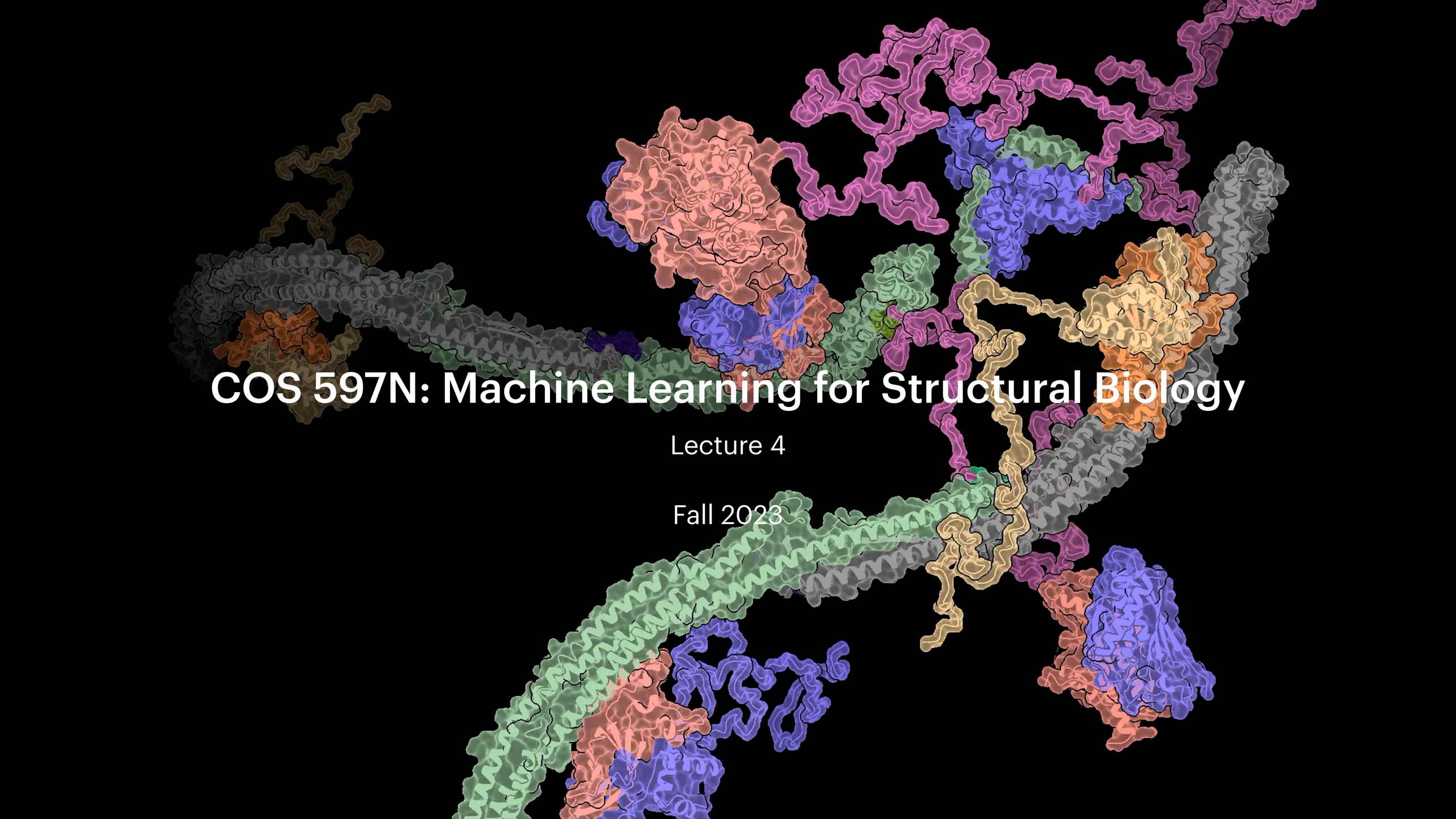


COS 597N: Machine Learning for Structural Biology

Lecture 4

Fall 2023

Course Logistics

- Optional student-only “precept”, Tuesdays at 4:30p in CS 401.
- Today:
  - Protein structure determination and cryo-EM reconstruction
- Next week 9/28: Protein language modeling — modified format!
  - Flash talks (groups of 1-2) + guest instructor (Adam Lerer)
  - (Very) short writing assignment
  - More details and paper sign up by the end of this week (end of day Friday): https://docs.google.com/spreadsheets/d/1WznSeVYRaCFk8cLzGpxKRhe5JM65TZLd-Ge29byuze4/edit#gid=0
- Oct 12: Protein design
- Oct 19 (fall break): No class + Project proposal due
  - Guidelines: https://docs.google.com/document/d/1bKyklL9v-N-Yac1tBQCNi8CGQHsN5wZ5BQM7ZDo4WN4/edit

This lecture

CryoDRGN ❄️🐉

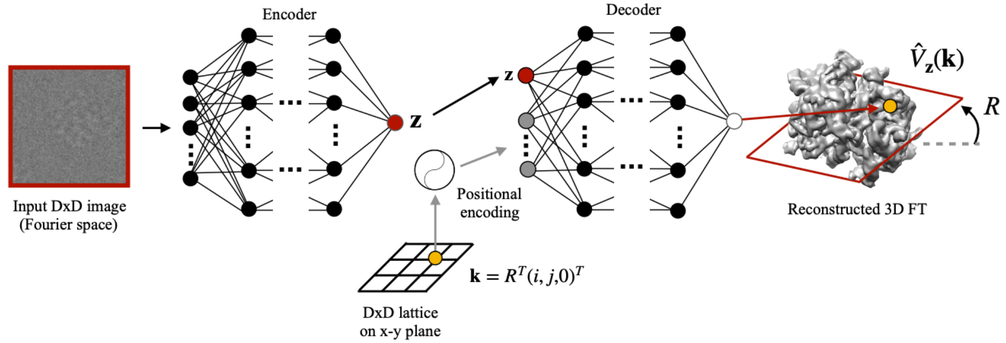
- (Recap): Who went to John Jumper's talk?
- CryoDRGN: Deep Reconstructing Generative Networks
  - Seminar
  - Figure by figure

- Questions:
  - What did you think of the papers?
  - What are the differences between conference vs. journal paper?
  - Who is familiar with NeRFs and implicit neural representations?
  - Any other thoughts/reflections?

[Submitted on 11 Sep 2019 (v1), last revised 15 Feb 2020 (this version, v3)]

Reconstructing continuous distributions of 3D protein structure from cryo-EM images

Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, Bonnie Berger



Article | Published: 04 February 2021

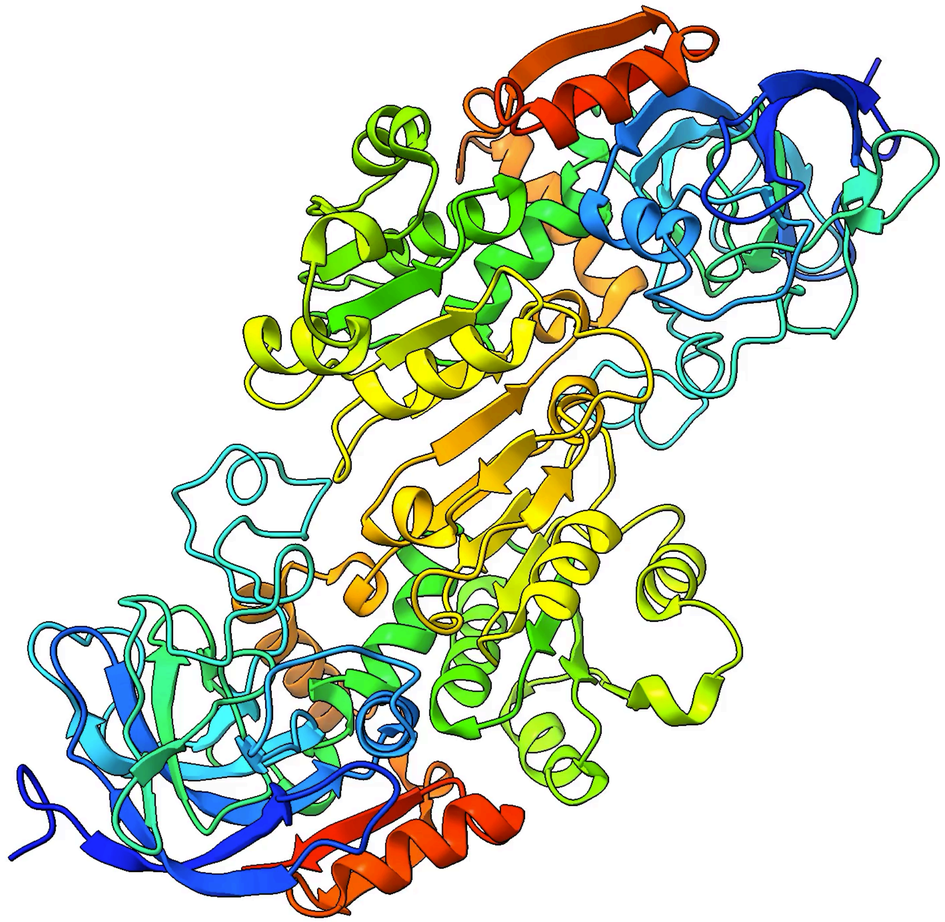
CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong, Tristan Bepler, Bonnie Berger ✉ & Joseph H. Davis ✉

Nature Methods **18**, 176–185 (2021) | Cite this article

Outline

- **Motivation:** Why do we care about protein structure?



