pose accuracy for homogeneous reconstruction of real cryo-EM datasets quantified by mean/median error between predicted in-plane translations $\{\hat{t}_i\}$ to the reference translations $\{t_i\}$. The translation error for image i is defined as $||t_i - \hat{t}_i||^2$ after a rigid 6-D global alignment of the set of images.

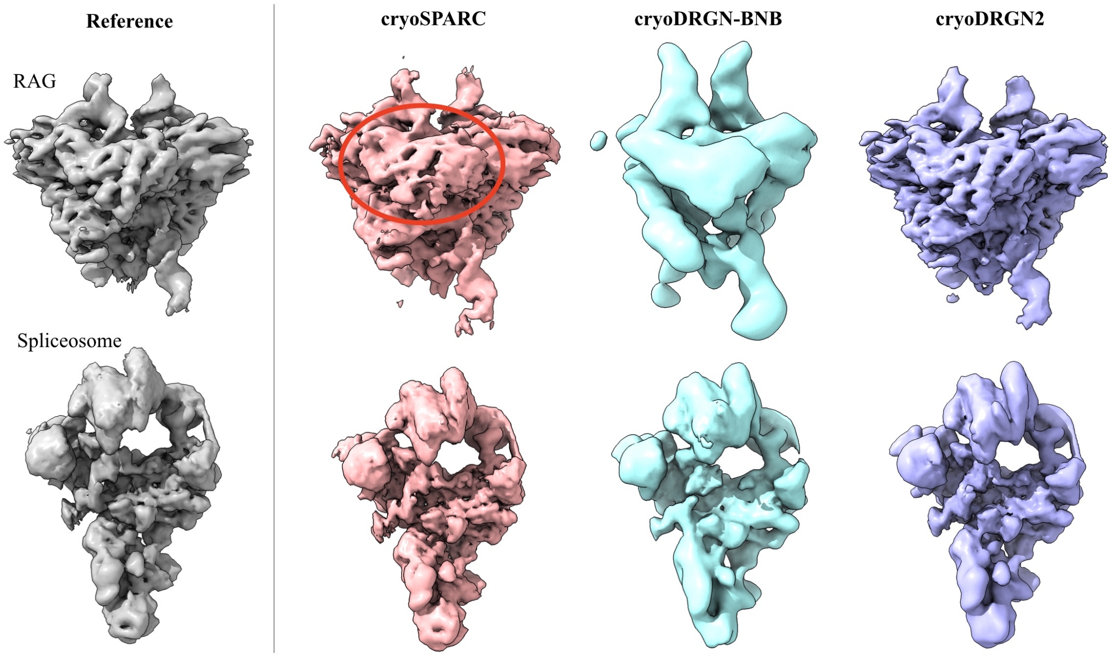


Figure 5-3: Reconstructed volumes from different homogeneous *ab initio* reconstruction algorithms and the reference volume.

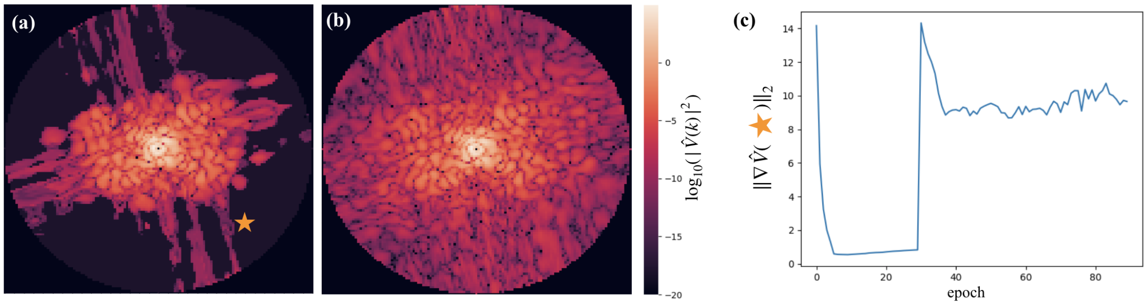


Figure 5-4: Power spectral density $|\hat{V}(k)|^2$ of a slice from the neural volume at different stages of training. **(a)** Model slice after 30 epochs of joint optimization of pose and $\hat{V}$. **(b)** Model slice after the model was reinitialized and trained for 30 epochs using fixed poses from **(a)**. **(c)** L2 norm of the gradient of a dummy loss computed at the starred coordinate in **(a)** with respect to last layer of weights of $\hat{V}$.

molecules, non-uniformity of pose distributions, and varying degrees of underlying structural heterogeneity and symmetry. As real datasets lack ground truth, to produce a reference model for comparison, we train a cryoDRGN coordinate-based MLP using published poses [55]. We note that the published structures were originally obtained using prior knowledge from other related complexes (as their initial models for refinement), and in the case of the spliceosome, also involved many rounds of hierarchical processing due to the heterogeneity of the complex. We baseline against cryoDRGN-BNB and cryoSPARC *ab initio* reconstruction followed by homogeneous refinement (Additional details in Section 5.6.2).

On real cryo-EM datasets, cryoDRGN2 is able to obtain high quality structures *ab initio*, matching that of the reference refinement and competitive with existing *ab initio* methods. We report the difference in the estimated poses to the reference poses in Tables 5.3 and 5.4. Visualizations of the reconstructed volumes of the RAG and spliceosome datasets are given in Figure 5-3. We additionally quantify that the reconstructed volumes match the ground truth using Fourier Shell Correlation (FSC) curves (Figures 5-10,5-11 and Tables 5.7,5.8).

80S: The 80S ribosome dataset is a common cryo-EM benchmark dataset with high contrast images and static structure, and all methods perform well with low pose error (Tables 5.3,5.4) and good FSC metrics (Tables 5.7,5.8, Figures 5-10,5-11).

RAG: The RAG complex is a much more challenging dataset, e.g. a replicate of cryoSPARC refinement using the same initial model of the published structure yielded a 0.91/0.03 mean/median rotation error between replicates. The discrepancy between the mean and median statistic in Table 5.3 is likely from the approximate 2-fold symmetry of the core of the complex. In the qualitative comparison of the reconstructed volumes, we observe that high resolution features are more resolved in the cryoDRGN2 volume than in cryoSPARC (Figure 5-3). CryoDRGN-BNB only produces an approximately correct low-resolution shape (Figure 5-3). We also observe improvements of the cryoDRGN2 volume relative to the *reference volume* in the (non-symmetric) DNA extensions of the complex, likely due to alignment of images to the correct symmetry copy by cryoDRGN2 (Figure 5-12).

Spliceosome: CryoDRGN2 produces a volume closely matching the reference with low pose error (Fig. 5-3). Relative to cryoDRGN2, cryoDRGN-BNB has higher pose error and captures an approximately correct, though much lower resolution shape. Initial results with cryoSPARC were poor (e.g. the median image alignment error was 5.8 for rotations), however once images were recentered based on published poses, cryoSPARC was able to produce a high quality consensus reconstruction (Figures 5-3,5-10,5-11).

Importance of model reset

We identified a pathological case of vanishing gradients which we hypothesize results from distributional shifts in pose variables when training on the RAG complex dataset (discussed in Section 5.3.3). We note that the RAG dataset pose distribution is highly skewed towards a preferred orientation. We observe that while the initial round of cryoDRGN2 training produced a low resolution structure matching the reference, the model output at high resolution (large $|k|$) was essentially zero (visualization in Fig. 5-4). We compute the norm of the gradient of the last layer during the stages of training (Fig. 5-4), showing disappearance of the gradient in the initial stage and the recovery of gradient information after model reset. Further refinement of the model with pose search (cryoDRGN2+r), is able to learn high resolution features with concomitant improvement in pose accuracy (Fig. 5-4 right, Fig. 5-1d). This

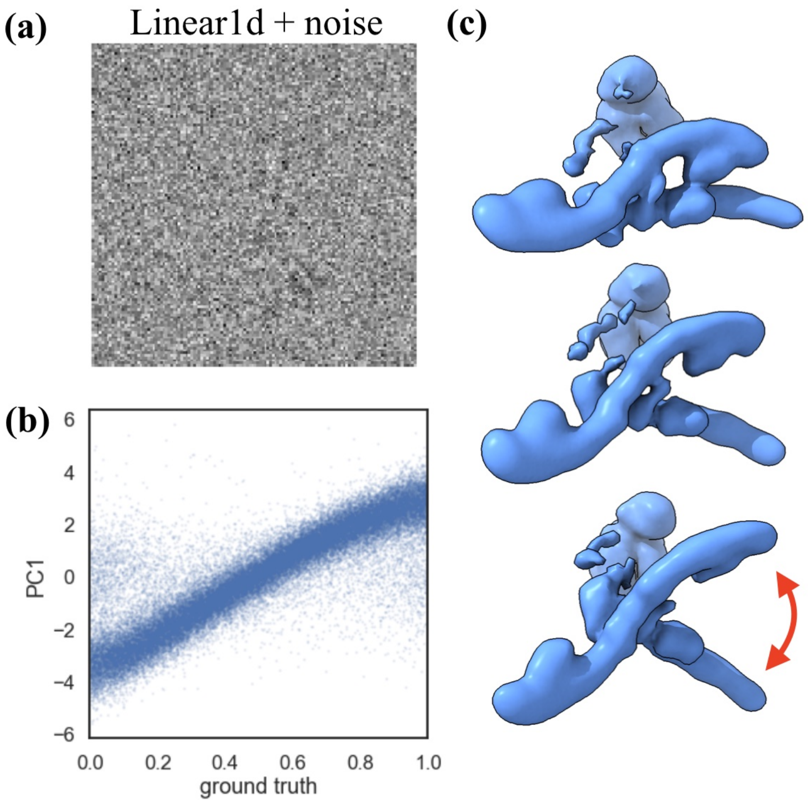


Figure 5-5: (a) Example images of the Linear1d dataset. (b) CryoDRGN2 latent embeddings of particle images. (c) CryoDRGN2 reconstructed structures along the PC1 axis of the latent embeddings.

observation motivates our multi-stage training procedure (cryoDRGN2+r), and may be relevant in other application domains of neural volume rendering.

5.4.3 Heterogeneous reconstruction

Data and setup: We perform *ab initio* heterogeneous reconstruction (i.e. joint inference of $\hat{V}_z$ and ϕ_is) on two datasets that contain large structural variations. The Linear1d dataset is a synthetic dataset containing a large continuous 1D motion [56]. We generate a dataset containing 50k images with CTF and noise (SNR=0.1, D=128) from 50 ground truth models simulating a continuous motion (Fig. 5-5a). We also test cryoDRGN2 on the pre-catalytic spliceosome dataset (EMPIAR-10180), which contains large continuous motions [27, 55, 33]. We compare against cryoDRGN-BNB as all other methods for reconstructing continuous heterogeneity require previously

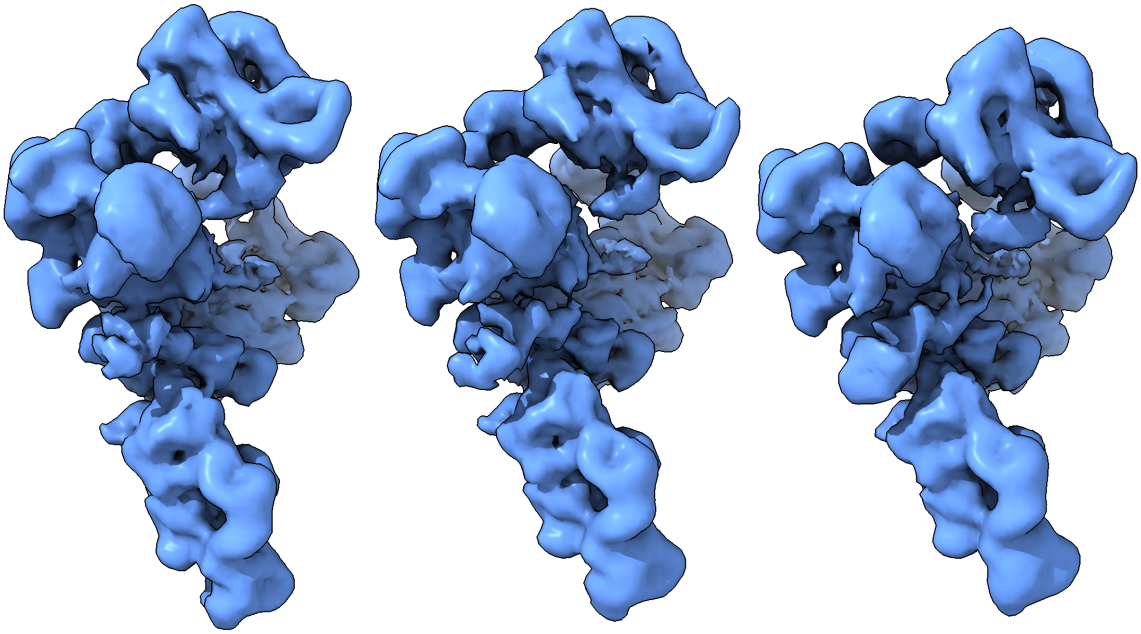


Figure 5-6: CryoDRGN2 reconstructed volumes of the spliceosome generated along the PC1 axis of the latent embeddings.

assigned poses.

We find that CryoDRGN2 is able to reconstruct the underlying continuous 1D motion of the synthetic dataset (Fig. 5-5, Fig. 5-13). Trained on the spliceosome dataset, cryoDRGN2 volumes sampled along the PC1 axis of the latent embeddings show large-scale flexing of the molecular complex (Fig. 5-6), similar to previous analyses with pose supervision [27, 55, 33]. CryoDRGN-BNB captures the same qualitative motions in these datasets however volumes are lower resolution (Figure 5-14). Visualizing the inferred pose distribution highlights local minima in the cryoDRGN-BNB pose search (Fig. 5-7).

5.5 Conclusion

We present cryoDRGN2, a method for reconstructing single or heterogeneous distributions of protein structure from unlabelled 2D cryo-EM images. By addressing inaccuracies and computational bottlenecks in earlier unsupervised optimization of cryoDRGN models, we demonstrate that neural models can achieve state-of-the-art

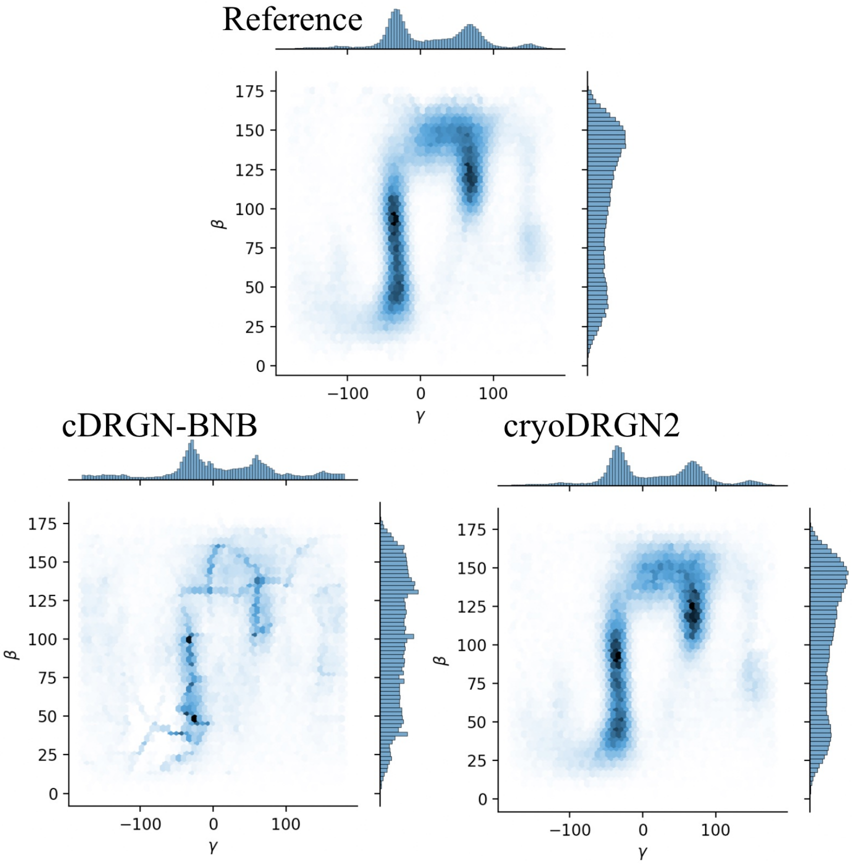


Figure 5-7: Pose distribution of the spliceosome dataset inferred from *ab initio* heterogeneous reconstruction with cryoDRGN2 and cryoDRGN-BNB.

accuracy for *ab initio* reconstruction of challenging, real cryo-EM datasets. Although we reanalyze publicly available datasets here, we are optimistic that this and future improvements will be fruitful for structure determination of novel datasets, especially for structurally heterogeneous complexes, for which no other reconstruction algorithms exist. The techniques shown here may be useful in other domains in computer vision, including graphics, inverse rendering, and robotics.

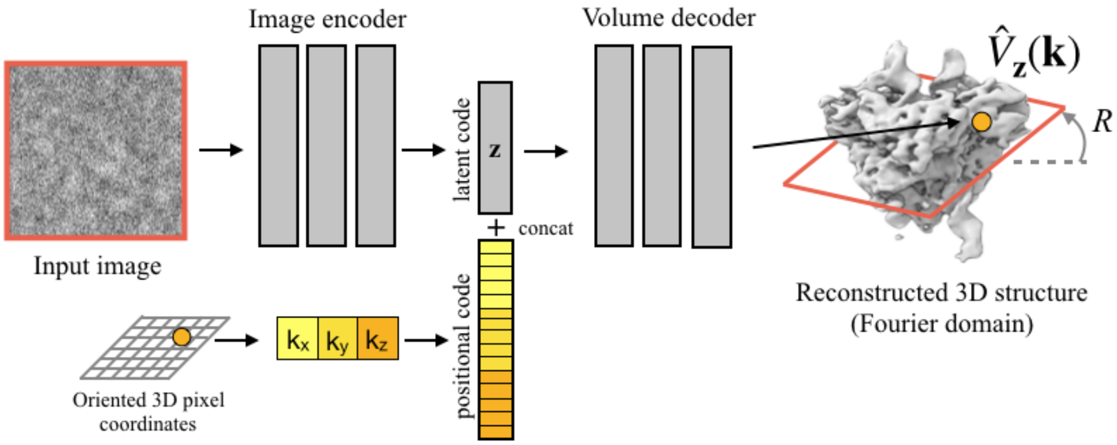


Image encoder
Volume decoder
$\hat{V}_{\mathbf{z}}(\mathbf{k})$
R
latent code
z
+ concat
Input image
positional code
k_x k_y k_z
Oriented 3D pixel coordinates
Reconstructed 3D structure (Fourier domain)

Figure 5-8: The cryoDRGN architecture for heterogeneous cryo-EM reconstruction. Adapted from Zhong et al [56].