Outline

- **Motivation:** Why do we care about protein structure?
- **Background:** Cryo-EM reconstruction & the heterogeneity problem

Continuously heterogeneous hyper-objects in cryo-EM and 3-D movies of many temporal dimensions

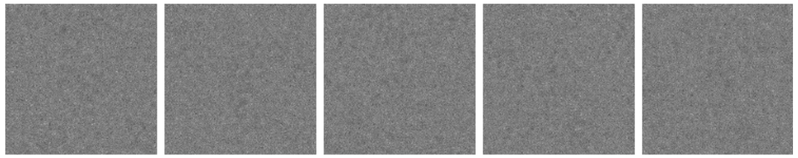
Roy R. Lederman* and Amit Singer†

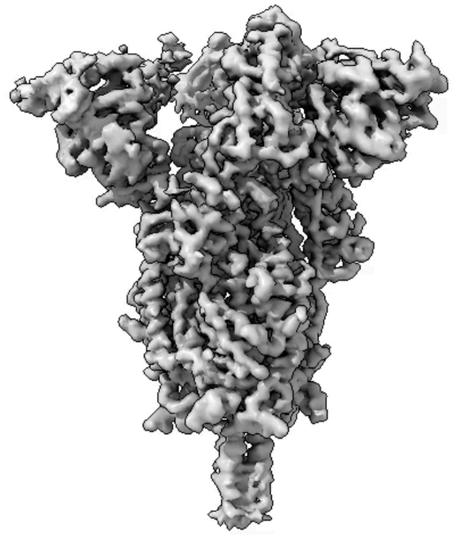
April 11, 2017

Abstract

Single particle cryo-electron microscopy (EM) is an increasingly popular method for determining the 3-D structure of macromolecules from noisy 2-D images of single macromolecules whose orientations and positions are random and unknown. One of the great opportunities in cryo-EM

The cryo-EM reconstruction task





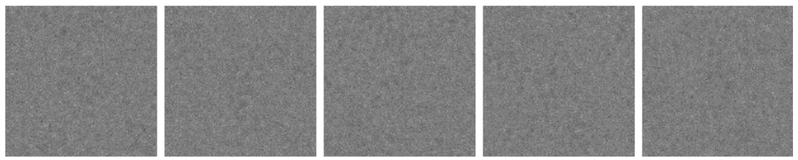
Walls et al, 2020

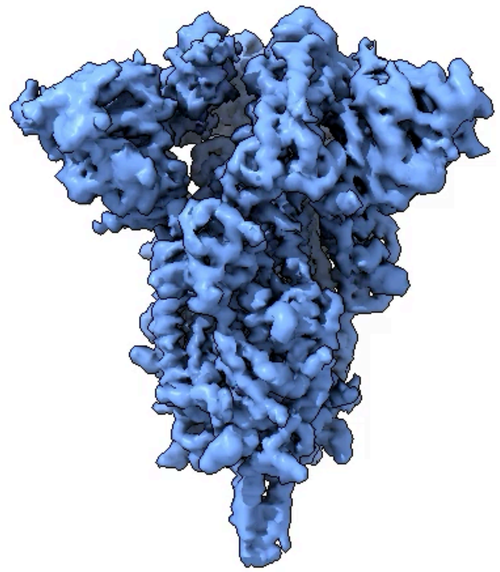
Cryo-EM structure of the SARS CoV-2 Spike protein

Outline

- **Motivation:** Why do we care about protein structure?
- **Background:** Cryo-EM reconstruction & the heterogeneity problem
- **CryoDRGN:** Neural 3D reconstruction of dynamic protein structure with cryoDRGN ❄️🐉

The cryo-EM reconstruction task





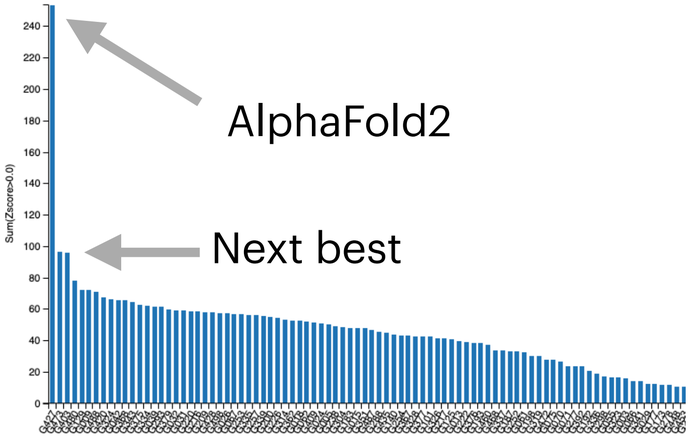
cryoDRGN trajectory of the SARS CoV-2 Spike protein

Zhong et al, Nature Methods 2021

Outline

- **Motivation:** Why do we care about protein structure?
- **Background:** Cryo-EM reconstruction & the heterogeneity problem
- **CryoDRGN:** Neural 3D reconstruction of dynamic protein structure with cryoDRGN ❄️🐉
- **Future directions:** Machine learning for structure determination at the proteome scale

CASP14 Results, Dec 2020

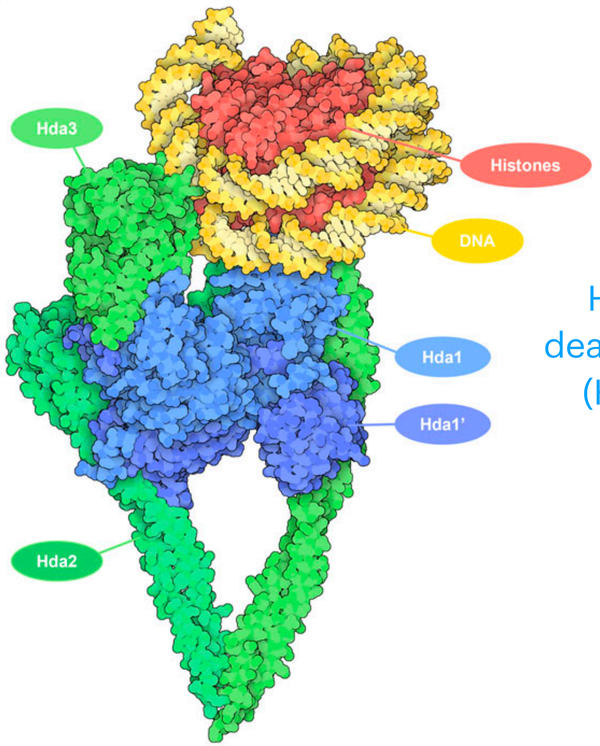


https://predictioncenter.org/casp14/zscores_final.cgi

https://pdb101.rcsb.org/sci-art/goodsell-gallery/escherichia-coli-bacterium

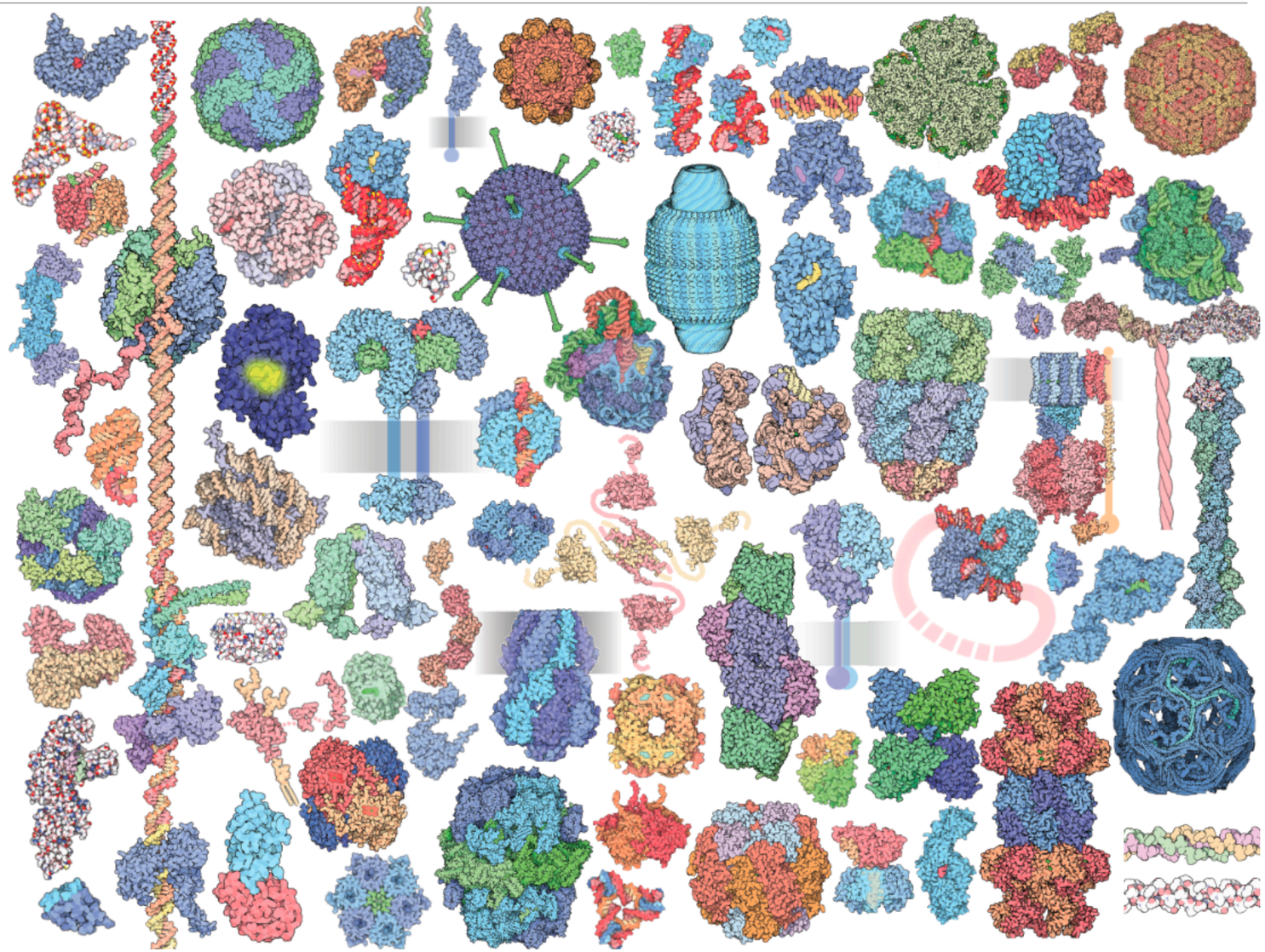
All essential biological processes are carried out by proteins and protein complexes