5.6 Supplemental details

5.6.1 Additional Methodological Details

cryoDRGN background and architecture

Figure 5-1c shows the architecture used in cryoDRGN's neural representation for 3D volume and Figure 5-8 shows an overview of the cryoDRGN method for heterogeneous cryo-EM reconstruction.

CryoDRGN directly approximates the volumetric density function, $V : \mathbb{R}^3 \to \mathbb{R}$, with a coordinate-based MLP where each input Cartesian coordinate is featurized with a fixed positional encoding function consisting of D sinuosids of varying frequency:

$$pe^{(2d)}(k) = cos(\gamma(d)k); d = 0, ...D/2-1 \tag{5.3}$$

$$pe^{(2d+1)}(k) = sin(\gamma(d)k); d = 0, ...D/2-1 \tag{5.4}$$

$$\gamma(d) = D\pi \left(\frac{2}{D}\right)^{d/(D/2-1)} \tag{5.5}$$

Without loss of generality, the volume is represented on a sphere centered at the origin with radius 0.5. Wavelengths of the positional encoding follow a geometric series from $2/D$ to 1, i.e. the Nyquist limit of the imaging dataset to the width of the volume.

For noisy (e.g. real) datasets, we follow prior work [56, 55] and scale all wavelengths of the featurization by 2π. Volume and image data are represented in Hartley space [15], which is closely related to Fourier space as the real minus imaginary component for real-valued signals. We use a white noise error model; the loss is computed as the mean square error between the input images and central slices from the 3D volume.

For heterogeneous reconstruction, we use the coordinate-based neural network to parameterize a generative model over volumes (Figure 5-8). While the latent pose variables are constrained as their geometric operations on the input coordinates, the unconstrained variables are directly input as additional dimensions to the MLP. We learn the generative model by amortized variational inference based on the standard variational autoencoder (VAE): An encoder predicts the approximate conditional posterior, $q(z|X)$, whose form is an isotropic Gaussian. A sample from this distribution is broadcast to all pixel coordinates of an (oriented) image. Then the decoder is then evaluated pixel-by-pixel to reconstruct the slice. The prior on the latent variables $p(z)$ is a standard normal. The image orientation is unknown for *ab initio* reconstruction; we therefore propose an efficient search algorithm over the 5-D space of poses performed within the training loop.

5-D pose search

Given a cryo-EM image and the current weights of the volume decoder, $\hat{V}$, we perform a grid search over the complete space of poses, $SO(3) \times \mathbb{R}^2$ (Algorithm 2). Our method exhaustively evaluates a discretization over the 5-D space at a base grid resolution and iteratively refines the top K poses ($K = 8$ by default), doubling the resolution of the grid for M iterations ($M = 4$ by default). When computing the error for a given pose at the base resolution, we band-limit the image and model slice to $|k| < k_{min}$ for computational efficiency and to prevent overfitting to high-frequency noise. During the refinement iterations, the frequency band-limit is linearly increased up to k_{max}. As an additional computational speedup, we initialize k_{max} to a low value and let it increase over multiple epochs of training as the volume representation in early epochs will be lower resolution.

The uniform, incremental grids on $SO(3)$ are parameterized using the Hopf fibration [53], which is composed of the product of the Healpix [11] grid on the sphere and an ordinary grid on the circle S^1. The cryoDRGN2 base grid on $SO(3)$ has 4,608 rotations with spacing of 15° by default. For in-plane translations, we use a grid centered at the origin with extent, $[-t_{ext}, t_{ext}]^2$ and a spacing of $2 * t_{ext}/T = 2 * (20 \text{ pixels})/14 \approx 2.86$ pixels for D=128 images by default. As detailed in Section 5.3.2, after the base resolution poses are evaluated, we select the top K unique rotations, and subdivide the grid to get $8K$ new rotations at the next incremental grid. We select the best rotation t^*, then generate a new grid centered at t^* with half the extent and the same number of grid points for the next incremental grid.

Algorithm 2 CryoDRGN2 pose search

```
procedure OPTPHI($\hat{X}, \hat{V}$)  ▷ Find the optimal pose for $\hat{X}$ given the current decoder $\hat{V}$
    $k_{min} \leftarrow 12,\ k_{max} \leftarrow min(48, D/2),\ M \leftarrow 4,\ K \leftarrow 8$
    $\gamma \leftarrow 15°$
    $\Phi \leftarrow (\phi, (0,0))$ for $\phi$ in $SO(3)$ rotation grid at resolution $\gamma$.
    $\Psi \leftarrow$ Translation grid in $\mathbb{R}^2$ at base resolution centered at the origin.
    $k \leftarrow k_{min}$                                   ▷ frequency cutoff for computing $err$
    for $iter = 1 \ \dots \ M$ do
        $errors \leftarrow \{\}$
        for $(\phi, t) \in \Phi$ do                          ▷ Compute error for all rotations
            $err, dt^* = \underset{dt \in \Psi}{\arg\min} \left|\left|\text{SLICE}(\hat{V}_z, \phi) - S_l(t + dt)\hat{X}\right|\right|^2$
            $t^* = t + dt^*$
            $errors \leftarrow errors \cup (err, \phi, t^*)$
        $\Phi^* \leftarrow$ the $K$ pairs $(\phi, t^*)$ with lowest $err$ from $errors$.
        $\gamma \leftarrow \gamma/2$                                   ▷ Halve rotation grid spacing
        $\Psi \leftarrow \Psi/2$                                   ▷ Halve translation grid spacing
        $\Phi_{new} \leftarrow \{\}$
        for $(\phi, t) \in \Phi^*$ do
            $\Phi_{new} \leftarrow \Phi_{new} \cup \text{SUBDIVIDE}(\phi_i, \gamma, t)$  ▷ Subdivide rotations to 8 new rotations at 2x resolution
        $\Phi \leftarrow \Phi_{new}$
        $k \leftarrow k + (k_{max} - k_{min})/(N_{iter} - 1)$          ▷ Increase frequency band limit
    return the min-err element of $\Phi^*$
```

Hyperparameter sweep

We perform a grid search over possible hyperparameter values, including the base resolution for rotations (γ), the base resolution for translations (T), the frequency band-limit bounds (k_{min}, k_{max}), the number of subdivisions of the grid (M), and the number of kept poses at each subdivision (K) (Table 5.5). Each of the settings are evaluated by computing the alignment error for 1000 held out images of the synthetic spike dataset, the real 80S, and the real RAG dataset, aligned on a cryoDRGN model pre-trained using either ground truth (synthetic) or reference (real) poses.

We find that for the tested datasets, using a base resolution beyond 15° does not lead to improved accuracy. Increasing the maximum band-limit frequency k_{max} or the number of poses to refine K leads to increased accuracy, though increased training time. Our chosen defaults ("base" in Table 5.5) are $\gamma = 15°$, $K = 8$, $M = 4$, $k_{min} = 12$, $k_m ax = 48$, $T = 14$, and $t_{ext} = 20$. With these settings the final grid resolution for poses is 0.94° for rotations and 0.18 pixels for translations.

5.6.2 Additional dataset details

Example images and reference volumes for the datasets used in our study are shown in Fig. 1a, Fig. 5-2, and Fig. 5-5. Dataset statistics are provided in Table 5.6.

Dataset	**D**	N_{images}	**Å/pixel**	**Noise?**	**CTF?**
Hand ideal	64	50,000	6	N	N
Hand noisy	64	50,000	6	SNR 0.1	N
Spike ideal	128	50,000	3	N	N
Spike noisy	128	50,000	3	SNR 0.01	Y
80S [51]	128	93,852	3.76875	Y	Y
RAG [39]	128	108,544	1.845	Y	Y
Spliceosome [16]	128	139,722	4.25	Y	Y
Linear1d	128	50,000	6	SNR 0.1	N

Table 5.6: Dataset statistics

Synthetic datasets

Synthetic datasets were created by following the standard image formation model: Given a ground truth voxel array, images were generated by rotating the volume by a rotation matrix R, where R is uniformly sampled from $SO(3)$, projecting the volume along the z-axis, then shifting the resulting image by t, where t is uniformly sampled from $[-t_{ext}, t_{ext}]^2$ pixels. For the Hand dataset, the volume depicting a hand

	Spike			**80S**			**RAG**		
	MSE	*MedSE*	*Time*	*MSE*	*MedSE*	*Time*	*MSE*	*MedSE*	*Time*
base	0.208	0.002	0:00:56	0.004	0.0006	0:00:56	3.436	0.164	0:00:56
legacy	0.799	0.007	0:00:26	0.648	0.0023	0:00:26	4.171	4.993	0:00:26
k=[12,20]	0.259	0.004	0:00:31	0.004	0.0011	0:00:30	3.512	0.363	0:00:31
k=[12,48]	0.165	0.001	0:01:50	0.003	0.0004	0:01:49	3.325	0.107	0:01:49
k=[12,64]	0.159	0.001	0:03:01	0.003	0.0003	0:03:00	3.353	0.097	0:03:00
k=[10,20]	0.271	0.005	0:00:28	0.019	0.0011	0:00:28	3.636	1.645	0:00:28
k=[10,32]	0.208	0.002	0:00:53	0.015	0.0006	0:00:53	3.508	0.242	0:00:53
k=[14,20]	0.260	0.004	0:00:33	0.004	0.0011	0:00:32	3.514	0.405	0:00:32
k=[14,32]	0.196	0.002	0:00:59	0.004	0.0006	0:00:58	3.437	0.164	0:00:58
$\gamma_0 = 30°$	0.872	0.004	0:00:51	0.706	0.0017	0:00:50	4.127	4.901	0:00:50
$\gamma_0 = 15°$	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.385	0.142	0:00:56
$\gamma_0 = 7.5°$	0.219	0.002	0:01:28	0.001	0.0005	0:01:32	3.469	0.200	0:01:27
M=1	0.213	0.011	0:00:23	0.019	0.0095	0:00:23	3.601	1.256	0:00:23
M=2	0.210	0.004	0:00:34	0.006	0.0030	0:00:33	3.414	0.166	0:00:34
M=3	0.208	0.003	0:00:44	0.004	0.0011	0:00:44	3.451	0.160	0:00:44
M=4	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.412	0.144	0:00:56
M=5	0.219	0.002	0:01:08	0.004	0.0005	0:01:07	3.385	0.129	0:01:07
M=6	0.225	0.002	0:01:19	0.003	0.0005	0:01:23	3.436	0.177	0:01:19
M=7	0.232	0.002	0:01:29	0.003	0.0005	0:01:29	3.496	0.275	0:01:29
K=2	0.373	0.003	0:00:17	0.037	0.0008	0:00:17	3.860	3.271	0:00:17
K=4	0.256	0.002	0:00:31	0.017	0.0006	0:00:30	3.477	0.214	0:00:30
K=8	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.400	0.141	0:00:56
K=16	0.196	0.002	0:01:50	0.001	0.0006	0:01:48	3.428	0.114	0:01:49
K=24	0.196	0.002	0:02:25	0.001	0.0006	0:02:42	3.420	0.124	0:02:43
$T = 7$	0.208	0.002	0:00:53	0.004	0.0006	0:00:52	3.425	0.155	0:00:52
$T = 28$	0.208	0.002	0:01:13	0.004	0.0006	0:01:12	3.407	0.133	0:01:12

Table 5.5: Hyperparameter sweep for pose search. The default cryoDRGN2 settings ("base") are $\gamma = 15°$, $K = 8$, $M = 4$, $k_{min} = 12$, $k_{max} = 48$, $T = 14$, and $t_{ext} = 20$, and other rows describe modifications to "base". The "legacy" settings approximate the default hyperparameters used in the cryoDRGN-BNB grid, which are $K = 8$, $M = 4$, $k_{min} = 12$, $k_{max} = 20$, $T = 7$, and $t_{ext} = 20$. Mean, median square error, and timing are computed for aligning 1000 images from each dataset aligned on a pretrained model. Rotation error is defined as the squared Frobenius norm of the rotation matrices after global rigid body alignment.

was manually generated on a 64^3 voxel array. For the Spike dataset, the volume was generated by simulating the electron scattering of PDB 6VYB [49] at 6 Åresolution using a grid spacing of 3 Å using the molmap command in UCSF Chimerax [10]. The volume was then padded to a final dimension of 128^3. For the Linear1d dataset, 50 volumes (D=128) sampled along a 1-D reaction coordinate were used as the ground truth. 1000 images were generated for each volume at a resolution of 12 Åand grid spacing of 6 Å/pixel. Image CTF parameters were sampled without replacement from EMPIAR-10028 [51] and applied to each image. Noise was added to each dataset to a signal to noise ratio (SNR) level as specified in Table 5.6, where we define the signal as the whole $D \times D$ image for Hand and Spike and a circle of radius 40 pixels for Linear1d. We note that the signal variance changes drastically depending on how much of the background is included in the SNR computation, thus we show example images for each of our synthetic datasets in Figures 5-2,5-5.

Real datasets

Real datasets for the *Plasmodium falciparum* 80S ribosome (80S) [51], RAG1-RAG2 complex (RAG) [39], and pre-catalytic spliceosome (Spliceosome) [16] were downloaded from EMPIAR at accession codes 10028, 10049, and 10180, respectively. We used the filtered datasets from Zhong *et al.* [55] available on Zenodo [54]. Images were downsampled to D=128 by clipping in Fourier space. A real space windowing function was applied to the images where the corners of the images are scaled to zero using a linear ramp from a radius of 85% to 95% of the image. We use previously published poses [54] as the "ground truth" poses for comparison. These poses were originally obtained via traditional refinement initialized from previously determined structures and thus include prior 3D information. As real datasets also lack a ground truth volume, to generate a reference volume for visual and quantitative comparison, we use the published poses [55] and train a cryoDRGN MLP with 3 hidden layers of width 256 for 30 epochs. We use the same architecture and amount of training to produce a volume that is more directly comparable to the cryoDRGN2 reconstruction (i.e. controls for model capacity).

5.6.3 Experimental setup

cryoSPARC

We use the cryoSPARC software package [34] as representative of state-of-the-art traditional, voxel-based reconstruction methods. CryoSPARC implements both an *ab initio* reconstruction algorithm using stochastic gradient ascent on the posterior probability distribution of the volume given the data to get an approximately correct low resolution structure, and an iterative refinement (E-M) algorithm which requires a roughly-correct initial model as input. In our baseline experiments, we perform *ab initio* reconstruction followed by homogeneous refinement using all default settings in cryoSPARC v2.15. We experimented with non-uniform refinement [35] in cryoSPARC, however it produced slightly worse pose errors and FSC resolutions for the 80S and RAG datasets and similar performance on the spliceosome dataset, and we report results from standard homogeneous refinement.

Pose-VAE

As a baseline for homogeneous reconstruction of synthetic datasets, we test a functional approach to pose prediction by training a VAE for SO(3)-valued pose variables. Briefly, we use the $S^2 \times S^2$ parameterization of SO(3) for the homeomorphic mapping of encoder outputs ($\mathbb{R}^6$) to an element of SO(3) and the modified KL-divergence described in [8]. A sample of pose from the approximate posterior is then used to transform a 2D coordinate lattice, which is then fed into the coordinate-based neural decoder to reconstruct the input image. We test on synthetic, centered images and use the VAE for inference of rotations only. As in our cryoDRGN experiments, networks are 256×3 MLPs with residual connections, and are trained for 30 epochs in minibatches of 8 images. We train on centered images.

Pose-GD

As a baseline for homogeneous reconstruction of synthetic datasets, we test the gradient-based optimization of pose variables. We randomly initialize pose variables (uniformly on SO(3)) and use a pre-trained cryoDRGN MLP on ground truth poses. We jointly update the volume and poses with backpropagation for 200 epochs in mini-batches of 512 images. We train for more epochs to get more updates of pose variables. We train on centered images.

cryoDRGN-BNB

We compare against prior work which performed *ab initio* reconstruction of cryoDRGN models [56]. We use the default settings for pose search and train the same 256x3 coordinate-based MLP architecture for 30 epochs with pose search performed every 5 epochs. Unlike in Zhong *et al.* [56], we do not train on tilt-series pairs.

cryoDRGN2

Unless otherwise specified, all cryoDRGN2 neural networks are instantiated as fully connected neural networks with 3 hidden layers of width 256 and ReLU activations. We use residual connections between layer input and output to whiten gradients, i.e. $y(x) = ReLU(fc(x) + x)$.

In our cryoDRGN2 experiments, we train for 30 epochs total and perform pose search every 5th epoch (where image poses and the model are jointly updated). We additionally test resetting the model with random weights, retraining the model for 30 epochs using the last round of estimated poses, then repeating 30 more epochs of training with pose search every 5 epochs (cryoDRGN2+r). Training is performed in minibatches of 8 images using the Adam optimizer [18] with a learning rate of 1e-4.

For heterogeneous reconstruction, we jointly train an image encoder along with a generated model of 3D volume. Encoder/decoder architectures are 3 layer MLPs of width 256, and we use a 8-D latent variable. We train for 30 epochs with pose search performed every 5 epochs.

Training times varied across datasets. For homogeneous and heterogeneous experiments on the three EMPIAR datasets, training cryoDRGN2 took between 7 and 15 hours (depending on the dataset) on a single V100 Nvidia GPU for 30 epochs of the above training schedule. For comparison, training cryoDRGN with fixed poses for 30 epochs ranged between 3 and 5 hours, and training cryoDRGN-BNB ranged between 18 and 38 hours on a single Nvidia V100 GPU.

5.6.4 Supplementary results – *Ab initio* homogeneous reconstruction

In this section, we provide more detailed experimental results for *ab initio* homogeneous reconstruction. To evaluate reconstruction quality, we focus on image pose error, which allows for direct comparisons of reconstruction accuracy between different model classes (voxel vs. neural). To compute the pose error, we first perform a 6-D rigid body alignment of the reconstructed volume into the frame of the reference volume and transform the predicted poses accordingly. Pose errors are reported as the square Frobenius norm between the reference and predicted rotation matrix, $||R_{ref} - R_{pred}||_F^2$, and the square L2 norm between translation vectors, $||t_{ref} - t_{pred}||_2^2$. Pose errors statistics are summarized in Table 5.2 for synthetic datasets and Tables 5.3 and 5.4 for real datasets.

Reconstructed volumes for the hand dataset for all tested methods are shown in Figure 5-9, and the reconstructed volumes for the RAG and spliceosome datasets are shown in Figure 5-3.

We also quantify the correlation between the reconstructed volumes and the reference volume using Fourier Shell Correlation (FSC) curves [2]. In Table 5.8, we report the map-to-map FSC with a 0.5 criterion (i.e. the resolution at which the FSC curve falls below 0.5) between the reference and reconstructed volumes. In Figure 5-10, we provide the full FSC curves. In Table 5.7, we report the gold standard half-map FSC [38] with a 0.143 criterion, where we performed independent reconstructions on random half subsets of the dataset and compare the resulting volumes, i.e. “half-maps”. In Figure 5-11, we provide gold standard FSC curves.