- Fundamental molecules of life
- Medicine and health
- Nanotech and biotech



Histone deacetylases (HDACs)

PDB-101 Molecule of the Month

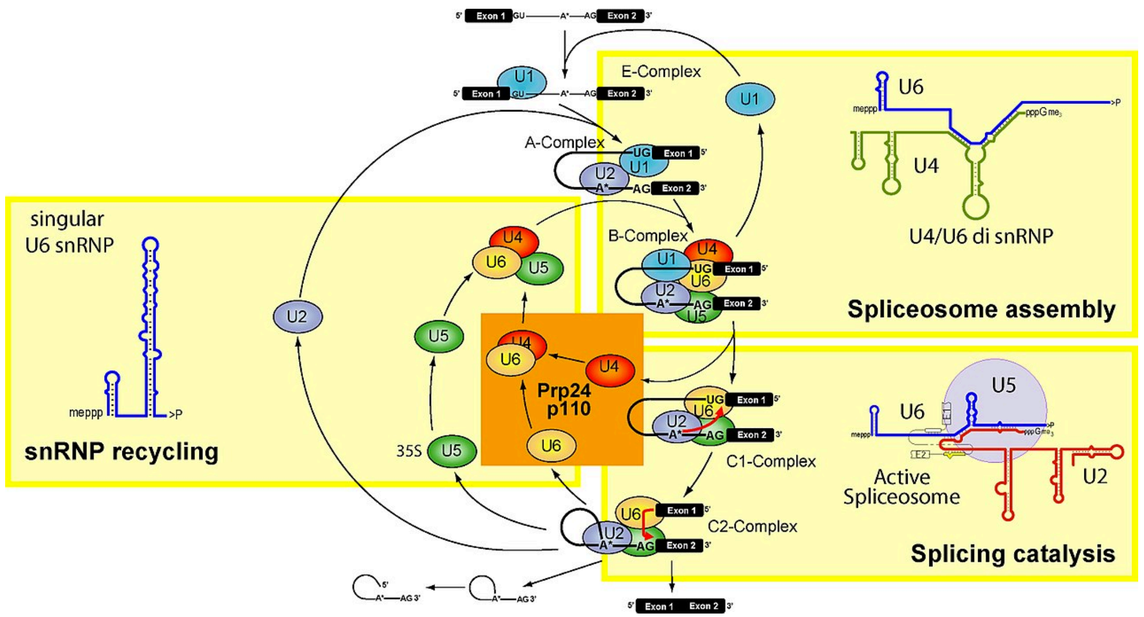


Goodsell et al. PLoS Biology 2015.

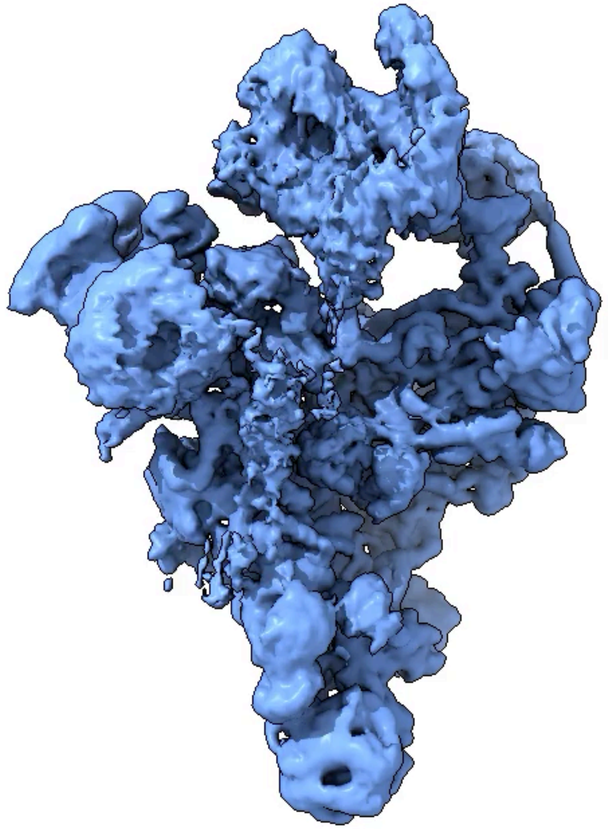
All essential biological processes are carried out by proteins and protein complexes

... which are dynamic macromolecular machines

Spliceosome splicing cycle



https://en.wikipedia.org/wiki/Spliceosome



cryoDRGN trajectory of the pre-catalytic spliceosome

Zhong et al, Nature Methods 2021

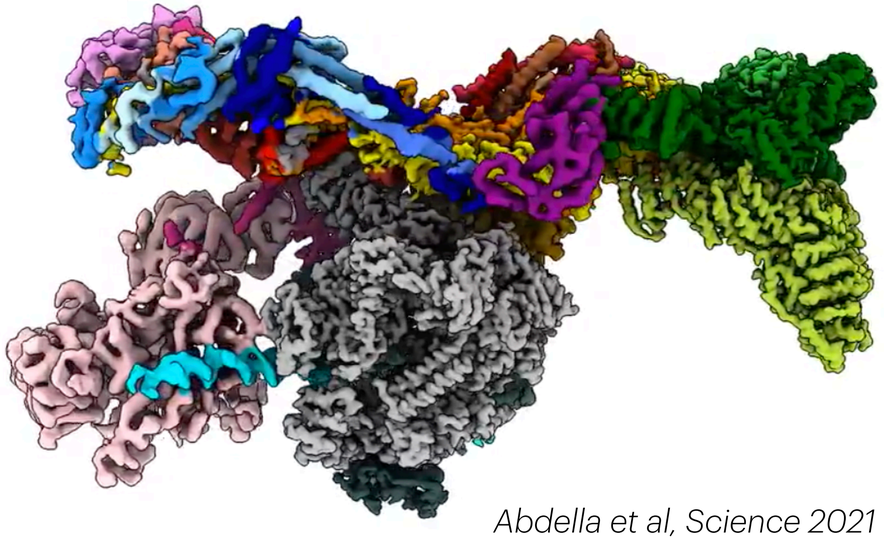
Techniques to study molecular motions are limited

- Nuclear magnetic resonance (NMR) spectroscopy
  - Small proteins (<100 AA in length)
- Electric field crystallography (EF-X), multi-temperature and XFEL crystallography
  - Requires sample crystallization
- Computational modeling
  - Molecular dynamics simulations
  - Hacking AlphaFold?
- **Cryo-electron microscopy (cryo-EM)**

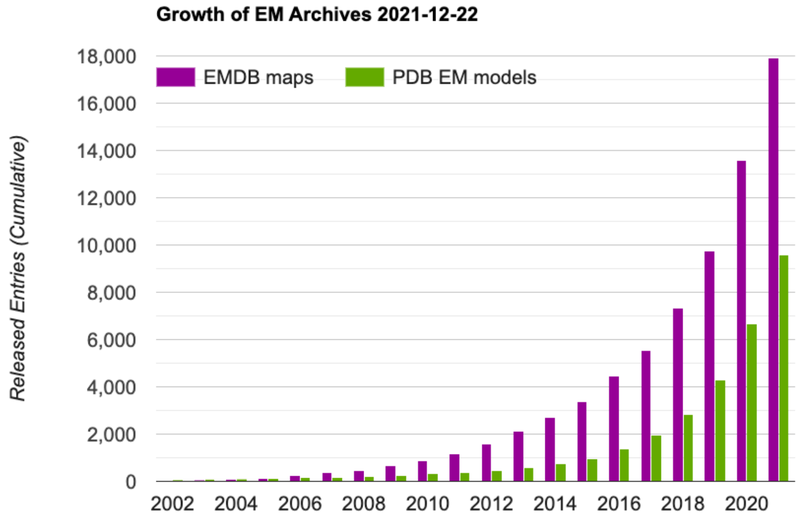
MD simulation of SARS CoV-2 Spike

The ongoing cryo-EM “resolution revolution”

- 2017 Nobel Prize in Chemistry
- Cryo-EM has opened up new areas of structural biology
- Recent hardware and software breakthroughs:
  - **Hardware:** direct electron detectors
  - **Software:** New reconstruction algorithms, GPU compute
  - **Faster:** Automation and democratization of cryo-EM imaging
- New computational challenges and opportunities