As described in Section 5.6.3, we train the model in three stages: standard pose

[2] To remain invariant to differences in absolute scale and normalization imposed by different algorithms and to capture resolution-dependent decay, cryo-EM volumes are typically compared as their correlation as a function of radial shells in Fourier space.

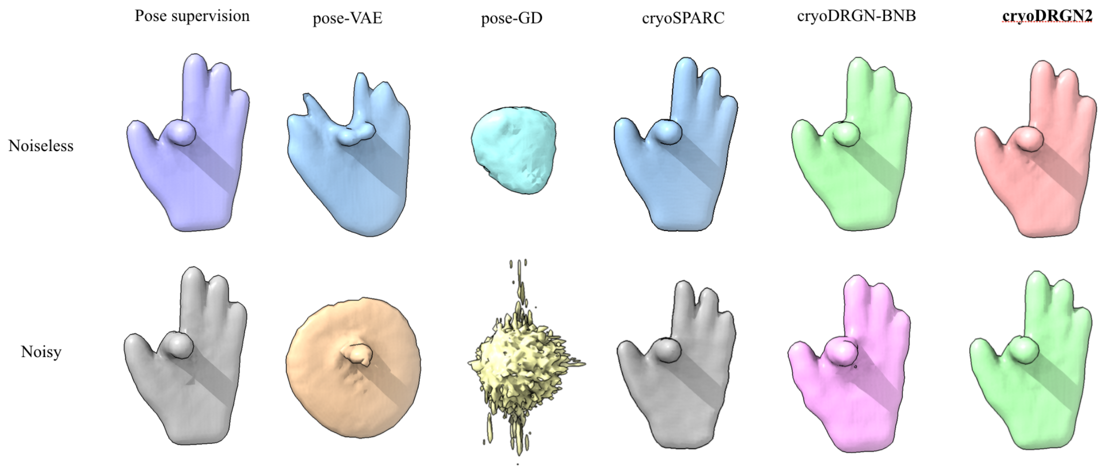


Figure 5-9: Reconstructed volumes of the synthetic Hand dataset from different homogeneous *ab initio* reconstruction algorithms. Noiselss - top row; Noisy - bottom row.

Method	**Hand**	**Spike**	**80S**	**RAG12**	**Spliceosome**
cryoSPARC	4.00	2.17	2.00	2.98	2.00*
cryoDRGN-BNB	4.27	2.21	2.06	4.57	3.88
cryoDRGN2	2.91	2.03	2.00	2.42	2.21
cryoDRGN2+r	N/A	N/A	2.00	2.21	2.06

Table 5.7: Quantitative comparison of the reconstructed volumes to the ground truth (synthetic) or reference volume (real) by an FSC=0.5 criterion. Lower values are better; the best possible is 2 pixels. “cryoDRGN2+r” refers to model reset followed by additional pose refinement. *CryoSPARC originally failed to produce the correct structure (resolution of 32 pix) before re-centering the images.

search interleaved with model updates (cyoDRGN2), we then reset the model and train on fixed poses from the last iteration (cryoDRGN2+r), followed by standard pose search interleaved with model updates (cryoDRGN2+r+ps). While low pose error is often achieved after the first stage, showing that poses may be accurately aligned on low resolution structures, further iterations are useful for converging the neural model. We note that the cryoDRGN MLP is a relatively small architecture (3 hidden layers of width 256; 296,193 parameters). On the 80S dataset, a large ribosomal complex, we repeated cryoDRGN 80S experiments with a larger MLP architecture (5 hidden layers of width 512; 1,510,913 parameters) (Figure 5-11,5-10).

Method	80S	RAG12	Spliceosome
cryoSPARC	2.00	2.42	2.00*
cryoDRGN-BNB	2.00	4.26	2.51
cryoDRGN2	2.00	2.97	2.09
cryoDRGN2+r	2.00	2.13	2.00

Table 5.8: Resolution of the reconstructed volumes from *ab initio* homogeneous reconstruction assessed with the gold standard FSC=0.143 criterion. Lower values are better; the best possible is 2 pixels. “cryoDRGN2+r” refers to model reset followed by additional pose refinement. *The cryoSPARC reconstruction originally failed to produce the correct structure before re-centering the images.

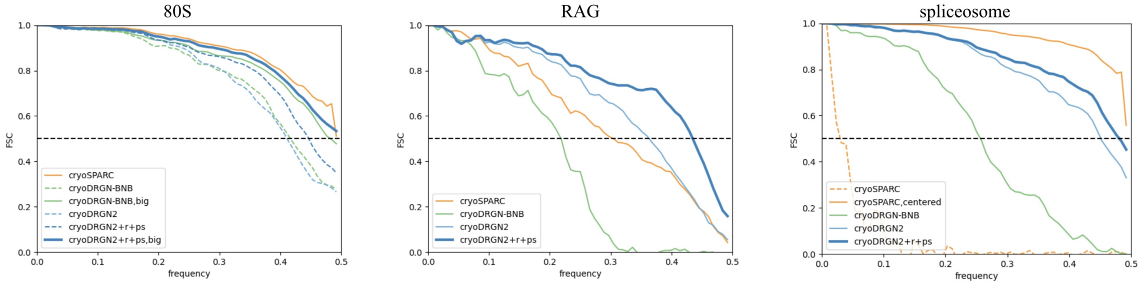


Figure 5-10: Map-to-map Fourier Shell Correlation (FSC) curves computed between the reference volume and the reconstructed volumes. “cryoDRGN2+r” refers to model reset and training on fixed poses from the last iteration. “cryoDRGN+r+ps” refers to model reset followed by additional iterations of pose search. On the 80S dataset, we show FSC curves after repeating *ab initio* reconstruction with a larger MLP architecture. We also show original cryoSPARC results on the spliceosome before image recentering.

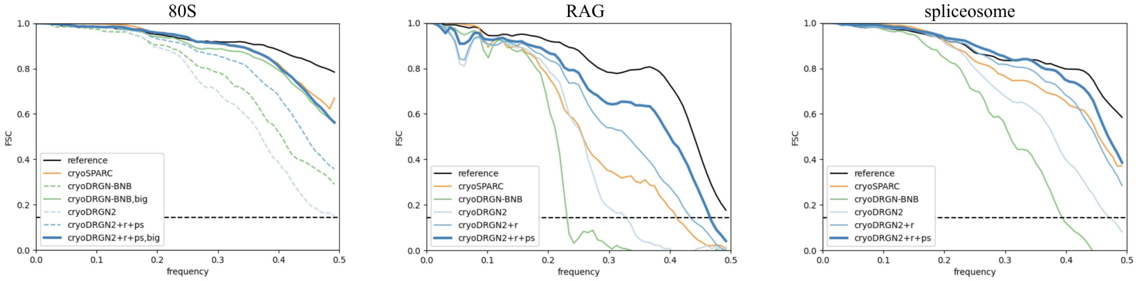


Figure 5-11: Gold standard Fourier Shell Correlation (FSC) curves from *ab initio* homogeneous reconstruction. “cryoDRGN2+r” refers to model reset and training on fixed poses from the last iteration. “cryoDRGN+r+ps” refers to model reset followed by additional iterations of pose search. On the 80S dataset, we show FSC curves after repeating *ab initio* reconstruction with a larger MLP architecture.

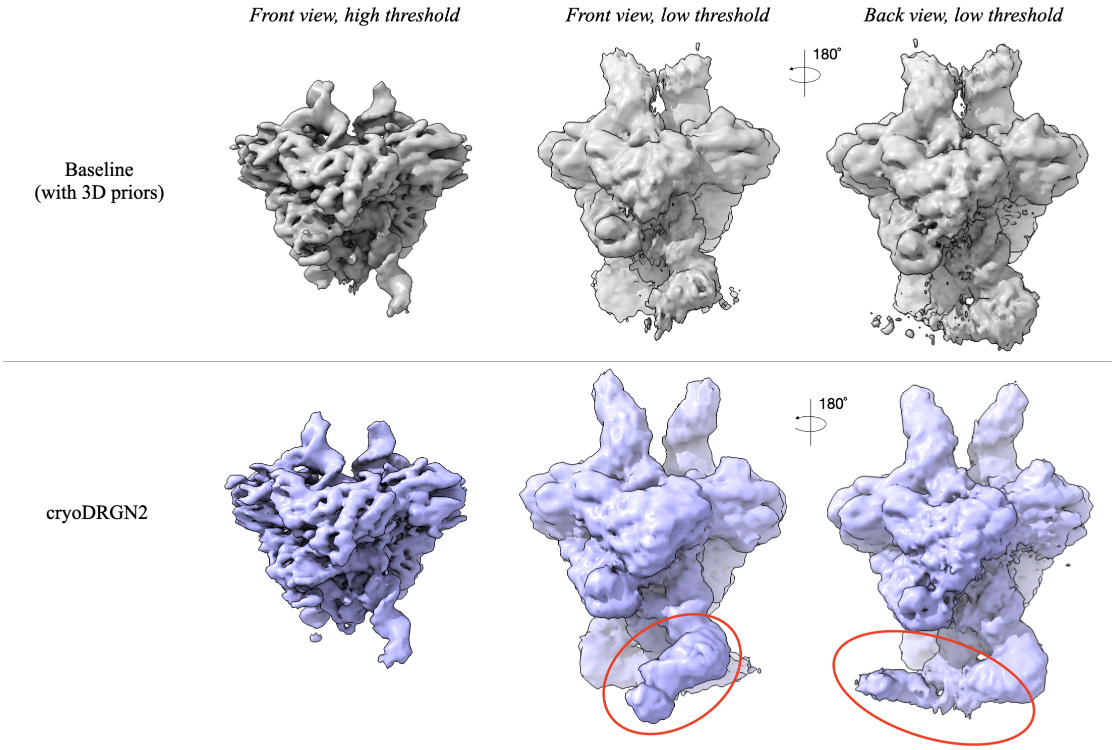


Figure 5-12: Additional views of the cryoDRGN2 reconstructed volume of the RAG dataset [39] and the baseline volume from traditional homogeneous refinement of the published volume. At a low threshold visualization of the volume, additional density of the DNA extensions is visible in the cryoDRGN2 volume (red circles), which are not resolved in the baseline structure.

5.6.5 Supplementary results - *Ab initio* heterogeneous reconstruction

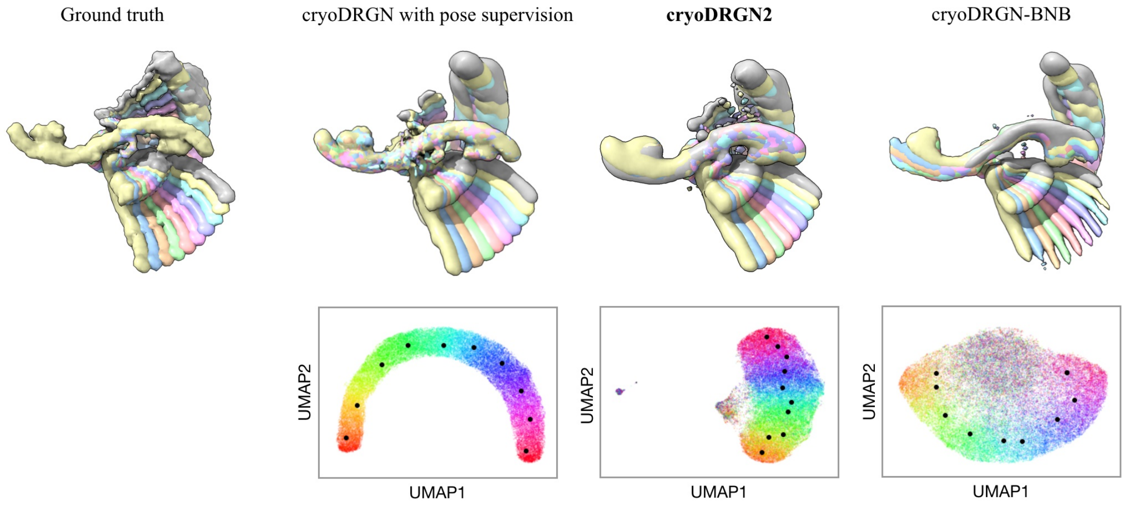


Figure 5-13: **Comparison of heterogeneous reconstruction algorithms on the Linear1d dataset.** *Top row:* Ten representative ground truth and reconstructed volumes along the 1D motion. *Bottom row:* UMAP visualization of the latent space embeddings of images from the dataset, colored by the ground truth reaction coordinate describing the motion.

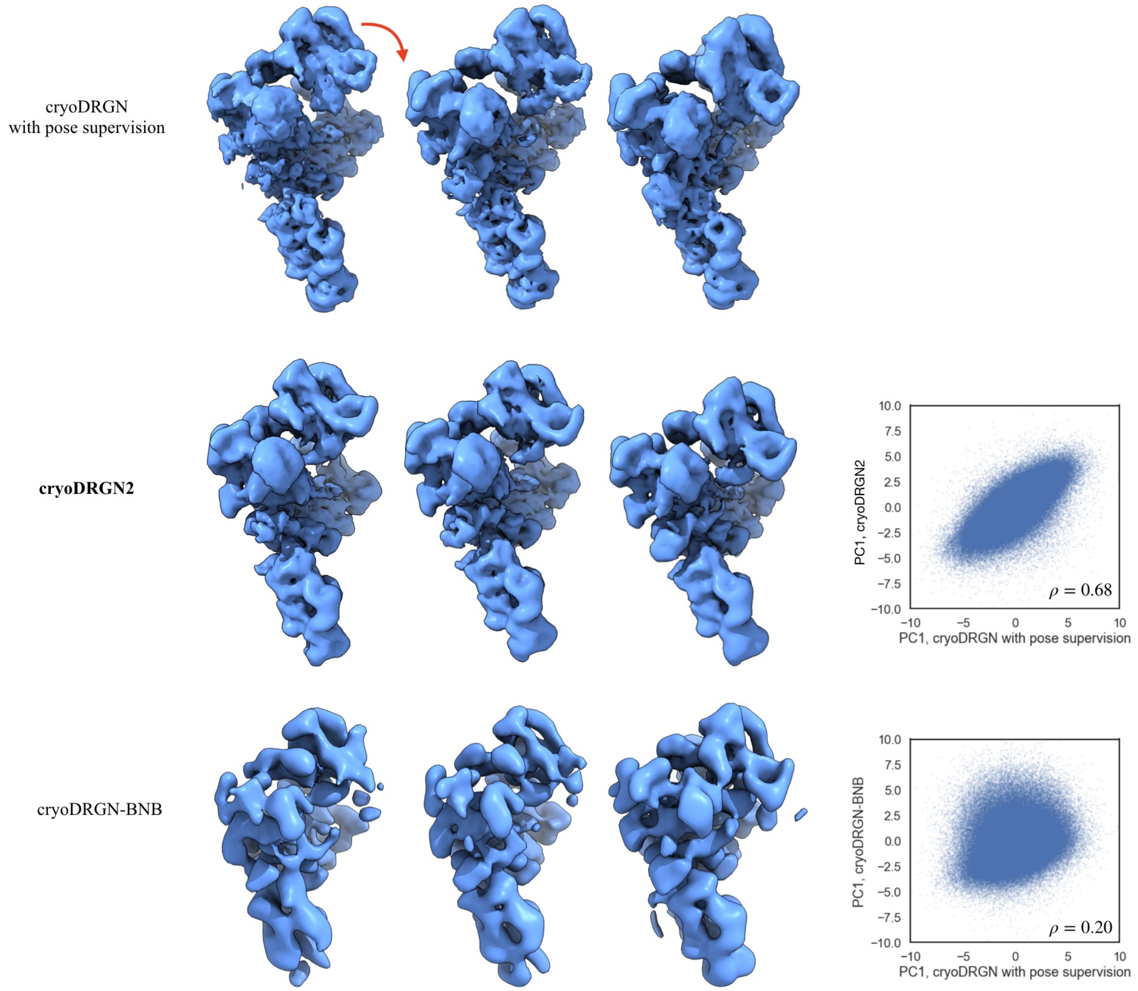


Figure 5-14: **Comparison of heterogeneous reconstruction algorithms on the pre-catalytic spliceosome dataset [EMPIAR-10180] [16].** Reconstructed volumes are generated along the first PC of the latent space embeddings and show a hinging motion of the complex (red arrow). Comparison of the latent embeddings from *ab initio* reconstruction to the previously published cryoDRGN reconstruction with pose supervision [54]. Spearman correlation noted.

Bibliography

[1] Jordan T Ash and Ryan P Adams. On warm-starting neural network training. *arXiv preprint arXiv:1910.08475*, 2019.

[2] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak. Rapid solution of the cryo-EM reconstruction problem by frequency marching. *arXiv.org*, October 2016.

[3] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

[4] Marcus A Brubaker, Ali Punjani, and David J Fleet. Building proteins in a day: Efficient 3D molecular reconstruction. April 2015.

[5] Muyuan Chen, Steven Ludtke, and Verna Marrs. Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. *arXiv*, 2021.

[6] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[7] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *arXiv preprint arXiv:1406.2572*, 2014.

[8] Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in Homeomorphic Variational Auto-Encoding. *arXiv.org*, July 2018.

[9] Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods (San Diego, Calif.)*, 100:61–67, May 2016.

[10] T D Goddard. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.*, 27, 2018.

[11] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

[12] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cisTEM, user-friendly software for single-particle image processing. *eLife*, 7:e14874, mar 2018.

[13] Miao Gui, Meisheng Ma, Erica Sze-Tu, Xiangli Wang, Fujiet Koh, Ellen D Zhong, Bonnie Berger, Joseph H Davis, Susan K Dutcher, Rui Zhang, et al. Structures of radial spokes and associated complexes important for ciliary motility. *Nature structural & molecular biology*, 2020.

[14] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. Cryo-GAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *bioRxiv*, 2020.

[15] Ralph VL Hartley. A more symmetrical fourier analysis applied to transmission problems. *Proceedings of the IRE*, 30(3):144–150, 1942.

[16] David Haselbach, Ilya Komarov, Dmitry E Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell*, 172(3):454–464.e11, January 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Kühlbrandt, Werner. Biochemistry. The resolution revolution. *Science*, 343(6178):1443–1444, March 2014.

[20] Roy R Lederman, Joakim Andén, and Amit Singer. Hyper-Molecules: on the Representation and Recovery of Dynamical Structures, with Application to Flexible Macro-Molecular Structures in Cryo-EM. *arXiv.org*, July 2019.

[21] Roy R Lederman and Amit Singer. Continuously heterogeneous hyper-objects in cryo-EM and 3-D movies of many temporal dimensions. *arXiv.org*, April 2017.