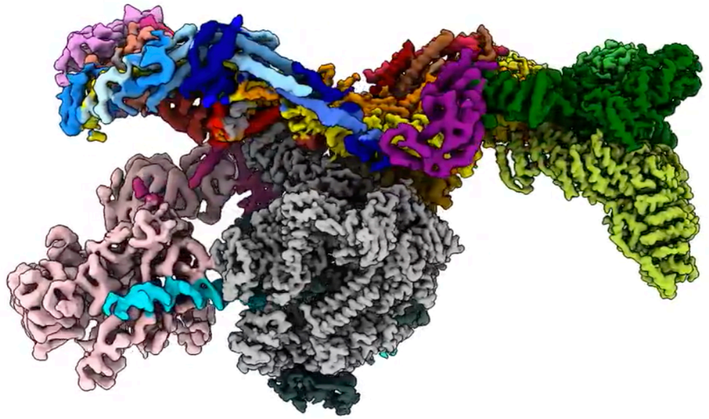


Abdella et al, Science 2021

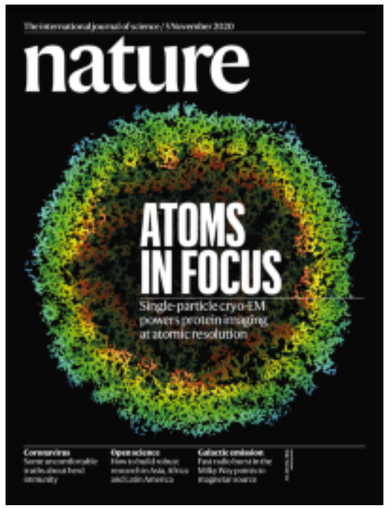


Frontiers of single particle cryo-EM

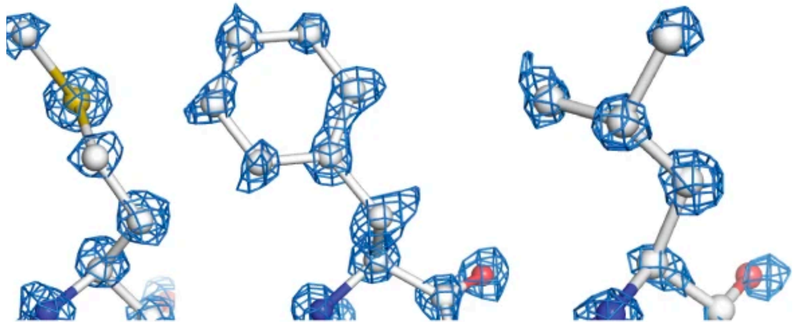
- Higher resolution structures
- Small proteins
- Time-resolved cryo-EM
- **Large, dynamic complexes**
  - (MDa scale, 10s-100s of proteins)



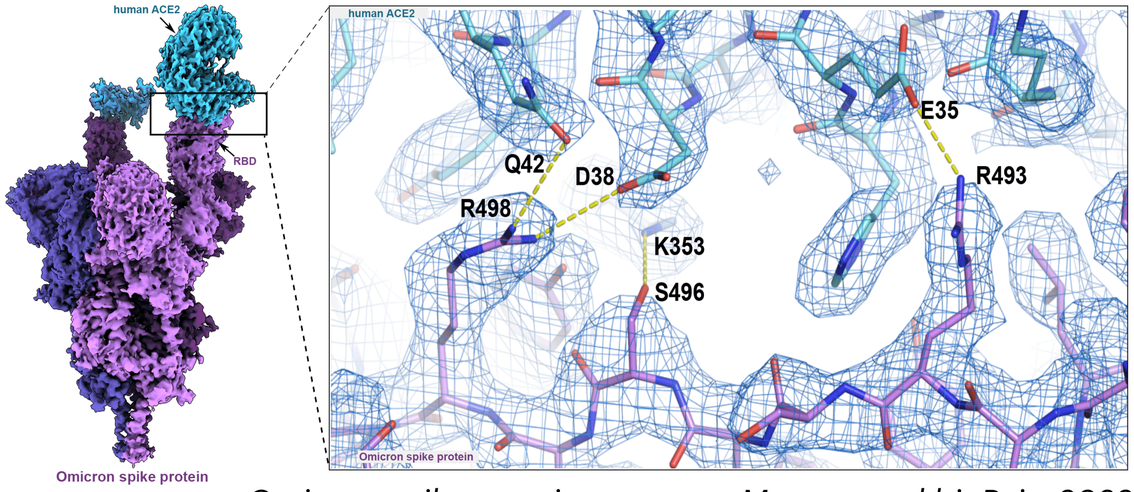
https://twitter.com/DanielHurdiss/status/1372659832780623872



November 2020



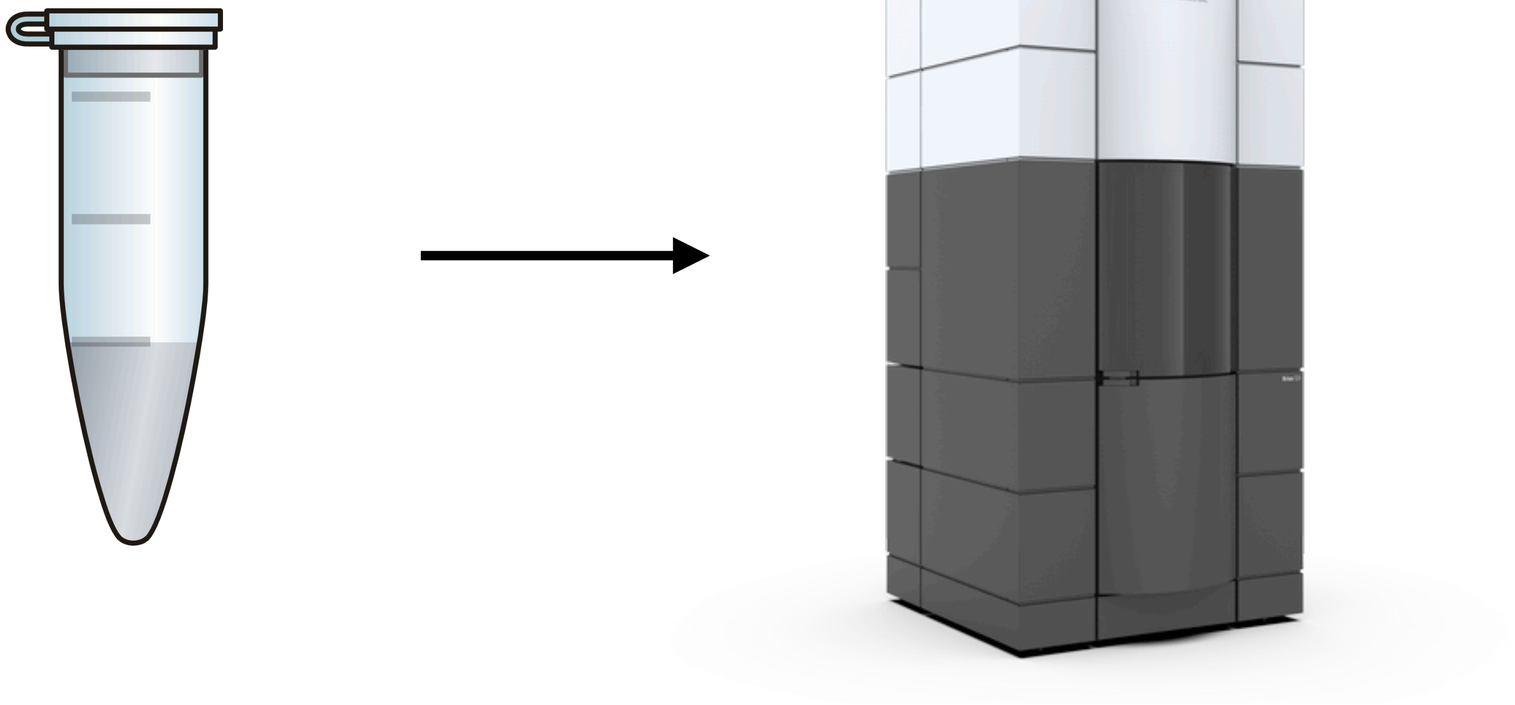
Nakane et al Nature 2020



Omicron spike protein structure. Mannar et al bioRxiv, 2022

The cryo-EM image processing pipeline: From micrograph to atomic coordinates

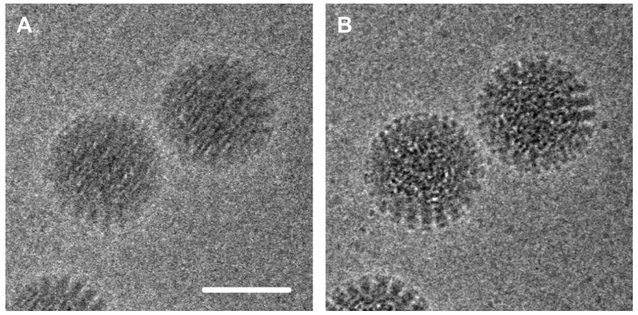
[Step 0) Sample preparation and imaging]



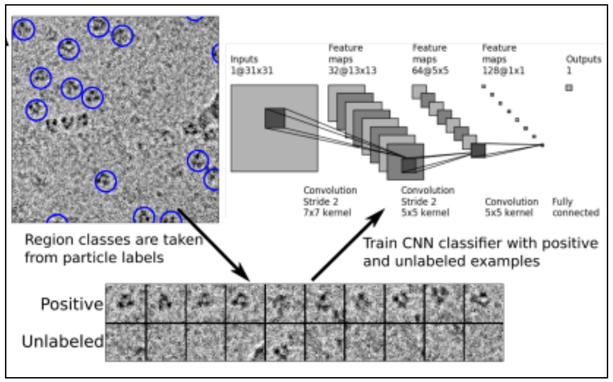
The cryo-EM image processing pipeline: From micrograph to atomic coordinates

[Step 0) Sample preparation and imaging]

1) Micrograph pre-processing



Grigorieff, 2013

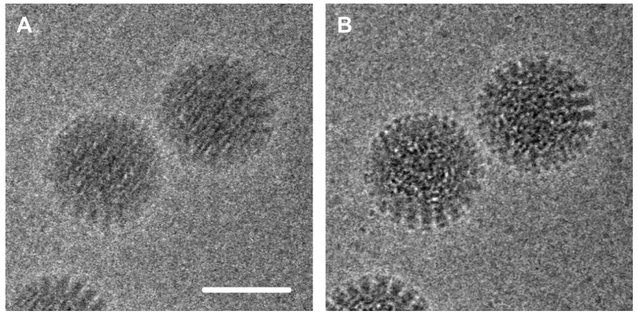


Topaz, Bepler et al, 2019

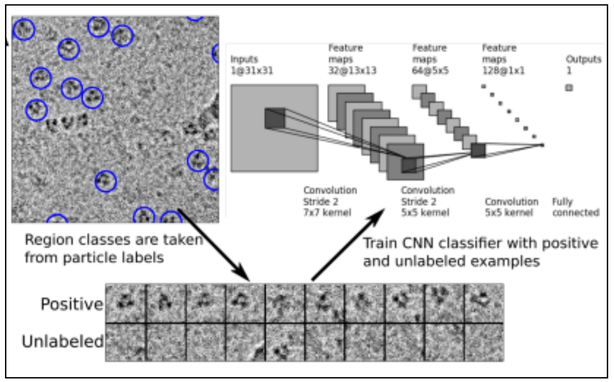
The cryo-EM image processing pipeline: From micrograph to atomic coordinates

[Step 0) Sample preparation and imaging]

1) Micrograph pre-processing



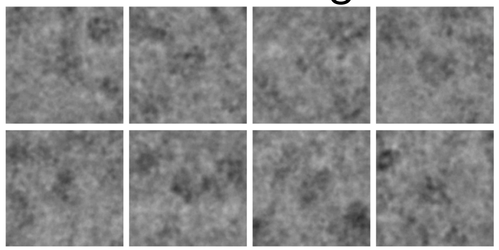
Grigorieff, 2013

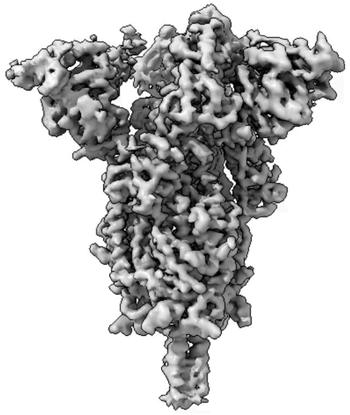


Topaz, Bepler et al, 2019

2) 2D to 3D reconstruction

10^4-10^7 images



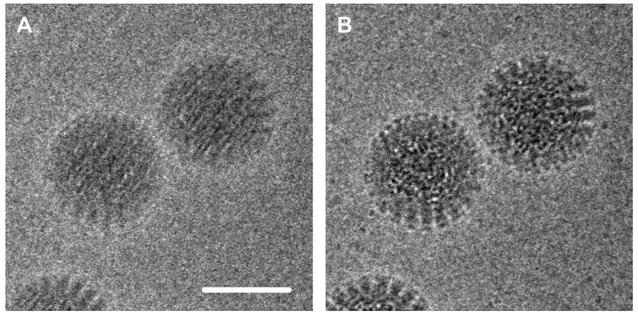


Walls et al, 2020

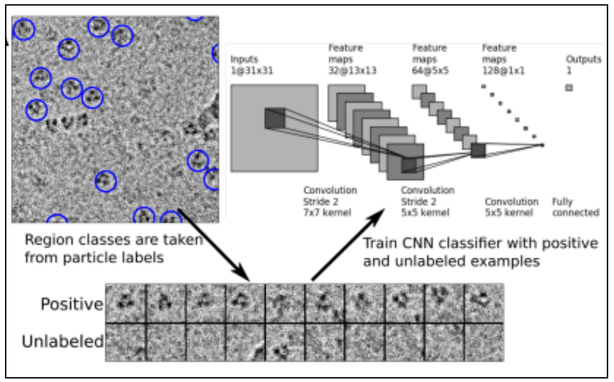
The cryo-EM image processing pipeline: From micrograph to atomic coordinates

[Step 0) Sample preparation and imaging]

1) Micrograph pre-processing



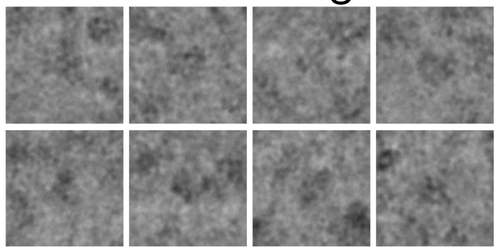
Grigorieff, 2013

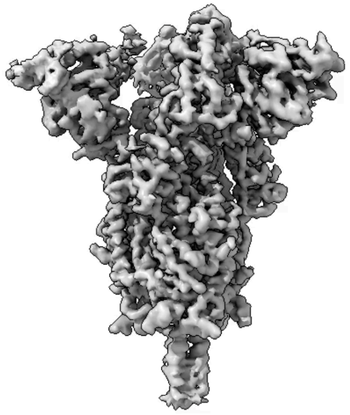


Topaz, Bepler et al, 2019

2) 2D to 3D reconstruction

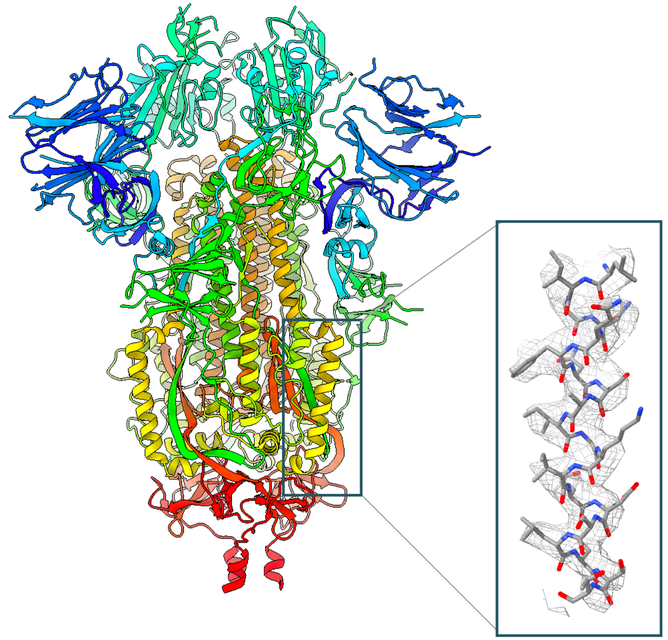
10^4-10^7 images





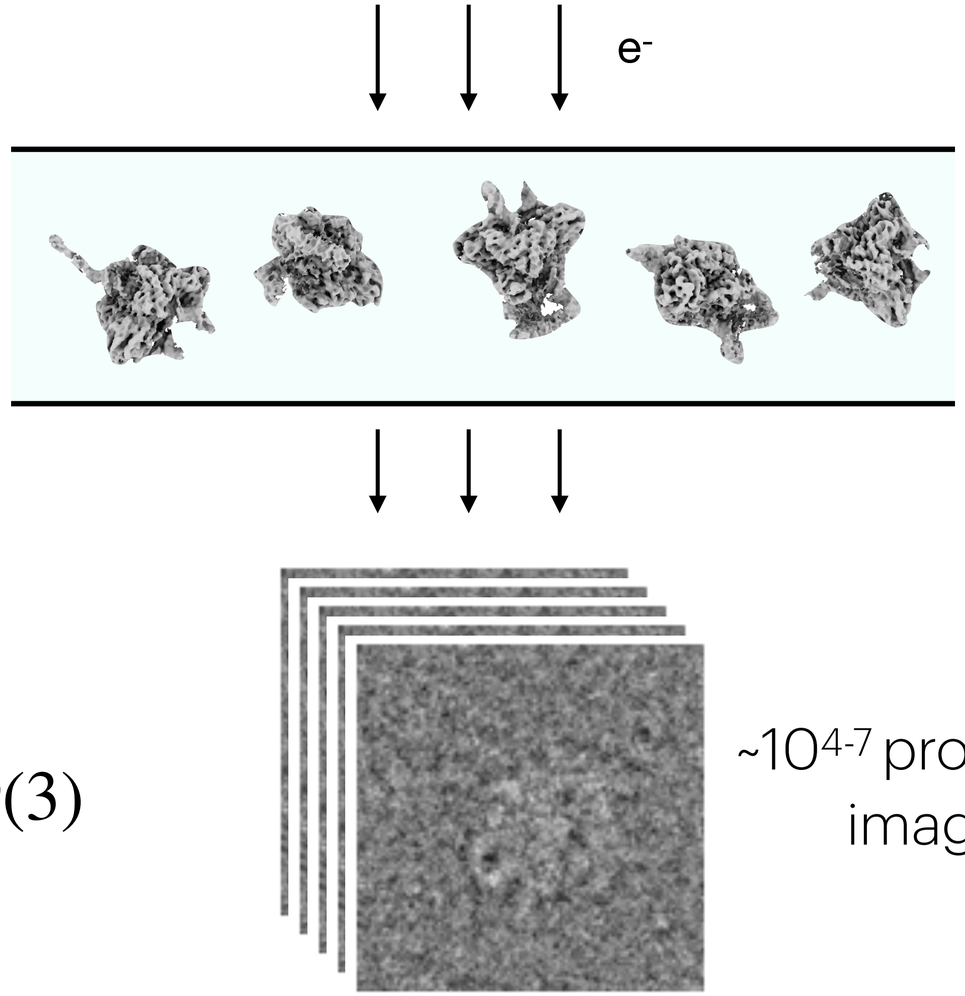
Walls et al, 2020

3) Atomic model fitting



Single particle cryo-EM image formation

- A purified solution of the molecule is fixed in a thin layer of vitreous ice
- Each cryo-EM image $X : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a tomographic projection of a volume $V : \mathbb{R}^3 \rightarrow \mathbb{R}$