[22] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Transactions on Graphics*, 38(4):14, jun 2019.

[23] Lyumkis, Dmitry, Brilot, Axel F, Theobald, Douglas L, and Grigorieff, Nikolaus. Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of structural biology*, 183(3):377–388, September 2013.

[24] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. March 2020.

[26] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *arXiv.org*, July 2019.

[27] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[28] Youssef S G Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-end simultaneous learning of single-particle orientation and 3D map reconstruction from cryo-electron microscopy data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, October 2021.

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

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. *CVPR*, 2019.

[31] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, jun 2017.

[32] Ali Punjani and David J Fleet. 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. *bioRxiv*, 2021.

[33] Ali Punjani and David J Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.*, 213(2):107702, June 2021.

[34] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, March 2017.

[35] Ali Punjani, Haowei Zhang, and David J Fleet. Non-uniform refinement: adaptive regularization improves single-particle cryo-em reconstruction. *Nature methods*, 17(12):1214–1221, 2020.

[36] Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-Ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature reviews. Drug discovery*, 17(7):471–492, July 2018.

[37] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch, S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes, 2021.

[38] Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, October 2003.

[39] H Ru. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell*, 163, 2015.

[40] Sjors H W Scheres. A Bayesian view on cryo-EM structure determination. *Journal of molecular biology*, 415(2):406–418, January 2012.

[41] Sjors H W Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3):519–530, December 2012.

[42] Sjors HW Scheres. Maximum-likelihood methods in cryo-em. part ii: Application to experimental data. *Methods in enzymology*, 482:295, 2010.

[43] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.

[44] Amit Singer and Fred J. Sigworth. Computational Methods for Single-Particle Electron Cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, jul 2020.

[45] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, Gordon Wetzstein, and Stanford University. Implicit Neural Representations with Periodic Activation Functions. *NeurIPS*, 2020.

[46] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning Persistent 3D Feature Embeddings. *arXiv.org*, December 2018.

[47] 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, January 2007.

[48] Karen Ullrich, Rianne van den Berg, Marcus Brubaker, David Fleet, and Max Welling. Differentiable probabilistic models of scientific imaging with the Fourier slice theorem. *arXiv.org*, June 2019.

[49] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 2020.

[50] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF- - Neural Radiance Fields Without Known Camera Parameters. *CoRR*, 2021.

[51] Wilson Wong, Xiao-Chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H W Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e01963, June 2014.

[52] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. *arXiv*, 2020.

[53] Anna Yershova, Swati Jain, Steven M LaValle, and Julie C Mitchell. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *The International Journal of Robotics Research*, 29(7):801–812, May 2010.

[54] Ellen D. Zhong. zhonge/cryodrgn_empiar: Initial release, January 2021.

[55] 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, 2021.

[56] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

[57] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021.

[58] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-em reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

[59] Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7, nov 2018.

Chapter 6

Landscape analysis

Previous chapters have described machine learning algorithms that provide new capabilities in reconstructing continuous distributions of protein structures. However, once trained, the downstream analysis of reconstructed 3D density maps is largely based on expert-driven, manual inspection, and thus does not scale to modern approaches that produce large ensembles (> 100) of density maps.

In this chapter, we present an efficient and automated volume analysis framework for analyzing the distribution of density maps from a trained cryoDRGN model. Existing analysis approaches described in Chapter 4 have focused on interpreting the low-dimensional latent variable representation of the dataset together with inspecting a small set of sampled volumes from the model. However, the latent space representation is challenging to interpret as it is nonlinearly related to the distribution of volumes. Here, we instead focus our analysis of the learned distribution entirely in the output space of volumes, where distances and dimensions are more directly interpretable. With this guiding principle, we first summarize the learned volume distribution with a sketching algorithm to enable tractable analysis. Then, the sketched volumes are clustered to characterize the modes of the distribution, which can be interpreted as *discrete* conformational states. In parallel, *continuous* reaction coordinates are estimated from the volume sketch. These reaction coordinates provide a natural embedding for volumes, which are then used to embed the full dataset and visualize a conformational landscape that is more interpretable than cryoDRGN's latent variable representation.

This chapter presents unpublished work performed jointly with Ashwin Narayan, Xue Fei, Bob Sauer, Joey Davis, and Bonnie Berger.

6.1 Introduction

Single particle cryo-electron microscopy (cryo-EM) is uniquely poised to visualize complex structural distributions of large, dynamic biomolecular complexes, and several advanced tools for heterogeneous 3D reconstruction have recently been proposed towards this promise [13, 11, 1, 9]. In the cryoDRGN method, heterogeneous reconstruction is framed as unsupervised learning of a deep generative model of 3D density maps parameterized by a neural field representation of structure that is trained from 2D particle images [14]. Central to the cryoDRGN approach is learning a generic latent variable model, which has been empirically shown to model both discrete and continuous forms of structural variability, for example compositional changes from co-factor binding during ribosome assembly [13] and large-scale continuous motions of dynein motor protein complexes [5]. In cryoDRGN's framework of generative modeling, once a model is trained, an arbitrary number of volumes may be reconstructed at sampled values of the latent variable, thus tools are needed to comprehensively explore the reconstructed distribution.

To accommodate the diverse sources of heterogeneity present in cryo-EM data, cryoDRGN possesses a number of interactive and automated processing approaches for analyzing cryoDRGN results. Existing approaches have focused on visualization of the low-dimensional latent embeddings coupled with user-guided exploration of the volume ensemble [13]. However, while the distribution of latent space embeddings may possess interpretable features reflective of the underlying structural ensemble, the objective function of training a cryoDRGN model aims to reconstruct the distribution of the imaged particles without any guarantees that their latent space representation is (visually) interpretable. Here, we instead focus our analysis of the learned distribution on the high-dimensional output space of volumes.

Briefly, we first summarize the volume distribution with a k-means-based sketching of the latent space to enable computationally tractable analysis. Two types of analyses are then performed on the sketch of volumes: 1) an aggolmerative clustering algorithm which produces a small number of summary volumes that can be interpreted as *discrete* conformation states and 2) principal component analysis (PCA), where the estimated eigenvectors (or, "eigenvolumes") define *continuous* reaction coordinates that may be used to interpret the full ensemble of particles. The rationale behind these choices is described in the next section.

Applied on a previously published dataset of the ClpXP protease from Fei *et al.* [3], we automatically identified a new substrate-engaged state comprising 1,255 particles (0.3% of the dataset) that was both missed in traditional 3D classification and was not immediately apparent from visualizing the cryoDRGN latent space. By applying PCA on the volume ensemble, we produced reaction coordinates that provide a more interpretable visualization of the ensemble than cryoDRGN's latent variable representation. A software tool which automates this "landscape analysis" is openly available in cryoDRGN software version 1.0.

6.2 Methods

In this section, we describe the computational pipeline and design choices for analyzing and interpreting a cryoDRGN model (Figure 6-1, top), using the ClpXP protease dataset from Fei *et al.* [3] as an instructive example (Figure 6-1, bottom).

6.2.1 Overview of cryoDRGN outputs

Given a dataset of single particle cryo-EM images, $\{X_1, ..., X_N\}$, cryoDRGN performs heterogeneous reconstruction by jointly training an inference model over images, $q_\xi(z|X)$, (the encoder) and a generative model over volumes, $p_\theta(V|z)$, (the decoder). Once trained, the model may be used to predict a latent variable representation for each image in the dataset, which we refer to as the "latent embedding":

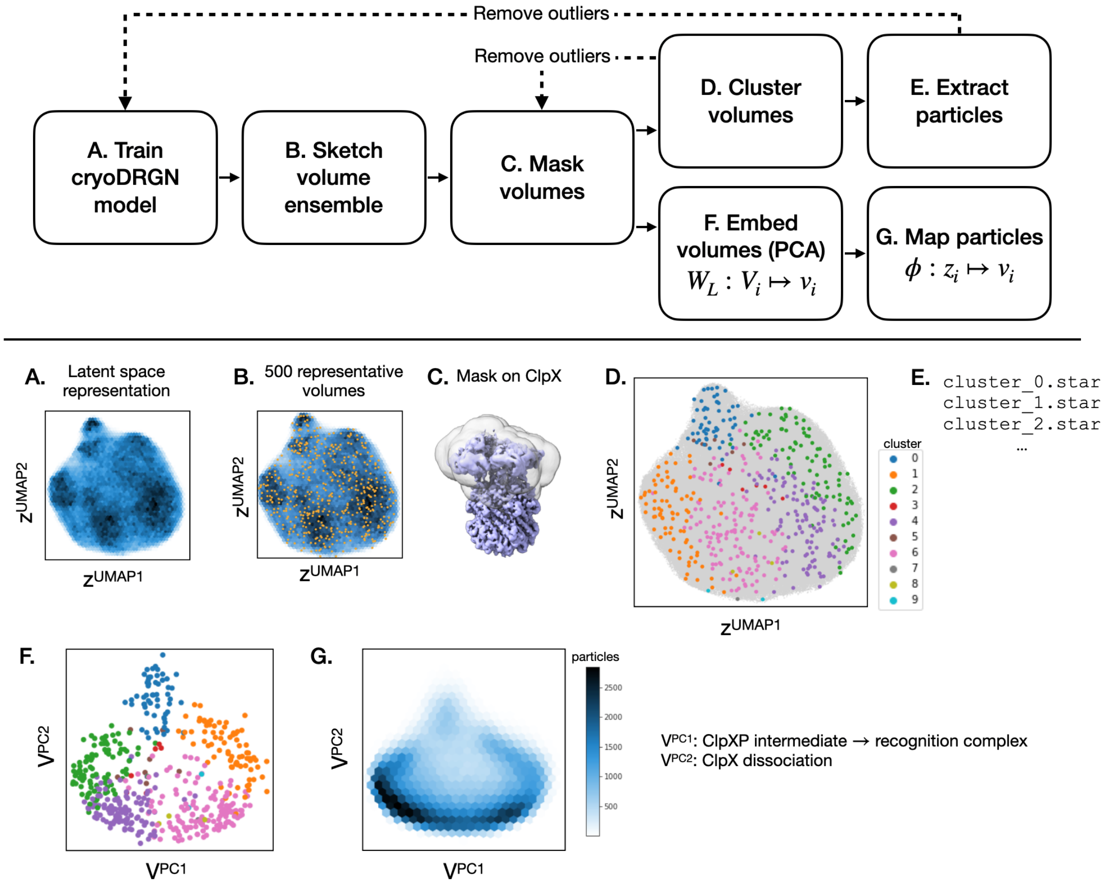


Figure 6-1: **Overview of the landscape analysis pipeline:** We show the general schematic of landscape analysis (top) and its application to the ClpXP protease from Fei et al. [3] (bottom). **A.** First, a cryoDRGN model is trained, so that a latent variable representation z_i can be generated for each image i in the original dataset. **B.** Because generating volumes for the entire dataset is intractable, we *sketch* the set of latent embeddings to find k representative volumes ($k = 500$ here). **C.** A *mask* is applied on the sketched volumes to reduce noise from the background and/or focus on a subset of the volume; the mask shown on ClpXP covers the ClpX complex, which is the part of the protein complex that moves. **D.** and **E.** The 500 sketched volumes are then clustered to summarize *discrete* conformational states and their associated particle lists for any downstream refinement. **F.** We apply principal component analysis (PCA) on the set of sketched volumes to produce a linear map W_L for estimating low-dimensional volume embeddings v_i; the principal components (PC) indicate high variance modes of continuous motion in the structure and can be used to interpret v_i. Cluster assignments from **(D)** are also plotted. **G.** We train a multilayer perceptron (MLP) ϕ to learn the mapping from latent space to the volume PC space. We apply this model to produce a density plot of the full dataset in volume PC space, which may be visualized as a conformational landscape with interpretable axes. Arrows: Clusters can be inspected for outlier junk structures, whose underlying volumes or particles can be excluded to re-analyze the volumes or retrain a cryoDRGN model, respectively.

$$z_i \sim q_\xi(z|X_i) \tag{6.1}$$

and generate an associated volume representation:

$$V_i \sim p_\theta(V|z_i) \tag{6.2}$$

In practice, we define the latent embedding as the *maximum a posteriori* estimate of the (Gaussian) posterior $q_\xi(z|X_i)$, which provides a low-dimensional representation of each image, i.e. $z_i \in \mathbb{R}^N$, where $N = 8$ is typical; the volume is rendered on a 3D lattice for downstream visualization tools, i.e. $V_i \in \mathbb{R}^{D\times D\times D}$, where $D = 128$ or $D = 256$ is typical.

6.2.2 Motivation of volume space analysis

The set of latent embeddings of the dataset $\{z_i\}$ gives a low-dimensional vector representation of the dataset that can be visualized in 2-D with dimensionality reduction algorithms such as PCA, t-SNE, or Uniform Manifold Approximate and Projection (UMAP) [8] (Figure 6-1A). As shown in Chapter 4, the resulting features of the latent embedding distribution can be reflective of structural heterogeneity, such as clusters that correspond to different compositional states. While this may suggest an interpretation of the latent embeddings as an energy landscape, (e.g. where regions of higher/lower particle density correspond to low/high energy states), this interpretation is flawed. Namely the layout of the latent space is arbitrary (hindering interpretation), distances in latent space are not meaningful (z are passed through a nonlinear decoder), and empty regions of latent space do not in general correspond to high energy configurations. Thus, the interpretation of the latent embeddings typically requires annotations from user-guided exploration of the volume ensemble. For example, after training a 8-D latent variable model on the ClpXP dataset (see Section 6.4 for methods), we visualized the final latent embeddings with UMAP (Figure 6-1A). Although there are “features” in the UMAP visualization that suggest interpretable structures (e.g. regions of higher and lower particle density), its interpretation requires expert-assigned inspection of volumes to annotate the various regions. Furthermore, the interpretability of UMAP distances is not reliable.

Our motivation here is to provide an automated and comprehensive analysis approach for the entire ensemble of volumes $\{V_i\}$ to facilitate interpretation of the trained model. However, it is computationally intractable to generate the entire ensemble of volumes $\{V_i\}$ associated with each particle in the dataset due to the computational cost of rendering a volume (seconds per volume) as well as the storage requirement for the voxel arrays (for $10^5 - 10^6$ volumes). Existing cryoDRGN analysis approaches typically generate tens of volumes from different regions of the latent space followed by manual inspection. However, this approach can be time-intensive (for the practitioner) and prone to missing (rare) states that are not sampled, especially because it requires the practitioner to decide *a priori* the regions worth deeper study.

6.2.3 Sketching the volume ensemble

We first generate a *sketch*, or a representative subsample, of volumes from the trained cryoDRGN model that will be used for the downstream structural landscape analysis. The general objective of sketching a dataset is to generate a subsample such that some important properties of the original dataset are preserved. For instance, one can consider naive uniform sampling as a method of sketching, where the property preserved is the probability density of the data. However, a major drawback of uniform sampling is that rare classes will often not end up in a sample unless a very large sketch size is used. For example, in order to capture at least one example of rare substrate-engaged state in ClpXP (see Figure 6-4), which occurs in 0.3% of the samples with a probability of 99%, more than 1,500 samples are needed. See [7] for a discussion of various sketching algorithms for the computational analysis of single-cell RNA-sequencing datasets.

Here, we use a k-means clustering algorithm for sketching: Given a desired sketch size k, we perform k-means clustering on the latent embeddings $\{z_i\}$ (Figure 6-1B). The latent embedding that is closest to each k-means cluster center is then used to generate a volume for the volume sketch. Because k-means attempts to minimize the total variance of all clusters, with sufficiently large k, most points in the dataset should be relatively close to a point in the sketch. We then validate that each each sketched latent embedding yields a representative *volume* for each cluster for a reasonable choice of k, e.g. $k = 500, 1000$. Since each cluster of latent embeddings also represents a set of volumes, we measure the volume-space homogeneity by computing the pairwise L2 distance of volumes for a randomly sampled cluster (Figure 6-6). We use a sketch size of $k = 500$ for all downstream analysis of the ClpXP dataset.

6.2.4 Masking for feature selection

Once the volume sketch is generated, we mask out the voxels of the sketched volumes that correspond to the either background or a user-defined region prior to performing the downstream clustering and dimensionality reduction analysis (Figure 6-1C). A benefit of working in volume space is that the "features" of the volume vector are voxels in the volume representation. This allows us to remove voxels that are known to be irrelevant (by default, the background), which reduces the variation introduced by noise in the masked out region. The remaining variance is thus more likely to be meaningful, which is especially important since variance is fundamental to both our downstream clustering (Ward linkage minimizes cluster variance, see section 6.2.6) and dimensionality-reduction (PCA finds the directions of maximum variance) analyses.

For each of the 500 sketched volumes, we define the region to exclude as the voxels whose density is less than half of the maximum density of the volume; the final binary mask applied to all volumes is the union of the masked out region for each volume. The user can also define the mask on their own, which is especially useful if there is prior information on part of the complex that should be focused on. In the ClpXP example, the ClpX subunit is dynamic, and so that region is manually chosen to analyze in our pipeline (Figure 6-1C). Masking also has the computational benefit of reducing the feature space from 2,097,152 voxels in the 128^3 volume to the 162,210 voxels that are in the mask.

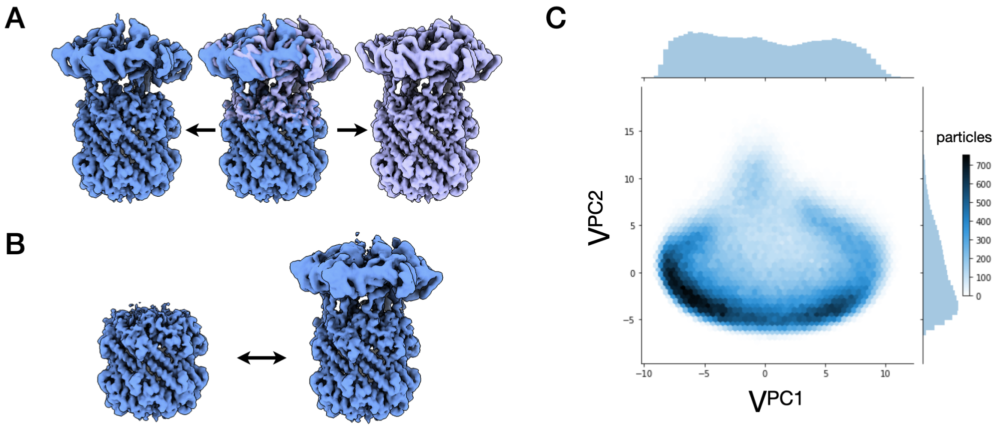


Figure 6-2: **Conformational landscape of ClpXP inferred from PCA of the cryoDRGN volume distribution: A.** Structures traversing principal component 1 shows the transition from the recognition to the intermediate complex. **B.** Structures traversing principal component 2 shows the dissociation of ClpX. **C.** A conformational landscape visualization of all particles mapped to the volume PC space. The methods and motivations for these analyses are described in Section 6.2.5

6.2.5 Visualizing a conformational landscape with PCA

Taking inspiration from previous work [6], we apply principal component analysis (PCA) on the set of masked volumes and use the resulting eigenvector decomposition to visualize the entire ensemble of reconstructed density maps (Figure 6-2). The PCA analysis provides two benefits: 1) the resulting eigenvectors (i.e. “eigenvolumes”) produce trajectories along the axes of maximum variation which can be used to summarize the major modes of motion (Figure 6-2A,B); and 2) the top N principal components (PCs) produce a low-dimensional *volume-space* embedding that can be used to visualize the entire cryoDRGN ensemble (Figure 6-2C).