$$X(x,y) = PSF * T_t * \int V(R^T(x,y,z)^T)dz \; + noise$$

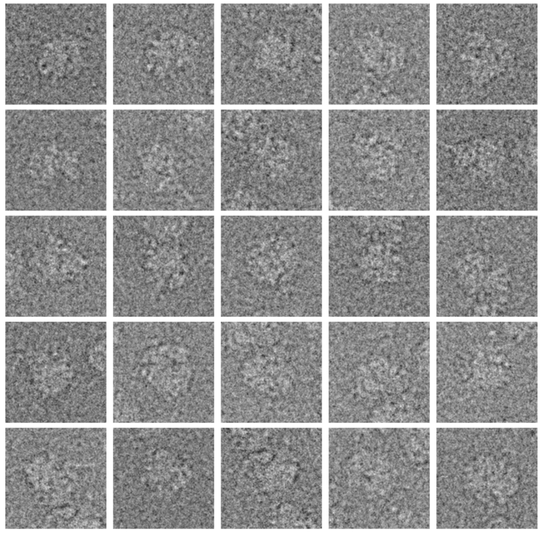
3D rotation by $R \in SO(3)$

In-plane shift by $t \in \mathbb{R}^2$

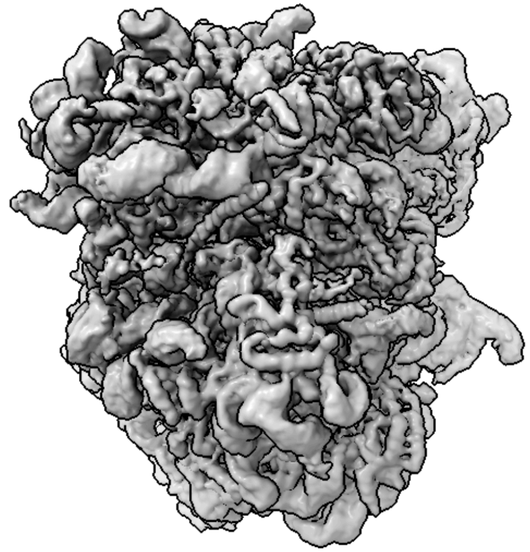
Microscope point spread function

The cryo-EM reconstruction task

- **Goal:** Reconstruct a volume $V : \mathbb{R}^3 \rightarrow \mathbb{R}$ describing a molecule's 3D structure from a set of noisy projection images $X_1, \ldots, X_N$ each containing a copy of V captured from an unknown pose $\phi_i \in (SO(3) \times \mathbb{R}^2)$

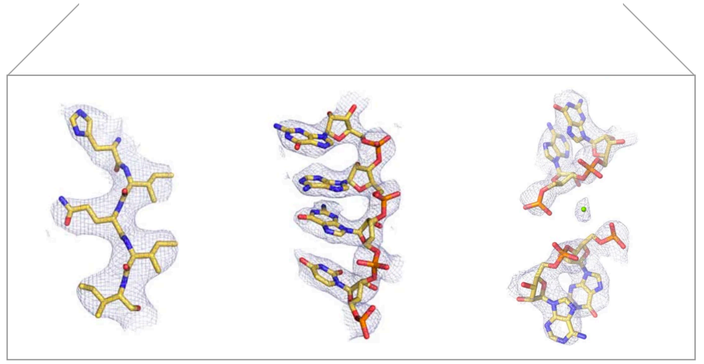


[EMPIAR-10028]



Challenges

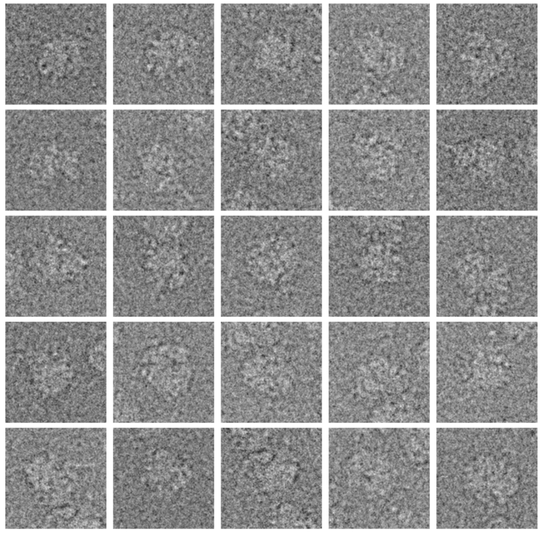
- Unknown particle poses
- Low signal to noise ratio
- Image degrading filters in microscopy
- Discretization of the measurements



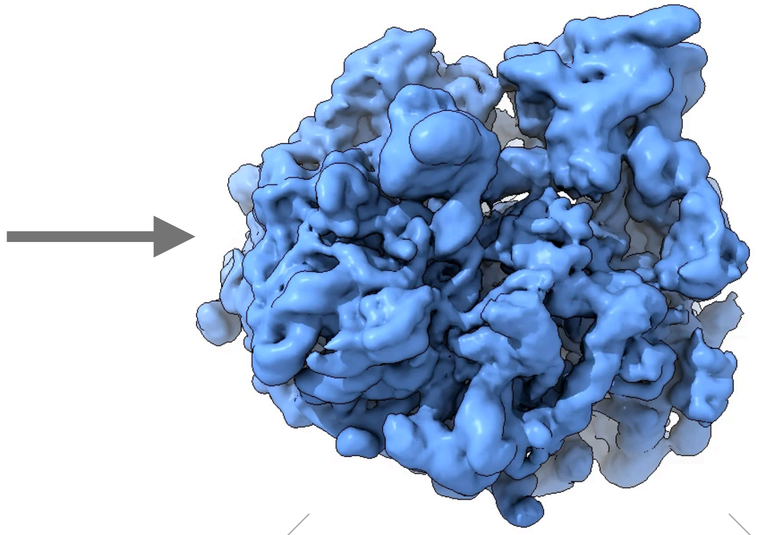
Wong et al. 2014

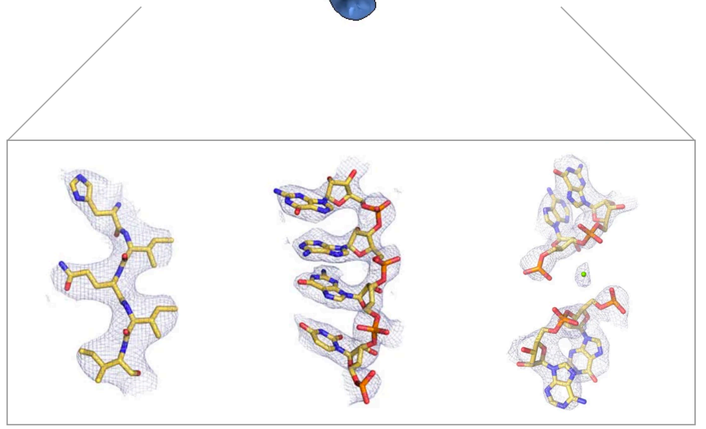
The cryo-EM reconstruction task

- **Goal:** Reconstruct a volume $V : \mathbb{R}^3 \to \mathbb{R}$ describing a molecule's 3D structure from a set of noisy projection images $X_1, \ldots, X_N$ each containing a copy of V captured from an unknown pose $\phi_i \in (SO(3) \times \mathbb{R}^2)$



[EMPIAR-10028]





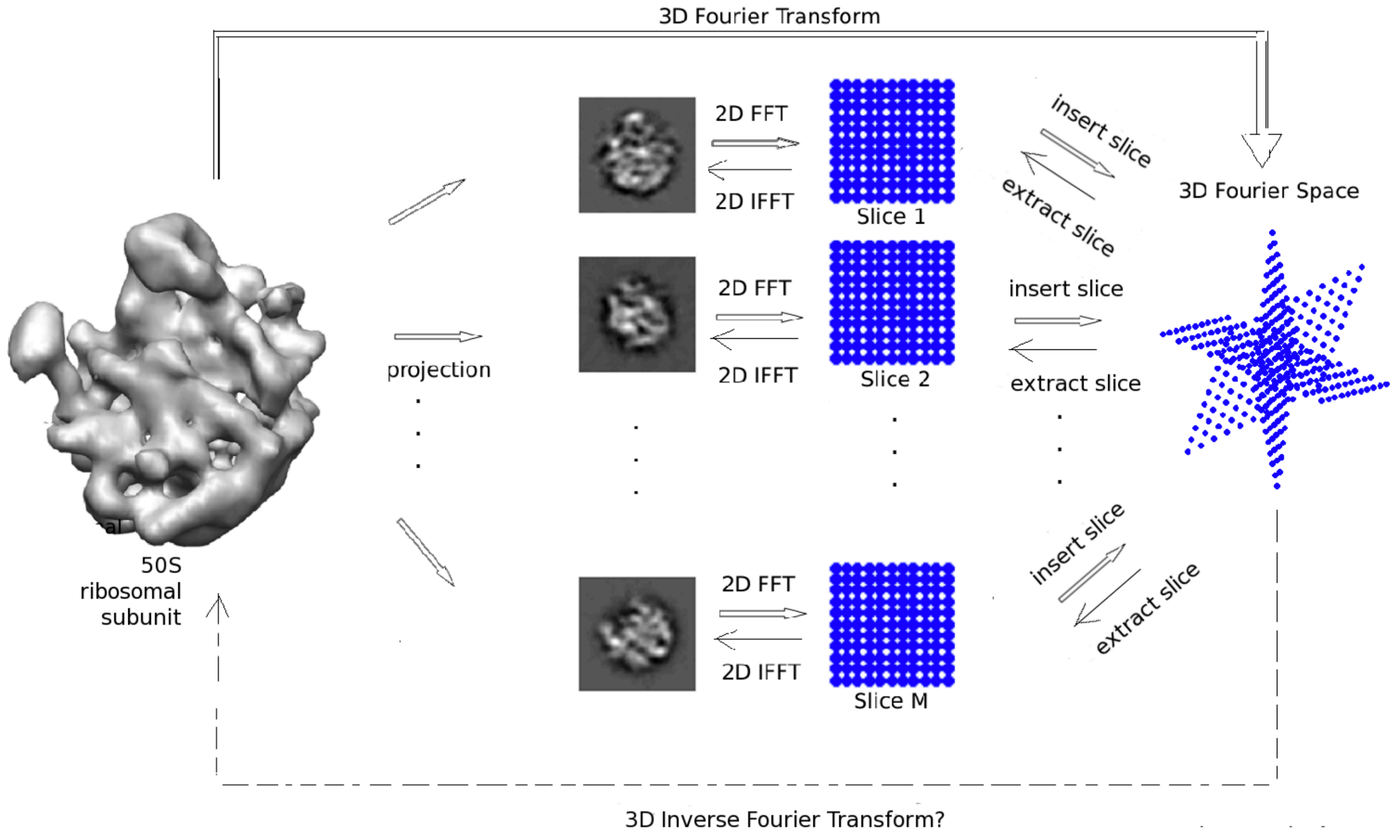
Wong et al. 2014

Challenges

- Unknown particle poses
- Low signal to noise ratio
- Image degrading filters in microscopy
- Discretization of the measurements
- *The heterogeneity problem*

The Fourier slice theorem

"The Fourier transform of a 2D projection of a volume is a central slice out of the 3D Fourier transform of the volume, perpendicular to the projection direction."