For the ClpXP protease, traversing the first PC in volume space corresponds to the transition between the ClpXP intermediate and recognition complexes (Figure 6-2A); the second corresponds to dissociation of ClpX (Figure 6-2B); and the third corresponds to appearance of the GFP substrate. Because the PCs form a linear, orthogonal basis, the location of each sketched volume in the PC embedding space can be more easily interpreted (i.e. as the linear combination of the basis vectors). This is in contrast to the latent variable representation (shown in Figure 6-9), where traversals along the PC axes in *latent* space or UMAP coordinates are not guaranteed to provide comprehensive summaries of the *volume* ensemble, due to the nonlinear nature of the decoder.

On the ClpXP protease, the top three principal components capture 65% of the variance in the data (Figure 6-5), making even a three-dimensional representation reasonable. The set of volumes in the sketch may be visualized as a scatterplot in the volume PC space. For example, Figure 6-3A shows a scatter plot of PC1 and PC2 for the sketched volumes, and Figure 6-4A shows PC2 and PC3.

As the PC decomposition is estimated on the 500 volumes in the sketch, we next apply this decomposition to the entire dataset of 344,069 volumes in order to visualize the entire conformational landscape of all the imaged particles (shown in Figure 6-2C). Instead of generating all 344,069 volumes, which is computationally intractable, we learn a function ϕ that maps latent space coordinates for each particle i *directly* to volume embedding space. Specifically, ϕ takes on the form of a simple multilayer perceptron (MLP) network. Having computed the transformation to principal components on the initial 500 sketched volumes, we then compute the volumes of an additional sample of 25,000 latent representations z_i, $1 \leq i \leq 25,000$ and map those into PCA space v_i; these pairs (z_i, v_i) are used to train the MLP ϕ (additional methods in Section 6.4). Once trained, the volume embedding representation of any point in the original dataset z can be computed as $\phi(z)$.

Finally, linear methods do have limitations for visualization, especially in cases where most of the variance is *not* contained in the top few PCs, or where the linear approximation to the underlying nonlinear motions is inaccurate. In those cases, nonlinear methods for visualization can be considered. We show two popular methods for visualization on the ClpXP volume sketch in Figure 6-7: multidimensional scaling (MDS) and UMAP.

6.2.6 Agglomerative clustering for identifying conformational states

We also cluster the volume sketch to summarize the major conformational states of the reconstructed ensemble. We use an agglomerative clustering algorithm and allow the user to vary the number of clusters M, the linkage criterion, and the distance metric. We note that different choices of the clustering hyperparameters emphasize different priors and definitions of what a "cluster" should be. We use agglomerative clustering with the goal that this bottom-up clustering algorithm may be effective at identifying outlier states, including rare states (or junk particles), which would "look different" (under the e.g. L2 distance metric) than the rest of the ensemble, and thus be agglomerated last. Unlike top-down clustering algorithms such as k-means, which first define the differences between clusters, agglomerative clustering does not impose any geometric priors on the shape or size of the clusters. On the ClpXP dataset, agglomerative clustering with $M = 10$ target clusters, a Euclidean distance metric, and an "average" linkage criterion, which minimizes the average distance between the two sets when merging clusters, yields five well-populated and five sparse clusters (Figure 6-3). An example of clustering results with $M = 20$ target clusters is shown in Figure 6-7 and Ward linkage (which minimizes the variance within each cluster) is shown in Figure 6-10.

We compute the centroid of each cluster (i.e. an average of the volumes in the cluster) as a representative structure; these summary states can be quickly and efficiently inspected to evaluate the diversity of the dataset (Figure 6-8). In the case of ClpXP, we find that cluster **0** represents the complex with the ClpX hexamer absent. Since we are interested in ClpX variability, the volumes from cluster **0** can be excluded to avoid the consideration of this state when re-running landscape analysis (not shown) or the underlying particles may be removed from any further cryoDRGN training.

After inspecting the sparse clusters, we found that cluster **8** reflects the substrate-engaged state of ClpXP (see Figure 6-4). This state was both missed in the original 3D classification of this dataset [3] and by expert-guided inspection of the cryoDRGN

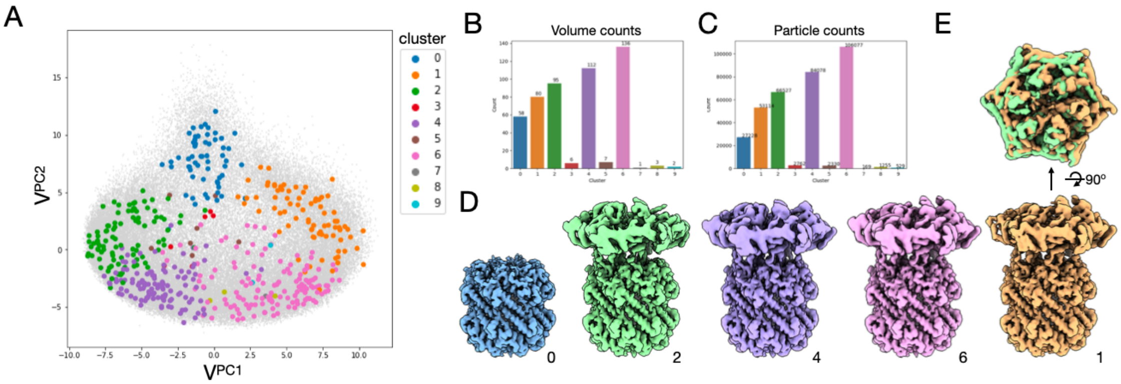


Figure 6-3: **Conformational states of ClpXP inferred from clustering the CryoDRGN volume distribution:** Agglomerative clustering ($M = 10$ clusters) produces five well-populated and five sparse clusters. **A.** The sketched datapoints are colored by their assigned cluster and plotted in volume PC space (from Figure 6-2). **B.** and **C.** The number of volumes and the number of particles for each cluster. Note that some clusters have very few counts, indicating they are outlier groups that might be artifacts or interesting rare conformations. **D.** Representative structures (the centroid of the cluster) for the five most populated clusters. Additional structures are shown in Figure 6-8. **E.** The top-down view of the cluster 1 and cluster 2 volumes from **D** superimposed, highlighting the conformational change between the ClpXP recognition and intermediate complex.

ensemble, yet is biochemically known to be in the sample. Since this cluster represents only 0.3% of the dataset, our focus on ensuring the diversity of possible conformations was covered in our sketching step was crucial for its discovery. We combined the 1,255 particles of the volumes associated with cluster **8** and performed a homogeneous refinement in RELION to validate the presence of this structure (Figure 6-3C). An atomic structure of the GFP substrate was able to be docked into the resulting density map (Figure 6-3D).

6.3 Discussion

The ability of cryoDRGN to model complex structural distributions has raised new questions on how its underlying deep generative model should be interpreted to yield testable structural hypotheses. In particular, the ability of cryoDRGN to reconstruct an arbitrary number of structures, rather than a single or discrete set of structures

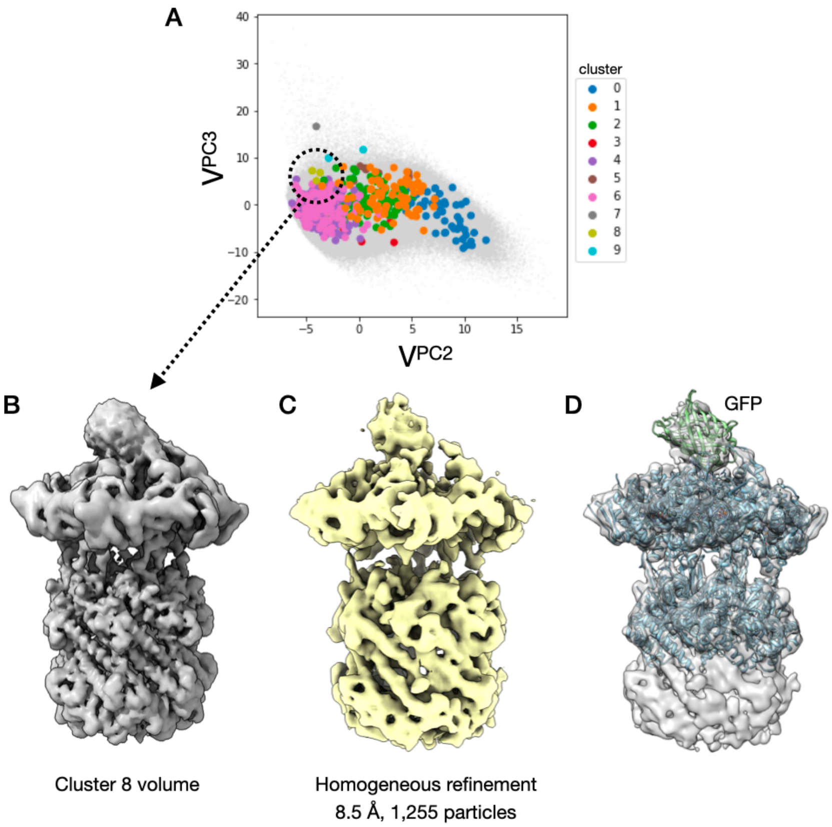


Figure 6-4: **Identification of the ClpXP substrate-engaged state:** Inspecting the cluster **8** structure from Figure 6-3 revealed the ClpXP substrate-engaged state. **A.** Cluster **8** can be identified on the volume PCA plot when comparing PC2 and PC3, where this cluster is more separated. **B.** The representative volume (cluster centroid) for cluster **8**. **C.** Homogeneous refinement in RELION of the 1,255 particles within this cluster. **D.** The density map from (C) with the atomic model docked. Although the GFP substrate is low-resolution, the density of GFP is well aligned with the atomic model.

(tens of structures), presents a novel challenge since examining each structure V_i of the dataset individually is both computationally and manually intractable. Here, we have introduced a "landscape analysis" pipeline that aims to summarize the full diversity of structures in a trained cryoDRGN model for the practitioner. The method is implemented as a tool in the cryoDRGN software for automated analysis, and we have found this approach to be useful for quickly analyzing models, especially in cases where the latent embeddings are visually uninformative.

This landscape analysis pipeline performs two separate but complementary approaches for summarizing the learned distribution: 1) as a small number of *discrete* conformational states (and their constituent particles for further refinement), including rare states of interest or 2) with *continuous* reaction coordinates inferred from PCA that provide an interpretable conformational landscape visualization of the full dataset. The interpretation as discrete or continuous variability, while seemingly at odds, work well together. For one, the choice of method can be tailored to specific structural hypotheses surrounding the dataset of interest. But more generally, the two interpretations may be complementary when both compositional and conformational heterogeneity are present in the dataset, such as in the ClpXP protease, e.g. dissociation of ClpX (compositional) and conformational transitions between the intermediate and recognition complexes (conformational). Even when there is a conformational continuum, it may be useful to discretize the continuum for summary structures. We emphasize that these analyses place different structural assumptions on the ensemble of volumes, and ultimately the choice of interpretation is made by the practitioner.

Unlike PCA-based approaches for reconstruction, such as in Tagare et al. [12] and 3D Variability Analysis [11], PCA is used here to summarize features of a full-rank set of volumes reconstructed by cryoDRGN. While cryoDRGN and other nonlinear methods for heterogeneity analysis can produce complex distributions of density maps, the latent representation is not directly interpretable due to the nonlinear nature of the mapping from latent space to volumes. Here, by separating reconstruction from the downstream volumetric analysis, we can take advantage of both cryoDRGN's powerful nonlinear representation of 3D density maps and established dimensionality reduction techniques to obtain interpretable features, here from PCA.

The analysis of large sets of vectorized volumes (i.e. high-dimensional vector arrays) is a general problem in large-scale, high-dimensional data analysis, and many other algorithms are transferable to this space. For example, this landscape analysis framework can be easily modified to use a different sketching algorithm, clustering algorithm, or volume embedding algorithm. This approach may also be tailored to analyze the results from other heterogeneous reconstruction methods that generate large ensembles of volumes, and may be especially relevant for the growing number of reconstruction methods based on deep learning [1, 10]. Finally, this approach is a purely data-driven approach for analyzing the ensemble of volumes (aside from any user-provided masks), and thus will be less biased, but perhaps less informative than other methods that guide the analysis of the ensemble based on an atomic model [4, 2].

6.4 Additional Methods

6.4.1 cryoDRGN training

CryoDRGN version 0.3.2 models were trained on 344,069 single-particle images of ClpXP from Xue et al. [3] downsampled to an image size of 128 × 128 (2.71875 Angstroms per pixel), with their corresponding poses assigned from a consensus reconstruction in RELION. All reconstructions used an MLP architecture with 3 hidden layers of dimension 1024 for the encoder and decoder networks. The latent variable dimension was 8. Training was performed in minibatches of 8 images using the Adam optimizer and a learning rate of 0.0001. Training was performed on a single V100 GPU and lasted 9 hours and 22 minutes.

6.4.2 Volume mapping

Given a PCA transformation W_L which keeps the top L components, we train a simple MLP network to learn the mapping from latent embeddings z_i to volume embeddings

$v_i = V_i W_L$ to avoid generating V_i for all images in the dataset. A training set of (z_i, v_i) pairs is first generated: 25,000 latent embeddings are sampled from the dataset and used to generate their associated volumes through the decoder. Each volume is generated on the fly and embedded to avoid storing 25,000 voxel arrays. The MLP is trained using a 3:1 training set to validation set split, where the loss on the held out validation set is monitored to prevent overfitting. The MLP is trained for 50 epochs in minibatches of size 64 with the Adam optimizer and a learning rate of 1e-3. The generation of the training set lasted 7 hours and 48 min, and training ϕ for 50 epochs lasted 4 minutes on a single Nvidia V100 GPU.

6.5 Supplemental Figures

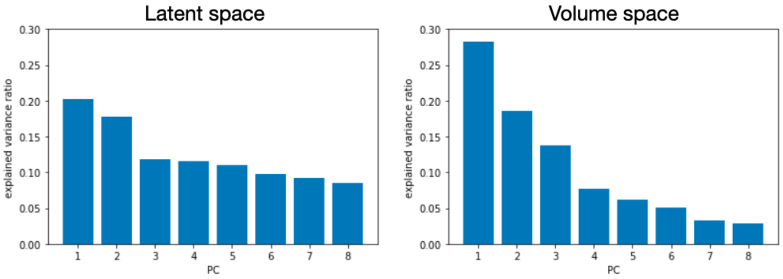


Figure 6-5: Explained variance of the top 8 principal components of the set of latent space embeddings and the volume sketch.

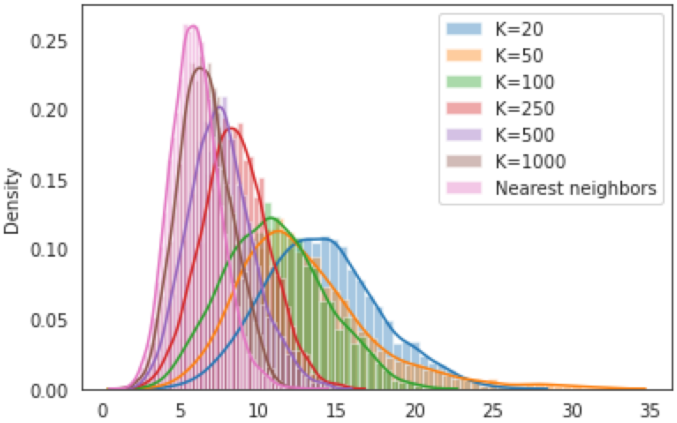


Figure 6-6: Distribution of pairwise L2 distances for the set of volumes in a sketched cluster for different values of k in k-means sketching.

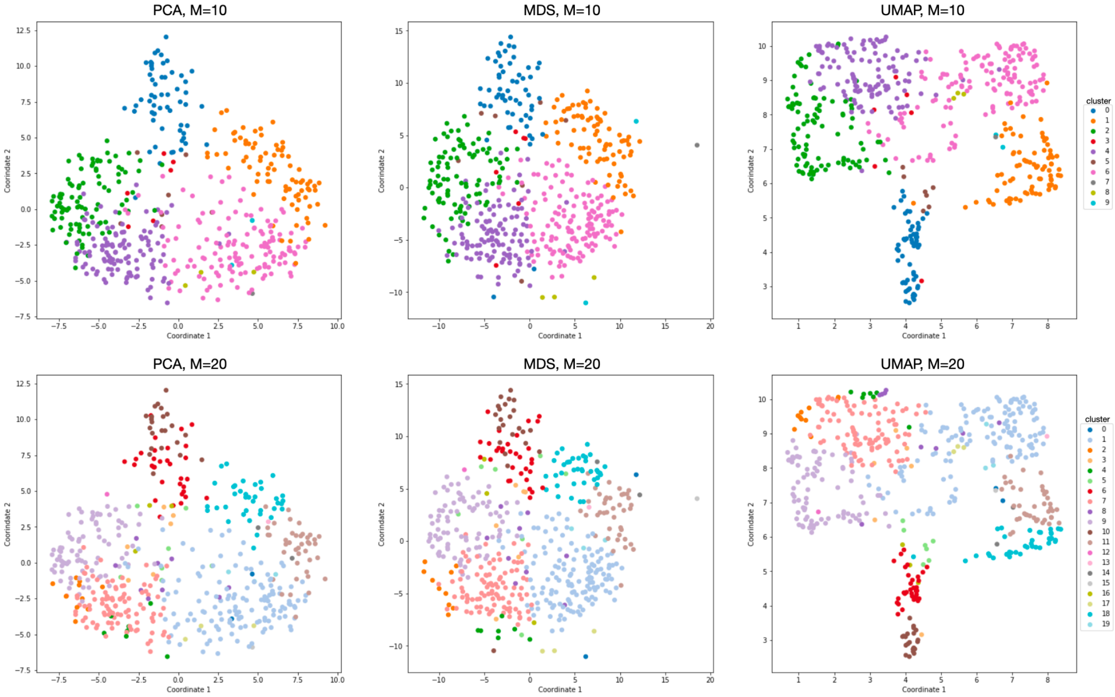


Figure 6-7: Different volume embedding algorithms applied on the sketch of volumes (PCA, MDS, UMAP from left to right). Different choices in in the number of clusters M (top row $M = 10$, bottom row $M = 20$)

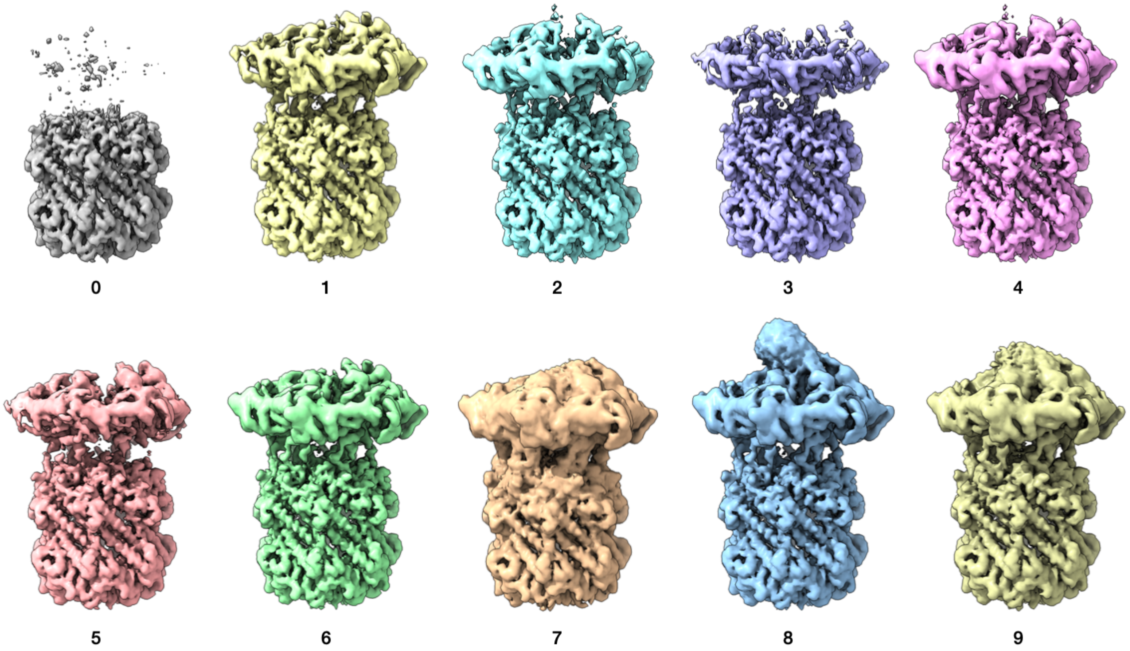


Figure 6-8: Cluster centroids after agglomerative clustering of the ClpXP volume sketch with $M = 10$, an average linkage criterion, and a Euclidean distance metric.

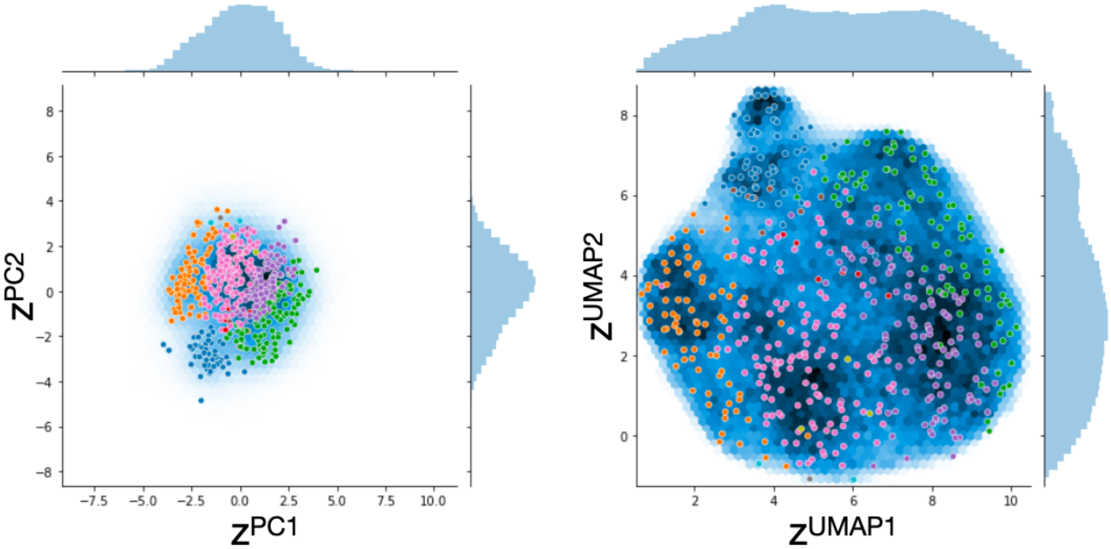


Figure 6-9: Clusters from Figure 6-3 visualized in the latent space representation of the dataset (PCA left, UMAP right)

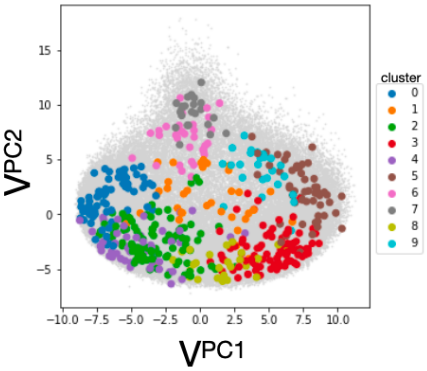


Figure 6-10: Agglomerative clustering of the volume sketch with $M = 10$, a Ward linkage criterion, and a Euclidean distance metric, visualized in the volume space representation of the dataset.

Bibliography

[1] Muyuan Chen, Steven Ludtke, and Verna Marrs. Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. *arXiv*, 2021.

[2] 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.

[3] Xue Fei, Tristan A Bell, Sarah R Barkow, Tania A Baker, and Robert T Sauer. Structural basis of ClpXP recognition and unfolding of ssra-tagged substrates. *Elife*, 9, October 2020.

[4] Julian Giraldo-Barreto, Sebastian Ortiz, Erik H Thiede, Karen Palacio-Rodriguez, Bob Carpenter, Alex H Barnett, and Pilar Cossio. A bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Rep.*, 11(1):13657, July 2021.

[5] Miao Gui, Meisheng Ma, Erica Sze-Tu, Xiangli Wang, Fujiet Koh, Ellen D Zhong, Bonnie Berger, Joseph H Davis, Susan K Dutcher, Rui Zhang, et al. Structures of radial spokes and associated complexes important for ciliary motility. *Nature structural & molecular biology*, 2020.

[6] David Haselbach, Ilya Komarov, Dmitry E Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell*, 172(3):454–464.e11, January 2018.

[7] Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. Geometric sketching compactly summarizes the Single-Cell transcriptomic landscape. *Cell Syst*, 8(6):483–493.e7, June 2019.

[8] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. February 2018.

[9] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[10] Ali Punjani and David J Fleet. 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. *bioRxiv*, 2021.

[11] Ali Punjani and David J Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.*, 213(2):107702, June 2021.

[12] Hemant D Tagare, Alp Kucukelbir, Fred J Sigworth, Hongwei Wang, and Murali Rao. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *J. Struct. Biol.*, 191(2):245–262, August 2015.

[13] 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, 2021.

[14] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

Chapter 7

Generative atomic models for cryo-EM reconstruction

In this chapter, we investigate the possibility of fitting the atomic protein structure directly during the reconstruction process. We have two motivations for this. First, atomic fitting is a labor-intensive step in cryo-EM post-processing, and in particular it is not clear how to perform atomic fitting for models of distributions of protein conformations learned during reconstruction with new tools such as cryoDRGN [28]. Second, modeling the atomic structure of the 3D volume provides a strong prior over structures. In fact, in many cases the protein sequence and an approximate reference structure are known beforehand, strongly constraining the space of feasible 3D volumes. These structural priors are especially important for models of heterogeneous distributions of molecules, because they constrain the conformational dynamics to those that approximate realistic protein motions. Without such priors, it is common to observe artifacts in the motion of the protein, e.g. mass appearing and disappearing between two distinct conformations with rarely-sampled transition states.

To this end, we propose a reconstruction process based on a coarse-grained atomic model for the cryo-EM volume. The model fits parameters including atomic coordinates, to maximize the likelihood of the dataset under a generative model that maps atomic structures to cryo-EM images. When initialized appropriately, this approach is able to learn both homogeneous structures and heterogeneous ensembles from synthetic

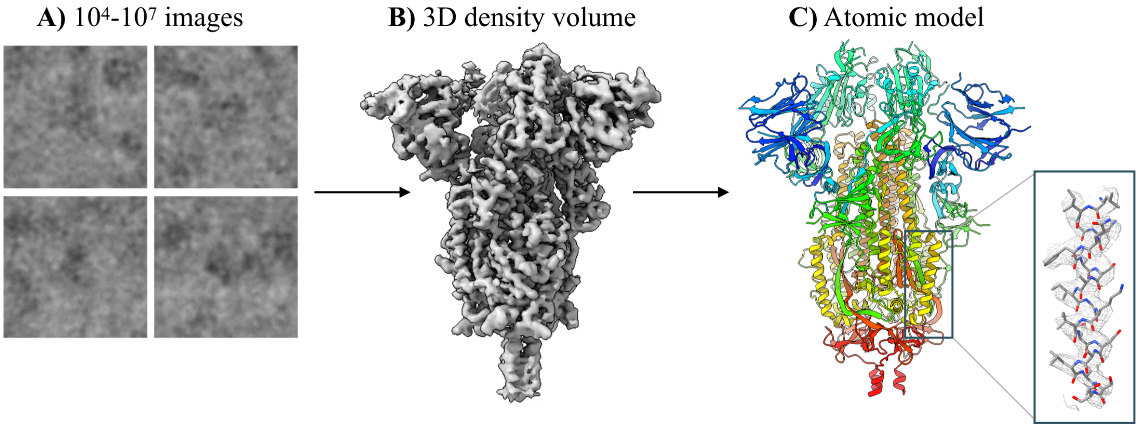


Figure 7-1: Structure determination via cryo-EM. Schematic of cryo-EM reconstruction (**A** → **B**) and atomic model fitting (**B** → **C**). In this work, we investigate the possibility of fitting the atomic protein structure directly during the reconstruction process. Dataset from Walls et al. [26].

cryo-EM images.

This chapter presents work described in [30] performed jointly with Adam Lerer, Joey Davis, and Bonnie Berger and presented at the 2020 NeurIPS Workshop on Machine Learning for Structural Biology.

7.1 Background

A cryo-EM experiment produces a dataset of 10^{4-7} noisy 2D projection images, each containing a unique molecule captured in a random, unknown orientation. The goal of cryo-EM reconstruction is to infer the 3D density volume $V : \mathbb{R}^3 \to \mathbb{R}$ that gave rise to the imaging dataset $X_1, ..., X_N$. As cryo-EM images are integral projections of the molecule in this imaging modality, 2D images can be related to the 3D volume by the Fourier slice theorem [4], which states that the Fourier transform of a 2D projection is a central slice from the 3D Fourier transform of the volume. Traditional methods approximate the volume as a voxel array $\hat{V}(\mathbf{k})$ in Fourier space [25].

To recover the desired structure, cryo-EM reconstruction methods must jointly solve for the unknown volume $\hat{V}$ and image poses $\phi_i = (R_i, t_i)$, where $R_i \in SO(3)$ and $t_i \in \mathbb{R}^2$ are the 3D orientation of the molecule and in-plane image translation, respectively. Expectation maximization and simpler variants of coordinate ascent are typically employed to find a *maximum a posteriori* estimate of $\hat{V}$ marginalizing over the posterior distribution of ϕ_i's [22].

A unique advantage of cryo-EM is its ability to image heterogeneous molecules. *Heterogeneous reconstruction* algorithms aim to reconstruct a distribution of structures from the dataset. A standard approach involves extending the generative model to assume that images are generated from a mixture model of K volumes $V_1, ..., V_K$ [23, 14]. More recently, cryoDRGN proposed a neural model to reconstruct heterogeneous ensembles of particles from cryo-EM data [29]. CryoDRGN represents a continuous n-dimensional distribution over volumes as a function $\hat{V} : \mathbb{R}^{3+n} \rightarrow \mathbb{R}$ approximated by a multi-layer perceptron (MLP) with positional encoding of Cartesian coordinate inputs [29]. To simplify reconstruction, cryoDRGN and other advanced reconstruction methods [28, 18, 8, 15] find it sufficient to use poses ϕ computed using a traditional reconstruction method, and focus on the volume reconstruction.

7.2 Related Work

A large body of work has investigated continuous deformations of protein structures produced from normal mode analysis of atomic, coarse-grained, or pseudo-atomic models [1, 11, 12, 24, 10]. The top N normal modes of a system where pseudo-atomic bonds are approximated by harmonic springs can summarize a molecule's flexing motions, which can then be used to model cryo-EM data during reconstruction [11, 12, 24, 10]. However, it is unclear how accurate the hypothetical motions that are generated from the underlying harmonic spring approximation are. In BioEM and other approaches, likelihood-based analysis of cryo-EM images or reconstructed maps based on an ensemble of atomic models (e.g. generated from molecular dynamics) has been used to investigate conformational heterogeneity [6, 3, 7].

Recently, deep learning has been used to incorporate prior beliefs from an existing dataset of atomic structures into the cryo-EM reconstruction process. Kimanius *et al.* train a convolutional denoising model on PDB structures that is integrated into the cryo-EM reconstruction process in the form of updates to regularization parameters during iterative refinement [13]. They show higher resolution reconstructions of synthetic datasets but observe artifacts from the learned general model in some cases. DeepEMhancer trains a network to perform volume post-processing (e.g. high frequency sharpening and solvent/background masking) from pairs of raw and postprocessed experimental density maps, where the postprocessed maps have been refined by the fitted atomic structure [21].

In concurrent work, Rosenbaum *et al.* propose a VAE with a similar RBF-based generative model over atomic coordinates [20]. While we model proteins at the residue level, Rosenbaum *et al.* model them at the atomic level [20] and also jointly infer image pose. They evaluate this model on synthetic data from a molecular dynamics simulation.

7.3 Method

7.3.1 Cryofold volume model

The central contribution of this work is a cryo-EM density model that is parameterized in terms of coordinates for a coarse-grained atomic model given a known atomic reference structure (Figure 7-2A). In this coarse-grained model, each amino acid is represented by two Gaussian radial basis functions (RBFs), one representing the backbone and the second representing the sidechain. Each RBF is parameterized by (μ_i, σ_i, a_i), where μ_i is the position of the ith RBF; a_i is the amplitude of the ith RBF; and σ_i is the width of the ith RBF. We tie $a_i = a_0 Z_i$ where a_0 is a global learned amplitude constant, and Z_i is the total number of electrons in the fragment represented by RBF i. Furthermore, we tie all σ_i to the same value σ. Thus, the full RBF model has $3K + 2$ parameters, where K is the total number of amino acids in the protein complex. The cryo-EM density can be computed as a function of these parameters as:

$$V(\mathbf{r}) = \sum_i (2\pi\sigma_i)^{-3/2} a_i \exp\left(\frac{-||\mathbf{r} - \mu_\mathbf{i}||^2}{2\sigma_i^2}\right) \tag{7.1}$$

As described in Section 2, reconstruction is performed in the Fourier domain. This makes the choice of Gaussian RBFs convenient, as V can be computed efficiently in Fourier space[1]:

$$\hat{V}(\mathbf{k}) = \sum_i a_i e^{-2\pi i \mu_\mathbf{i} \cdot \mathbf{k}} \exp\left(\frac{-\pi^2 \mathbf{k}^2 \sigma_i^2}{2}\right) \tag{7.2}$$

To impose physical constraints on the RBF model, we add a set of fixed harmonic bond terms $\mathcal{B}$ between consecutive backbone RBF centers. Each bond term $(i, j, k, l) \in \mathcal{B}$ is specified by a pair of RBF indices i, j and a bond strength k, which we set to 0.1. The bond length l is set to 3.8 Å, the distance between protein C_α backbone carbons.

We constrain the side-chain RBFs to be located close to their backbone RBF using one-sided harmonic restraints $\mathcal{C}$. These are similar to the bond terms, but only induce a loss when the distance between RBF centers exceeds the max length l. We set l to the maximum distance between a backbone C-alpha carbon and its side chain center of mass observed in the reference structure.

For homogeneous reconstruction, we fit $\{\mu_i\}, \sigma, a_0$ directly using stochastic gradient descent (SGD). The overall loss function, given a set of N images $\mathbf{X}$ with poses ϕ, bond terms $\mathcal{B}$ and side-chain constraints $\mathcal{C}$ is:

$$\begin{aligned}\mathcal{L}(\mu, \sigma, a_0|\mathbf{X}) = \frac{1}{N} \sum_{(x,\phi) \in X} ||\hat{X}(\phi|\mu, \sigma, a_0) - X||^2 &+ \sum_{(i,j,k,l) \in \mathcal{B}} k(||\mu_i - \mu_j|| - l)^2 \\ &+ \sum_{(i,j,k,l) \in \mathcal{C}} k \ \max\left(||\mu_i - \mu_j|| - l, 0\right)^2\end{aligned} \tag{7.3}$$

[1] Projections of Gaussian kernels can also be computed analytically in real space, obviating the need for the Fourier slice theorem altogether. This real space formulation allows for algorithms that scale better with the number of RBFs since the RBFs are localized in real space, allowing for spatial decomposition via gridding, KD-trees, etc.

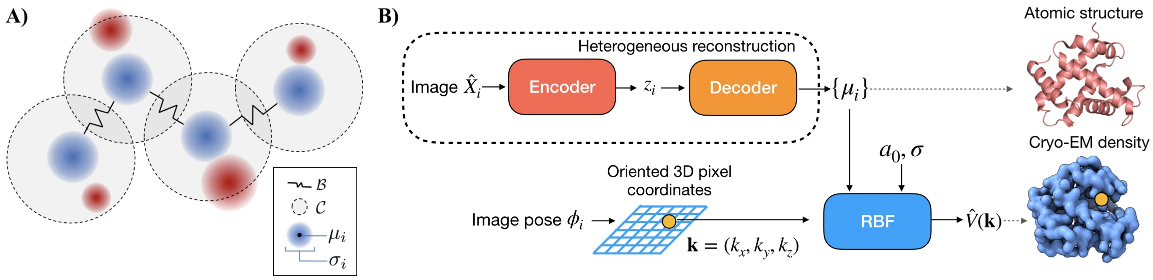


Figure 7-2: Cryofold model and architecture. **A)** The physics-inspired cryo-EM density model consists of set of Gaussian RBFs, two for each amino acid in the reference structure. One RBF represents the backbone (blue) and one represents the sidechain (red). Backbone RBFs are connected with harmonic bond terms $\mathcal{B}$, and sidechain RBFs are connected to their backbone with max-distance harmonic constraints $\mathcal{C}$. **B)** Architecture for heterogeneous reconstruction. We use a VAE to learn z-dependent offsets of RBF centers μ_i, given an unlabelled imaging dataset of image, pose pairs $(\hat{X}_i, \phi_i)$.