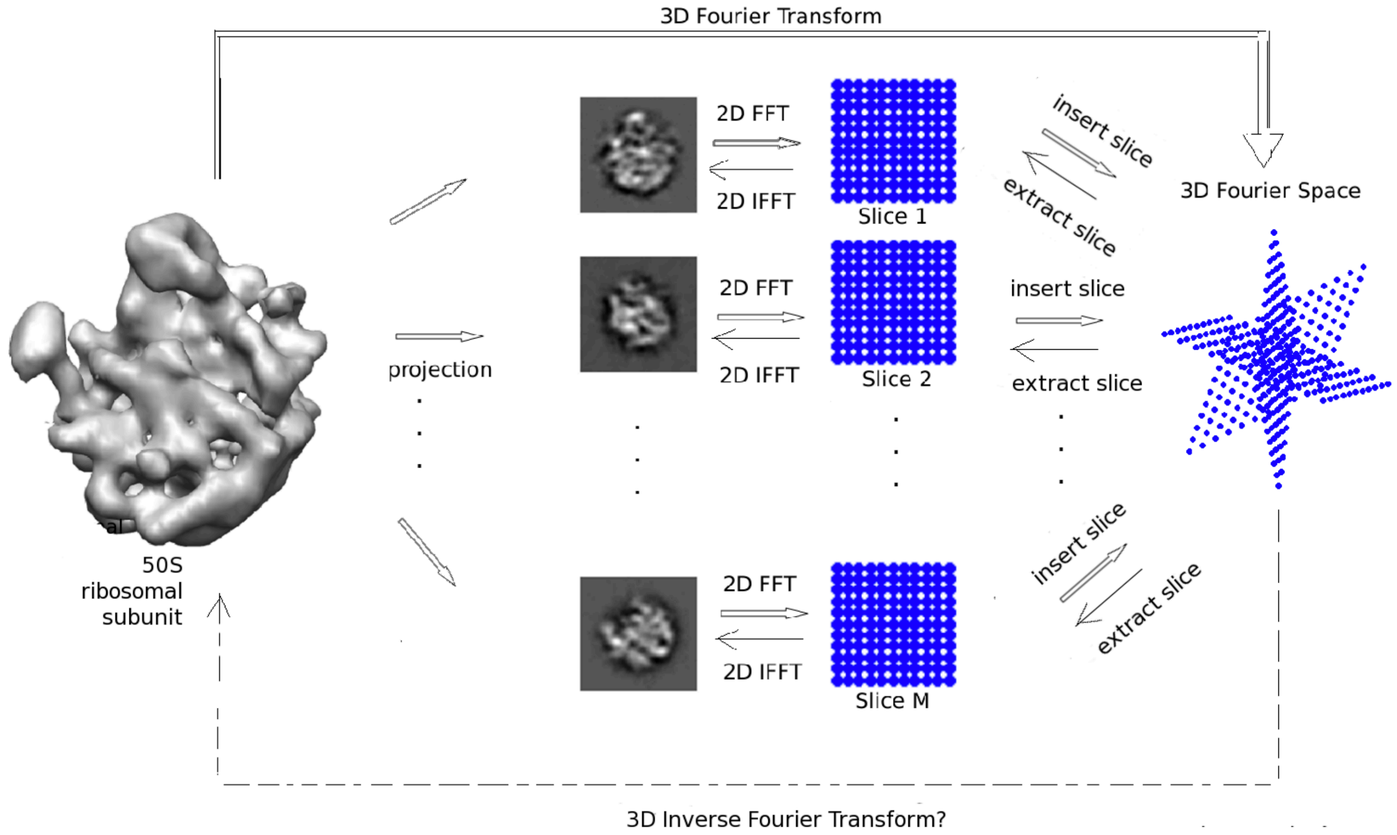
Wang, Shkolnisky, & Singer arXiv. 2013

Traditional homogeneous reconstruction algorithms

Goal: Find the 3D structure V_θ that maximizes the likelihood of data $\mathbf{x} = \{x_1, \ldots, x_N\}$, marginalizing over unknown poses $\{\phi_i\}$

$$p(\mathbf{x} \mid V_\theta) = \prod_i^N \int_{SO(3)\times\mathbb{R}^2} p(x_i, \phi_i \mid V_\theta) d\phi_i$$

- E-step: Estimate $\{\phi_i\}$ with fixed V_θ
- M-step: Estimate V_θ with fixed $\{\phi_i\}$



Wang, L., Shkolnisky, Y., & Singer, A. arXiv.org. 2013

Traditional homogeneous reconstruction algorithms

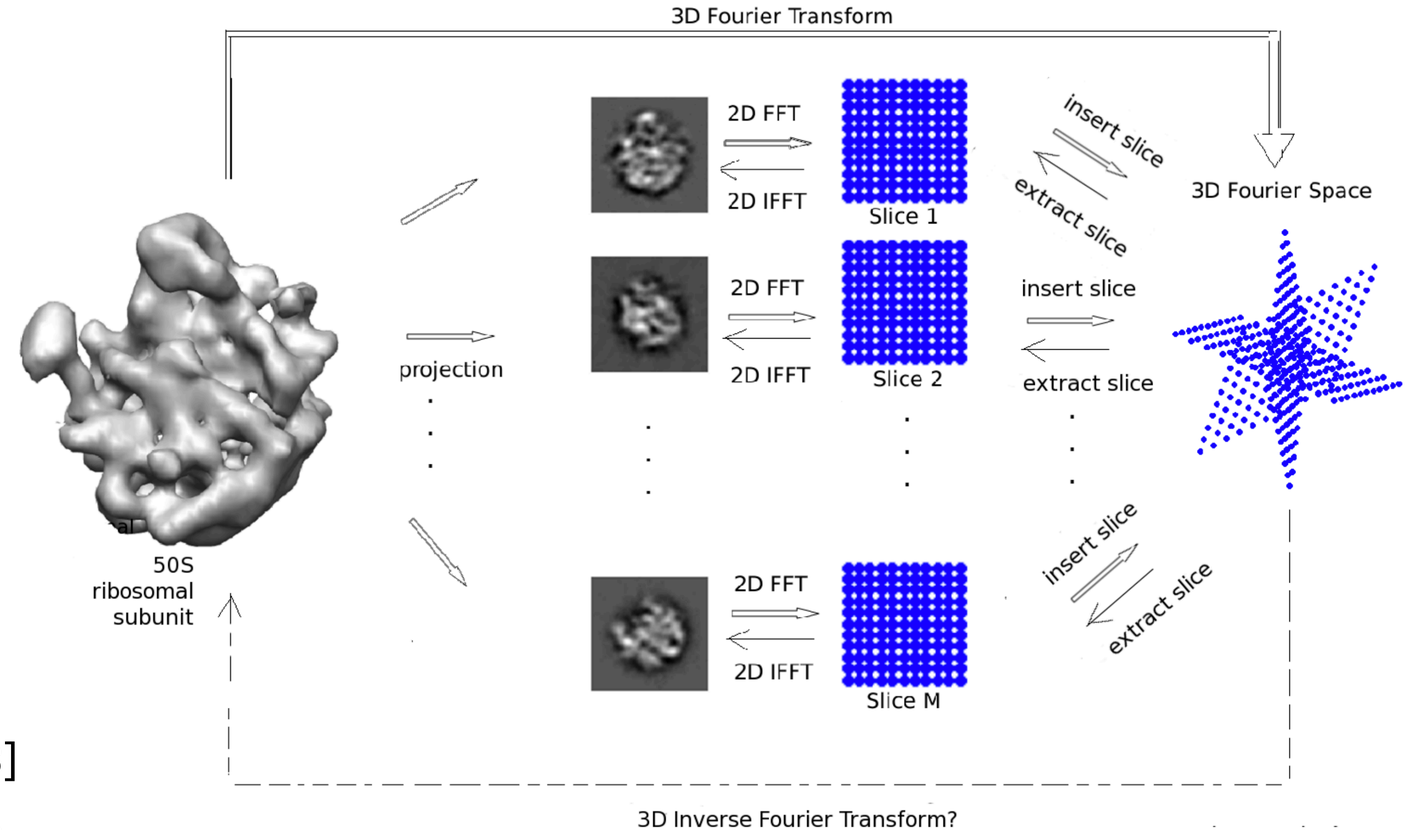
Goal: Find the 3D structure V_θ that maximizes the likelihood of data $\mathbf{x} = \{x_1, \ldots, x_N\}$, marginalizing over unknown poses $\{\phi_i\}$

$$p(\mathbf{x} \mid V_\theta) = \prod_i^N \int_{SO(3)\times\mathbb{R}^2} p(x_i, \phi_i \mid V_\theta) d\phi_i$$