7.3.2 Heterogeneous reconstruction

For heterogeneous reconstruction, we learn a continuous latent variable model of conformational heterogeneity expressed through motion of RBF centers (Figure 7-2B). Unlike standard heterogeneous cryo-EM reconstruction algorithms that use an unconstrained volume representation (e.g. voxel arrays or positionally encoded MLPs), the RBF model constrains the effect of the latent degrees of freedom to motion of the underlying atomic structure.

An image encoder E with parameters θ_E predicts a latent $z \sim E(\hat{X}|\theta_E)$; A decoder network D with parameters θ_D predicts a z-dependent translation of the RBF centers $\mu_{het}(z) = \mu + D(z|\theta_D)$. We optimize $\mu, \sigma, a_0, \theta_D, \theta_E$ together end-to-end with SGD.

7.3.3 Local minima

A major shortcoming of this atomic reconstruction approach is the existence of many local minima of the loss function that do not approximate the true atomic coordinates. This is in contrast to voxel-based models, where SGD converges to the global minimum of the convex loss given the poses, and in contrast to large-scale generic neural models (e.g. [29, 28]) which typically do not suffer from problematic local minima in practice [5]. We address the local minima problem primarily by initializing RBF centers μ_i

from a reference structure that is a close approximation of the imaged structure (homogeneous) or some point in the distribution of structures (heterogeneous). Such reference structures are often available when studying a variant of a known structure, or heterogeneous protein dynamics.

7.4 Results

Here, we present results for reconstructing coarse-grained atomic structures from synthetic cryo-EM image data with Cryofold. We first explore the effect of reference structure initialization in homogeneous reconstruction, when initialized from either an approximate reference structure (fitting), the exact structure (cheating), or from randomly initialized locations (folding). We then turn to heterogeneous cryo-EM datasets, where we evaluate the ability of Cryofold to reconstruct continuous distributions of structures and their atomic coordinates. Finally, we assess choices in the design of the RBF model by ablating key components in the homogeneous setting.

Datasets. We generate homogeneous and heterogeneous cryo-EM datasets using a 141-residue atomic model (PDB 5NI1) as the ground truth structure. To generate homogeneous data, we simulate 50,000 noisy projection images based on the cryo-EM image formation model. To model heterogeneity, we introduce a bond rotation in the backbone of 5ni1 to create a continuous 1D motion, and generate cryo-EM images sampling along the ground truth reaction coordinate. Further details on dataset generation are given in the Appendix.

Architecture and training. For all experiments, we train for 10 epochs using the Adam optimizer in minibatches of 8 images and a learning rate of 1e-4. We initialize σ and a_0 to 3.71 Å and 0.1, and perform one epoch of "warm-up" to refine these global parameters after which we reset the atomic coordinates to their initial values. For homogeneous reconstruction, we directly optimize all RBF parameters. For heterogeneous reconstruction, we use a VAE to predict the z-dependent offsets of the RBF centers from their reference values. Both the encoder and decoder are 3-layer MLPs of width 256 and residual connections, and a 1-dimensional latent variable. We

Initialization	**Initial RMSE** (Å)	C_α **RMSE** (Å)	**% within 3Å**	**Volume NMSE**
Exact 5NI1	0.00	0.843	99.29%	0.28
Approximate 5NI1	5.15	3.81	77.30%	0.32
5NI1 + Uniform[6 Å]	6.50	3.19	53.19%	0.35
Random	34.87	17.16	2.12%	0.43

Table 7.1: Comparison of different choices of initial atomic coordinates when training the model.

use ground-truth poses ϕ for training. In real applications, the poses would be inferred from traditional cryo-EM tools [19, 32, 9]. The model is implemented in PyTorch [16].

7.4.1 Homogeneous reconstruction

To explore the effect of reference structure initialization, we compare the reconstruction accuracy of Cryofold when initialized from the ground truth coordinates ("Exact 5NI1") and from three alternate starting configurations: 1) an approximate reference structure generated by evolving the system under a molecular dynamics simulation ("Approximate 5NI1"), 2) the ground truth structure with 6 Åuniform noise added to the ground truth coordinates of each C_α ("5NI1 + Uniform[6 Å]"), and 3) a random initialization of each C_α RBF in the 64 Å^3 region in the center of the box ("Random"). Side-chain RBFs are initialized to their corresponding C_α RBF centers.

We report the root-mean-square-error (RMSE) of model backbone coordinates to the C_α of the true structure, the percent of C_α backbone atoms predicted within 3 Å of the true structure, and the normalized mean-square-error (NMSE) of the reconstructed volume to the true structure (Table 7.1). We find that our atomic model is quite sensitive to initialization. While the model performs adequately when initialized at nearby reference structures, it makes some mistakes due to local minima, and performs much better when initialized with the ground truth coordinates. When initialized from random coordinates, SGD is unable to recover the ground truth atomic coordinates whatsoever. Instances of atomic local minima include "mismatched buttoning" of the amino acid backbone where multiple backbone RBFs are trapped in the same location, as well as incorrect tracing of the protein backbone through the volume (Figure 7-3).

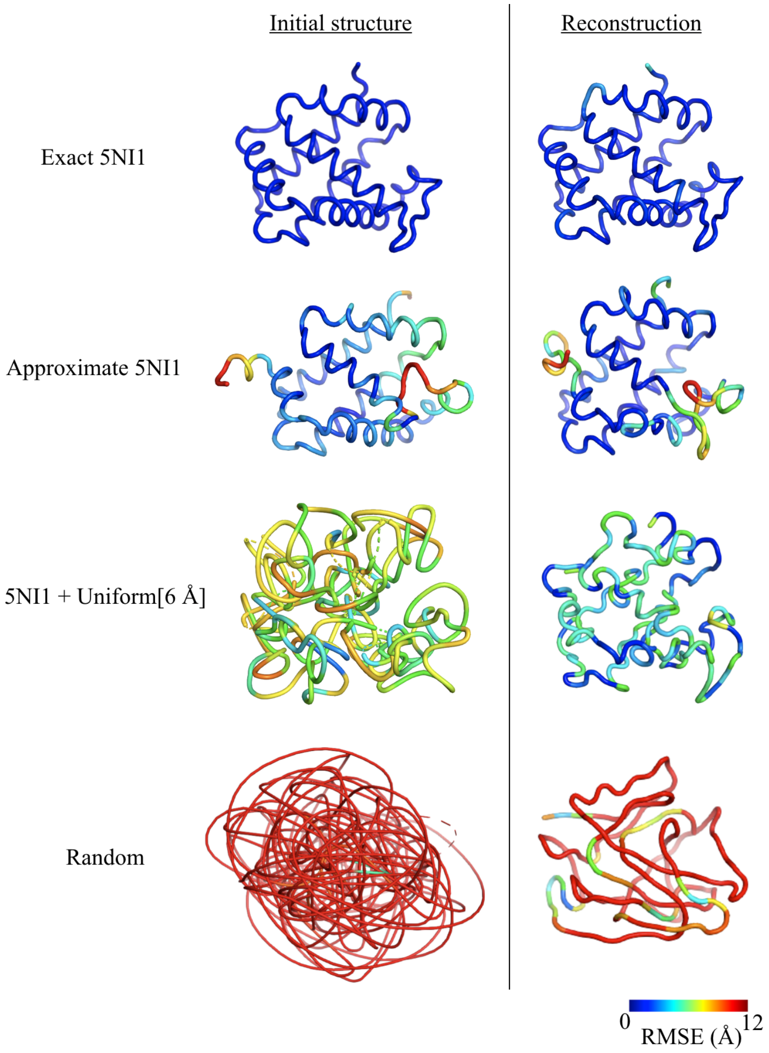


Figure 7-3: Reconstructed atomic structures with RBF centers initialized either 1) at the exact ground truth values of 5NI1, 2) at an approximate structure generated from evolution by molecular dynamics, 3) at 5NI1 coordinates randomly perturbed by Uniform[6 Å] noise, or 4) from random initial values. Structures are colored by C_α RMSE to the ground truth structure (top left).

7.4.2 Heterogeneous reconstruction

The main motivation behind proposing this model is to provide inductive bias from known atomic structures when reconstructing heterogeneous protein conformations. As opposed to the previous section where we reconstruct a single static structure, here we attempt to reconstruct a continuous manifold of structures from a heterogeneous cryo-EM dataset.

We consider a synthetic dataset of noisy projection images, where the underlying protein structure possesses a 1D continuous motion of the 5NI1 protein. A bond in the center of the protein is rotated, leading to a large-scale global conformational change (Figure 7-7). In our experiments, we initialize the RBF model to the structure at one end of the reaction coordinate, and train a heterogeneous RBF model with a 1-dimensional latent variable on the imaging dataset with ground truth poses.

The RBF model is able to correctly reconstruct the full distribution of 3D structures containing this large conformational change. The latent encodings of the images are well correlated with the true reaction coordinate (Figure 7-4, left), and the RBF atomic coordinates from traversing the latent space nearly exactly reconstruct the underlying protein motion (Figure 7-4, right).

As a baseline, we perform heterogeneous reconstruction with cryoDRGN [28], which learns an unconstrained neural representation of cryo-EM volumes. Similarly, we provide the ground truth poses and train a 1-D latent variable model with identical encoder architecture and training settings. While the latent space is well correlated with the ground truth motion similar to the RBF model, the reconstructed volumes from cryoDRGN contain noise and blurring artifacts in the mobile region whereas Cryofold volumes are regularized using structural priors from the underlying atomic model (Figure 7-4, bottom).

We also measure reconstruction accuracy across the reaction coordinate for four distributions of images across the reaction coordinate: a uniform distribution, two non-uniform distributions corresponding to an energy barrier of different heights between the end states; and two discrete clusters with no images in the middle (Figure

7-5). For each value of the reaction coordinate, we approximate the atomic coordinates using the median latent from the images at that coordinate, and measure its C_α RMSE with the ground truth. We see that when there is even a small probability mass across the reaction coordinate (Figure 7-5, top row), the atomic model learns the full distribution of conformations with high accuracy. However, if transition states are nearly or completely unobserved in the image distribution (Figure 7-5, bottom row), reconstruction accuracy is poor.

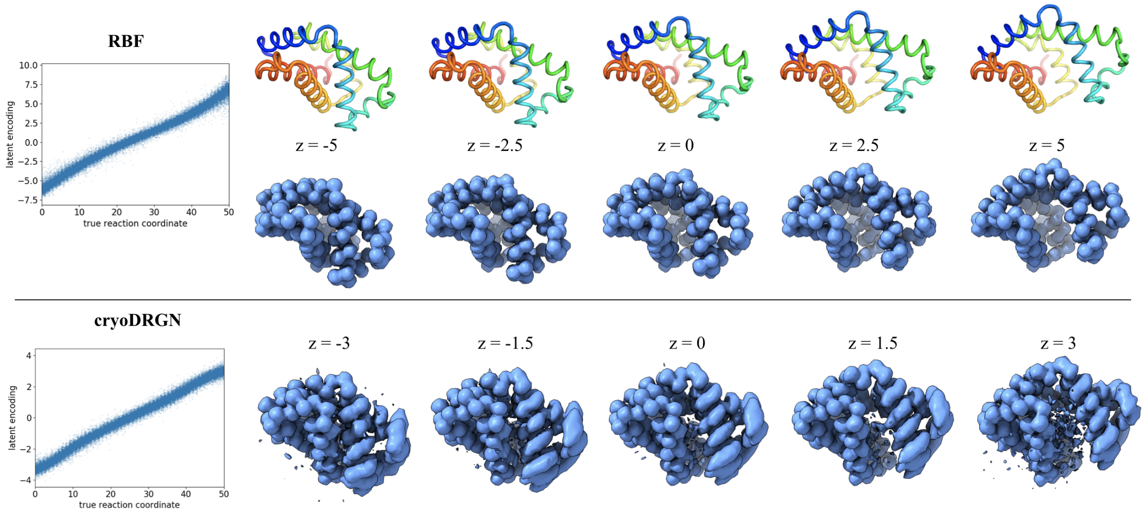


Figure 7-4: Heterogeneous reconstruction results of an unlabeled dataset containing a uniform distribution of images across the ground truth reaction coordinate. Predicted 1D latent encoding z plotted against the ground truth reaction coordinate (left), and reconstructed structures at the specified values of z. Cryofold directly reconstructs atomic coordinates (top). The unconstrained neural volume representation (cryoDRGN) contains noise and blurring of moving atoms in the reconstructed volumes (bottom) and does not produce atomic coordinates.

7.4.3 Model ablations

Using the homogeneous dataset and an initialization from the approximate 5NI1 structure, we explore various choices in design of the RBF model by ablating the side chain RBF, bond constraints, and the cryo-EM supervision (Table 7.2). We find that removing the sidechain RBFs and modeling each amino acid as a single RBF slightly degrades quality, whereas removing the internal bond terms leads to a dramatic degradation. Atomic accuracy is substantially worse but not completely

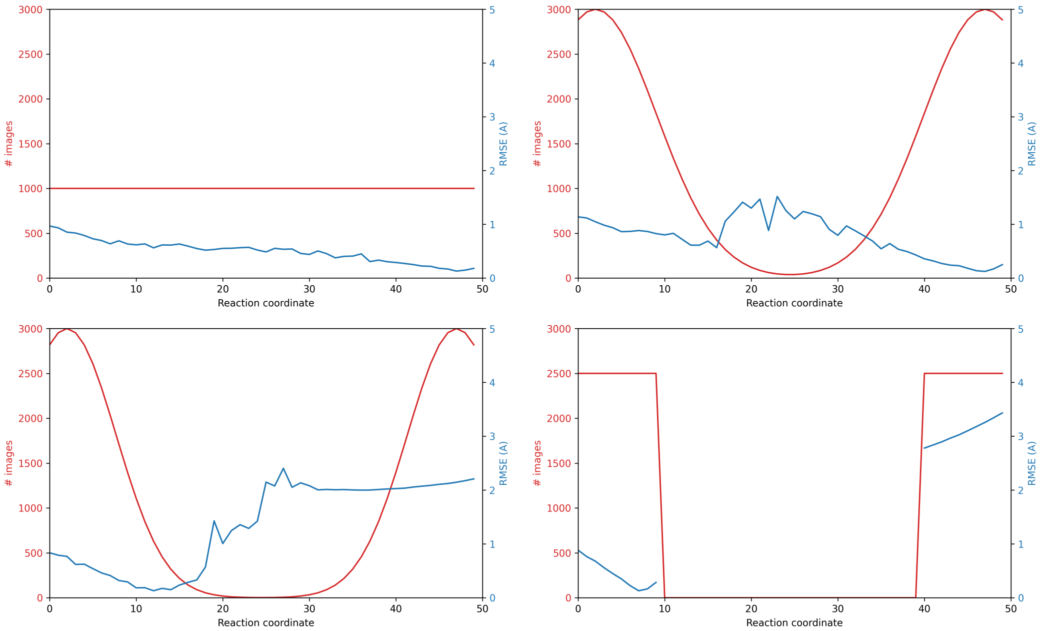


Figure 7-5: C_α RMSE (blue) for heterogeneous reconstruction of a large synthetic conformational change of 5ni1, with different distributions of images (red) across the reaction coordinate. The reference structure used for initialization is the ground truth atomic structure at reaction coordinate 0.

degraded even when ignoring the cryo-EM images due to the initialization; however, as expected there is low overlap with the true volume in this case. We expect the volume NMSE to be higher than unconstrained approaches (e.g. cryoDRGN), as the coarse-grained RBF model does not exactly match the underlying all-atom generative model even with correct coordinates (Figure 7-8).

Model	C_α **RMSE (Å)**	**% within 3Å**	**Volume NMSE**
Full Model	**3.81**	**77.30%**	**0.32**
No Sidechain RBFs	4.10	70.21%	0.49
No Bonds	10.42	37.59%	0.36
No Cryo-EM Loss	5.15	74.47%	1.82
CryoDRGN	N/A	N/A	0.08

Table 7.2: Ablations of model components. We remove the sidechain RBFs, the bond terms between RBFs, and the cryo-EM reconstruction loss; each degrades the quality of reconstruction.

7.5 Discussion

This work presents Cryofold, a simple framework for incorporating prior information provided from an atomic model into cryo-EM reconstruction. By constraining the generative model to act on a (coarse-grained) atomic model, we constrain the model output to a submanifold of coordinate positions, which both provides strong regularization to reduce artifacts seen in unconstrained methods and directly yields interpretable atomic coordinates. Our experiments suggest that such models are a promising direction for incorporating structural priors into cryo-EM reconstruction, especially for heterogeneous structures. However the experimental validation is preliminary: we worked entirely with *synthetic* cryo-EM datasets of a *small* protein using *exact poses*. Follow-up work is required to understand how these techniques behave with real cryo-EM images and volumes, using realistic protein complexes of interest, and using approximate poses generated by existing tools. For larger protein complexes, follow-up work will investigate whether a coarser model granularity may be more appropriate, especially when modeling large-scale heterogeneous dynamics.

Results on homogeneous datasets suggest that local minima in optimization space are a major problem for this class of methods if coordinates are not initialized near the ground truth structure. These local minima problems could potentially be ameliorated with a combination of improved modeling of structural priors such as additional bonded terms and steric interactions, and optimization methods such as Hamiltonian dynamics or Markov Chain Monte Carlo that can escape local minima. There is a rich literature on these topics in the domain of molecular dynamics simulation which could carry over to cryo-EM reconstruction [2, 27].

However, results on heterogeneous datasets suggest that we can correctly learn distributions with large continuous conformational changes when initialized from some structure in the distribution. Even structural changes that are too large to be modeled correctly when only the endpoint structures are observed (as in homogeneous reconstructions) were successfully reconstructed from heterogeneous datasets. Additional work is required to more fully characterize exactly when neural models of heterogeneous structures converge to the correct distribution.

It should be noted, however, that prior information *biases* the model output. In particular, this approach will not be able to model other sources of image heterogeneity that can not be expressed as motion of the RBFs (e.g. compositional variation) whose presence are not known *a priori*. Importantly, the separation of atomic modelling from volume reconstruction in existing cryo-EM pipelines provides a crucial mechanism for validation of the volume reconstruction accuracy. This and related approaches that incorporate atomic priors should be carefully employed only in settings where prior beliefs are strongly held, and we believe validation of these methods is an open problem. Nevertheless, future extensions of this method that can flexibly integrate biophysical simulation, learning on existing structures, and experimental single particle cryo-EM data is a promising direction for improved protein structure determination.

7.6 Supplemental details

7.6.1 Dataset generation

Synthetic cryo-EM datasets were generated from an atomic model of PDB 5NI1 according to the image formation model as follows: starting from the deposited atomic model of 5NI1, the 141-residue A-chain subunit of the complex was extracted. A cryo-EM volume was generated with the ‘molmap‘ command in Chimera [17] at 3 Å resolution and grid spacing of 1 Å. The volume was zero-padded to a cubic dimension of 128^3. 50,000 projection images were generated at random orientations of the volume uniformly from SO(3). Images were then translated in-plane by t uniformly sampled from $[-10, 10]^2$ pixels. We omit the application of the CTF for simplicity in this synthetic dataset. Gaussian noise was added leading to a signal to noise ratio (SNR) of 0.1, where we define the whole image as the signal (Figure 7-6). As real cryo-EM datasets have variable resolution and thus resolvability of the atomic structure, we investigate the effect of volume resolution and the included atoms in generating the dataset's ground truth volume/images (Table 7.3). We find that our atomic modeling is quite robust to the resolution and which atoms we model in the synthetic data, which suggests that it may transfer well to real cryo-EM images.

To generate the “Approximate 5NI1” reference structure, we evolve a coarse-grained C_α-only model of the 5NI1 A chain under a short Hamiltonian Monte Carlo simulation near the system's melting temperature. We use the implementation from [31], and run 10k Monte Carlo moves where each move consists of 2.25 ps of molecular dynamics at 320 K. The final frame of the trajectory is used as the approximate reference structure.

To create synthetic heterogeneous datasets, a dihedral angle in the backbone of the atomic model was rotated through 0.25 radians with a structure generated every 0.05 radians along the motion (Figure 7-7). The resulting 50 structures are used as the ground truth structures for approximating the continuous motion. For simplicity of heterogeneous dataset generation, we use the 5NI1 atomic models with C_α backbone atoms only. We generate multiple datasets of 50k total images following the homogeneous dataset image formation process, each with a different distribution of images along the reaction coordinate as shown in Figure 7-5.

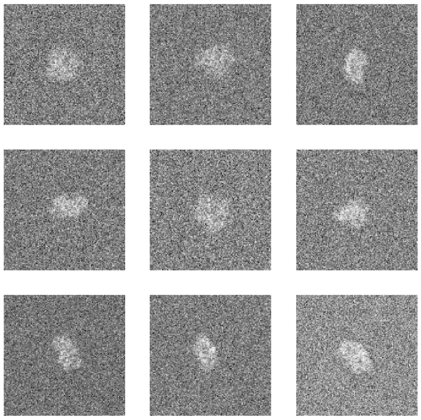


Figure 7-6: Example projection images.

Modeled Atoms	**Resolution**	C_α **RMSE (Å)**	**% within 3Å**
Cα	3Å	3.21	80.8%
Cα & Cβ	3Å	3.29	77.3%
All Atoms	3Å	3.75	78.0%
Cα	5Å	3.20	80.9%
Cα & Cβ	5Å	3.87	73.1%
All Atoms	5Å	3.80	74.5%
Cα	8Å	3.80	74.5%
Cα & Cβ	8Å	3.87	73.8%
All Atoms	8Å	4.20	78.7%

Table 7.3: Homogeneous reconstruction model accuracy for different settings of generating the dataset of cryo-EM images (resolution and which atoms are included). The model performs similarly across these choices. For other homogeneous experiments, we use all atoms at 3Å. For heterogeneous experiments, we use C_α at 3Å.

7.6.2 cryoDRGN baseline

For homogeneous reconstruction, we train a cryoDRGN positionally-encoded 3-layer MLP of width 256 for 20 epochs. For heterogeneous reconstruction, both the encoder and decoder networks are 3-layer MLPs of width 256, and are trained for 20 epochs.

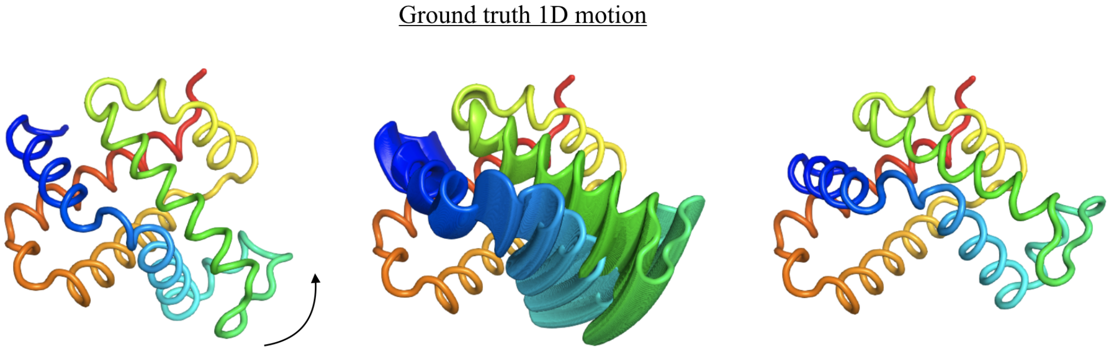


Figure 7-7: Ground truth structures of the heterogeneous datasets simulating a 1D continuous motion transitioning from left (5NI1) to right. All generated structures are shown in the center.

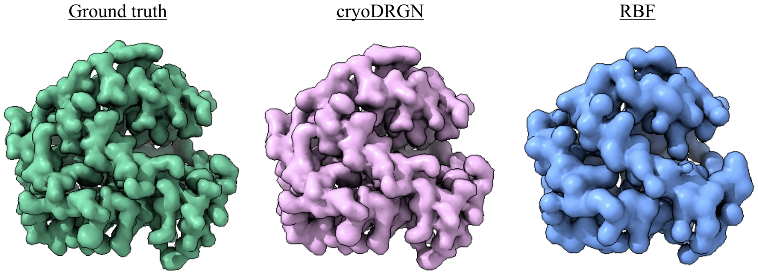


Figure 7-8: Ground truth and reconstructed volumes with a neural network representation of density (cyroDRGN) and with our RBF model initialized from exact coordinates. The RBF volume reconstruction is somewhat worse than that of cryoDRGN because even with exact coordinates, it cannot match the all-atom generative model.

Bibliography

[1] Ivet Bahar and A J Rader. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, 15(5):586–592, October 2005.

[2] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[3] Massimiliano Bonomi, Riccardo Pellarin, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. *Biophysical journal*, 114(7):1604–1613, 2018.

[4] Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2):198–217, 1956.

[5] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. 2015.

[6] Pilar Cossio and Gerhard Hummer. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of structural biology*, 184(3):427–437, 2013.

[7] Pilar Cossio and Gerhard Hummer. Likelihood-based structural analysis of electron microscopy images. *Current opinion in structural biology*, 49:162–168, 2018.

[8] Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods*, 100:61–67, May 2016.

[9] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cisTEM, user-friendly software for single-particle image processing. *eLife*, 7:e14874, mar 2018.

[10] Mohamad Harastani, Carlos Oscar S Sorzano, and Slavica Jonić. Hybrid electron microscopy normal mode analysis with scipion. *Protein Science*, 29(1):223–236, 2020.

[11] Konrad Hinsen, Nathalie Reuter, Jorge Navaza, David L Stokes, and Jean-Jacques Lacapère. Normal mode-based fitting of atomic structure into electron density

maps: application to sarcoplasmic reticulum ca-atpase. *Biophysical journal*, 88(2):818–827, 2005.

[12] Qiyu Jin, Carlos Oscar S Sorzano, José Miguel De La Rosa-Trevín, José Román Bilbao-Castro, Rafael Núñez-Ramírez, Oscar Llorca, Florence Tama, and Slavica Jonić. Iterative elastic 3d-to-2d alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure*, 22(3):496–506, 2014.

[13] Dari Kimanius, Gustav Zickert, Takanori Nakane, Jonas Adler, Sebastian Lunz, C-B Schönlieb, Ozan Öktem, and Sjors HW Scheres. Exploiting prior knowledge about biological macromolecules in cryo-em structure determination. *IUCrJ*, 8(1), 2021.

[14] Lyumkis, Dmitry, Brilot, Axel F, Theobald, Douglas L, and Grigorieff, Nikolaus. Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of Structural Biology*, 183(3):377–388, September 2013.

[15] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors Hw Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7:e36861, June 2018.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

[17] 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. *Journal of Computational Chemistry*, 25(13):1605–1612, oct 2004.

[18] 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, 2021.

[19] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, March 2017.

[20] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch, S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes, 2021.

[21] R Sánchez-García, J Gomez-Blanco, A Cuervo, J M Carazo, COS Sorzano, and J Vargas. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *bioRxiv*, 2020.

[22] Sjors H W Scheres. A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2):406–418, January 2012.

[23] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of Molecular Biology*, 348(1):139–149, 2005.

[24] Sandra Schilbach, Merle Hantsche, Dmitry Tegunov, Christian Dienemann, Cristoph Wigge, Henning Urlaub, and Patrick Cramer. Structures of transcription pre-initiation complex with tfiih and mediator. *Nature*, 551(7679):204–209, 2017.

[25] Amit Singer and Fred J. Sigworth. Computational Methods for Single-Particle Electron Cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, jul 2020.

[26] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 2020.

[27] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of chemical physics*, 151(7):070902, 2019.

[28] 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, 2021.

[29] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR*, 2020.

[30] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-em reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

[31] Ellen D Zhong and Michael R Shirts. Thermodynamics of Coupled Protein Adsorption and Stability Using Hybrid Monte Carlo Simulations. *Langmuir*, 30(17):4952–4961, 2014.

[32] Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7, nov 2018.

Chapter 8

Conclusions

One of the central tenets of structural biology is that the visualization of the 3D atomic structure of biomolecular complexes will yield direct insight into the mechanisms by which these molecular machines function. In pursuit of this goal, single particle cryo-EM has emerged as a mature structural biology technique uniquely poised to resolve not only static structures at atomic resolution, but also the conformational ensembles of massive protein complexes that carry out essential biological processes.

This thesis presents new algorithms that address the computational bottlenecks at the frontier of structure determination of these dynamic macromolecular machines. The main contribution of this thesis is a deep learning system, cryoDRGN, for heterogeneous cryo-EM reconstruction that both innovates on neural modeling techniques and addresses a major open challenge in the field of reconstructing continuous forms of heterogeneity from cryo-EM image data.

We first described the neural network design principles in the cryoDRGN method (Chapter 3). In developing cryoDRGN, we framed heterogeneous cryo-EM reconstruction as unsupervised learning of a deep generative model of 3D density maps from 2D cryo-EM imaging data. The resulting model of cryo-EM structure is (in principle) able to represent any distribution of structures that can be approximated by a deep neural network, a broad class of function approximators for continuous, nonlinear functions [3]. Underpinning this model is a neural representation of 3D structure $V_\theta : \mathbb{R}^3 \to \mathbb{R}$ parameterized as an MLP with sinusoidal featurization of input coordinates. This representation has since shown broad applicability in other domains of computer vision, and future work could explore adapting extensions of this technique in computer vision to cryo-EM.

A guiding motivation in this body of work has been to develop, validate, and apply cryoDRGN to solve new structures and visualize continuous dynamics from real cryo-EM data (Chapter 4). Extended to real experimental datasets, cryoDRGN has enabled 3D reconstruction of both compositionally and conformationally heterogeneous targets, helping to realize the potential of cryo-EM for uncovering complex conformational landscapes of biomolecular complexes. As cryoDRGN has a relatively expressive model for heterogeneity compared to existing and contemporary approaches, it has been used to discover new structures in previously published datasets [8], reconstruct dynamic motions in experimental collaborations [1], and used as a tool by external research groups through our open-source software release (`cryodrgn.csail.mit.edu`).

It is worthwhile to consider the limitations of this method, which lead naturally to directions for further improvement. Chapters 5 through 7 describe three such directions for extending the cryoDRGN approach: *ab initio* reconstruction, the downstream interpretation of generative models, and exploiting prior information from the underlying atomic structure.

In the application of cryoDRGN to real data, high quality reconstructions were achieved by using poses assigned from an upstream homogeneous reconstruction. Chapter 5 describes cryoDRGN2, which revisits the problem of *ab initio* reconstruction, i.e. reconstruction without any prior information on image poses or the volume representation. Techniques are proposed that dramatically speed up and increase the accuracy of pose estimation when optimizing neural models of volumes, enabling state-of-the-art cryo-EM reconstruction of protein structures from real datasets. While there are currently many exciting new ideas and exploratory paradigms for pose inference in cryo-EM reconstruction (e.g. gradient-based learning [6, 5] or distribution matching with adversarial learning [2]), at the time of this writing, all existing tools for heterogeneous reconstruction use previously estimated, fixed poses[1]. Thus, cryoDRGN2

[1] aside from 3D classification

provides new capabilities for heterogeneous reconstruction, potentially broadening the scope of single particle cryo-EM to new classes of heterogeneous protein and biomolecular complexes. Future work will explore how well the method transfers from the benchmark systems shown in Chapter 5 to these more challenging heterogeneous datasets.

In applying cryoDRGN across a variety of systems of diverse sizes and sources of heterogeneity, we discovered that the interpretation of deep generative models is itself a challenge. To facilitate interpretation, in Chapter 6, we describe a set of analysis tools that aim to comprehensively characterize the manifold of structures from a trained cryoDRGN model. We take a data-driven approach and characterize both *discrete* conformational states and a *continuous* conformational landscape, where the interpretation should be driven by the hypotheses of the practitioner. This flexibility in structural interpretation emphasizes the new exploratory nature of structural biology that is made possible by cryo-EM.

Finally, the 3D reconstruction stage of cryo-EM structure determination typically ends with a single or an ensemble of 3D density volume(s), after which an atomic model is fit into the volume for structural interpretation. Chapter 7 explores the possibility of including the underlying atomic model of the imaged protein directly in the reconstruction process. Two benefits of an atomic model parameterization include regularizing the reconstruction, especially for continuous dynamics, and automatically fitting an ensemble of atomic models, which facilitates interpretation. However, it is clear that this model can easily frustrate the optimization landscape, which is a challenge that future work exploring this paradigm must address.

8.1 Future Directions

8.1.1 Structured deep generative models

While cryoDRGN has introduced the power of deep learning-based function approximation to heterogeneous reconstruction, a nascent field remains to be shaped in machine learning for single particle cryo-EM image analysis. Future work could entail developing methods with more structured generative processes specifically tailored to known sources of heterogeneity (as opposed to cryoDRGN's generic latent variable model). For example, a model that explicitly accounts for out of distribution "junk" (common in real datasets due to experimental contamination) or tailored around known sources of compositional vs. conformational heterogeneity would both require new approaches for representation learning and be extremely impactful in real application settings. As the field grows, methods developers would also benefit from community-wide resources, for example, establishing benchmarks for heterogeneous reconstruction. These benchmarks will need to be reflective of the diverse tasks and goals in heterogeneous reconstruction.

8.1.2 Atomic resolution structure determination

Traditionally, resolving an atomic resolution cryo-EM structure requires obtaining extraordinarily large datasets at significant cost. I believe next generation reconstruction methods will break the firewall between the volumetric density model produced by reconstruction algorithms and the underlying atomic model of the protein structure. Methods that incorporate prior information based on an atomic model are currently in their infancy, but hold promise especially for resolving protein dynamics at high resolution. Whether these methods include priors based on our biophysical knowledge of proteins (e.g. physics-based energy functions) or an existing corpus of protein structures (e.g. AlphaFold-like systems) remains to be explored. Furthermore, these methods will help transition the current paradigm in structural biology from confirmatory to exploratory, hypothesis-generating experiments, and thus necessitate alternative biophysical and biochemical approaches for validation.

8.1.3 Commoditizing cryo-EM

Cryo-EM has undergone transformative improvements in quality and speed, leading to dramatic growth in its adoption. However, processing cryo-EM datasets still requires considerable manual skill and time to determine a single structure or a small set of structures. General machine learning systems may play a crucial role in next generation cryo-EM data processing software suites and enable the automatic determination of all high-resolution conformations from a given dataset. This will be highly nontrivial and require methods that can integrate information across datasets, while remaining invariant to extrinsic sources of variability. Ultimately, the commoditization of cryo-EM experiments will be transformative for protein structure *prediction* methods, whose accuracy is limited for data-sparse problems such as predicting large protein complexes and protein conformations and dynamics that can be resolved by cryo-EM.

8.1.4 Cryo-ET and *in situ* visual proteomics

Cryo-electron tomography (cryo-ET) is an upcoming area with tremendous promise in visualizing *in-situ* structural complexes: instead of single particles suspended in vitreous ice, cryo-ET produces noisy images of thin slices from whole cells. Alone, this data is incomplete to resolve high resolution 3D structures. We will need systems that can reason over our existing understanding of structural biology components (i.e. high accuracy prediction systems), and integrate this information with cryo-ET imaging data at scale to be able to reconstruct the 3D landscape of whole cells. In the future, it may possible to automatically detect the presence and atomic resolution structure of all protein components from cryo-ET imaging data, simultaneously providing their conformations, interactions, and their subcellular localization, to achieve an unprecedented resolution into molecular biology. These methods have the potential to connect our bottom-up mechanistic view of biomolecular structures to a systems-level understanding of cellular function and physiology.

8.1.5 Future outlook

At the time of this writing, the field of structural biology, the study of proteins and other biomolecules through their 3D structure, is undergoing an enormous transformation. The advent of high accuracy deep learning systems for protein structure prediction (i.e.

AlphaFold) has unlocked the ability of the scientific community to broadly reason about 3D protein structure [4, 7]. Simultaneously, a revolution in the capabilities of cryo-EM has significantly accelerated the discovery of new structures of large biomolecular complexes that are beyond the scope of current structure prediction methods. I am optimistic that there will be major advances in both structural biology and in machine learning via methods development at this intersection: computational methods for structure prediction and structure determination have fundamental limitations in isolation that may be addressed by one another, which if successful, will enable radically new capabilities in the visualization of cellular and molecular biology.

Bibliography

[1] Miao Gui, Meisheng Ma, Erica Sze-Tu, Xiangli Wang, Fujiet Koh, Ellen D Zhong, Bonnie Berger, Joseph H Davis, Susan K Dutcher, Rui Zhang, et al. Structures of radial spokes and associated complexes important for ciliary motility. *Nature structural & molecular biology*, 2020.

[2] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. Cryo-GAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning. *bioRxiv*, 2020.

[3] K Hornik, M B Stinchcombe, and H White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2, 1989.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, July 2021.

[5] Youssef S G Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-end simultaneous learning of single-particle orientation and 3D map reconstruction from cryo-electron microscopy data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, October 2021.

[6] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W. Senior, John Jumper, Carl Doersch, S. M. Ali Eslami, Olaf Ronneberger, and Jonas Adler. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes, 2021.

[7] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A A Kohl, Anna

Potapenko, Andrew J Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, July 2021.

[8] 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, 2021.