- E-step: Estimate $\{\phi_i\}$ with fixed V_θ
- M-step: Estimate V_θ with fixed $\{\phi_i\}$

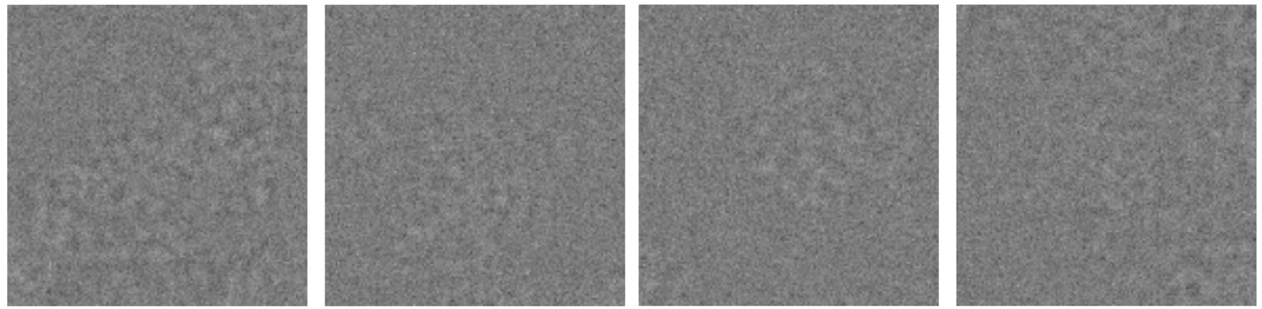
Many state of the art software packages:

- RELION: Bayesian formulation for MAP estimation proposed by Sjors Scheres [JSB 2013]
- CryoSPARC: Stochastic optimization techniques proposed by Punjani, Rubinstein, Fleet, Brubaker [CVPR 2016, Nat Methods 2017]



Wang, L., Shkolnisky, Y., & Singer, A. arXiv.org. 2013

“The heterogeneity problem”

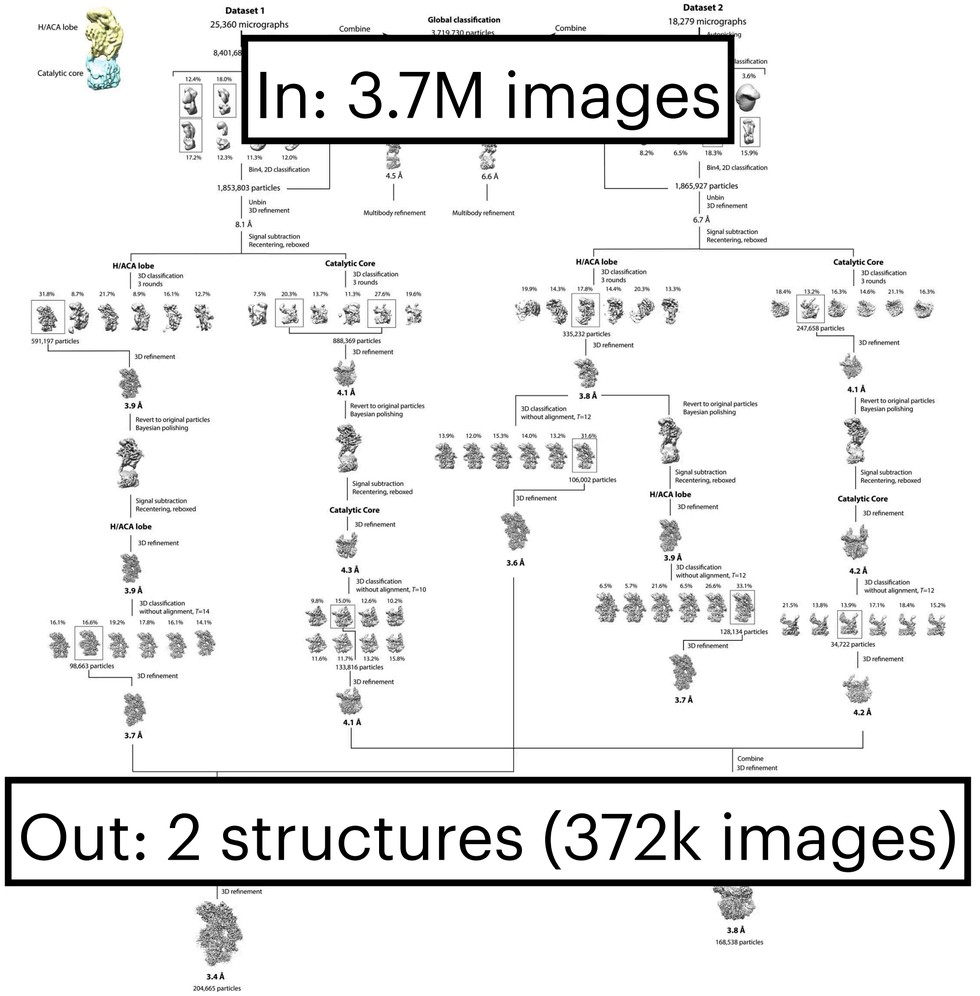


Each image contains a unique molecule

- The ability to image heterogeneous structures is a major opportunity in cryo-EM
- Standard approaches for heterogeneous structures include:
  - Discarding heterogeneous data
  - Multiclass reconstruction — a *discrete* mixture model of K independent structures
- The identification and analysis of heterogeneity — especially continuous forms — is an **open problem** in cryo-EM reconstruction.

Continuous heterogeneity: See Lederman & Singer 2017

A typical cryo-EM processing workflow

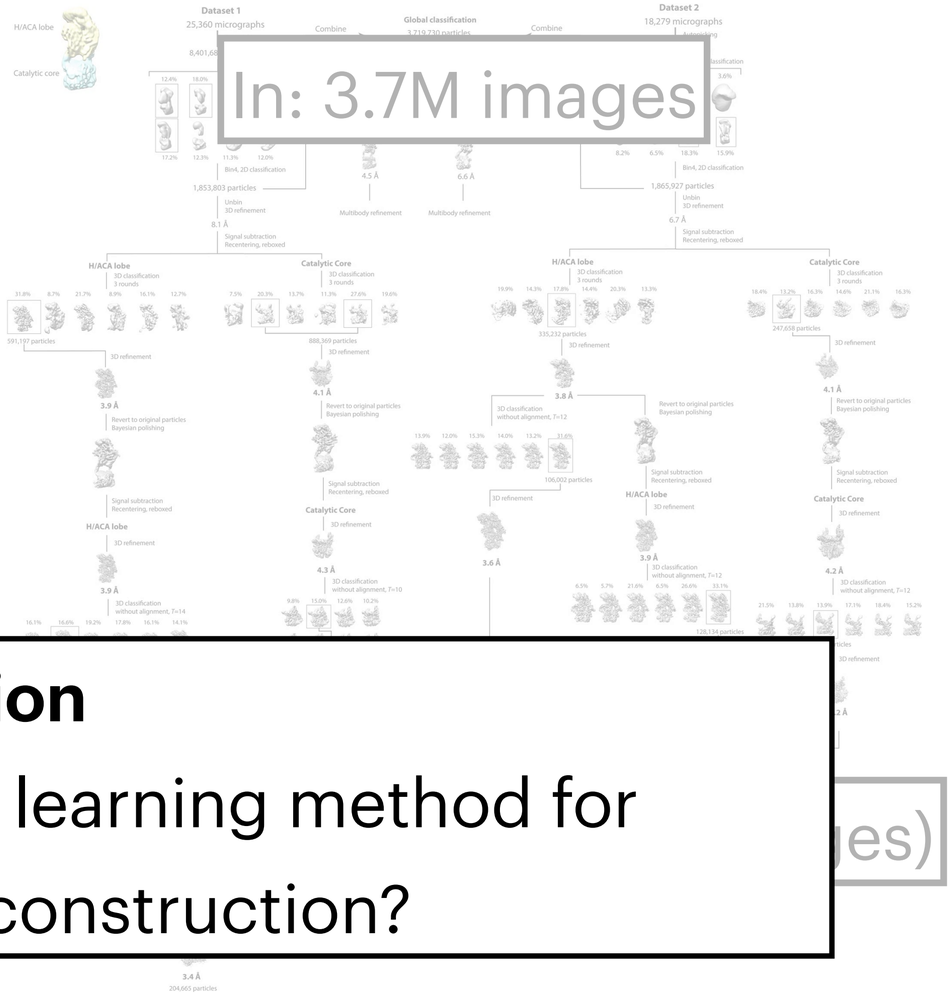


Ghanim et al. Nature 2021

“The heterogeneity problem”

Each image contains a unique molecule

A typical cryo-EM processing workflow



- The ability to image heterogeneous structures is a major opportunity in cryo-EM
- Standard approaches for heterogeneous structures include:
  - Discarding heterogeneous data
  - M
    in
- The
  con
  reconstruction.

Research Question

Can we design a modern machine learning method for heterogeneous cryo-EM reconstruction?

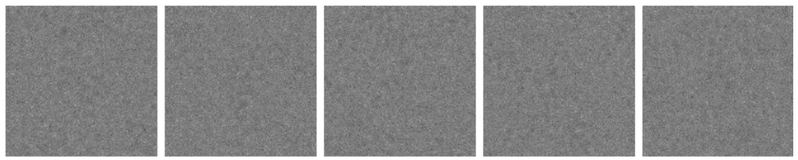
ges)

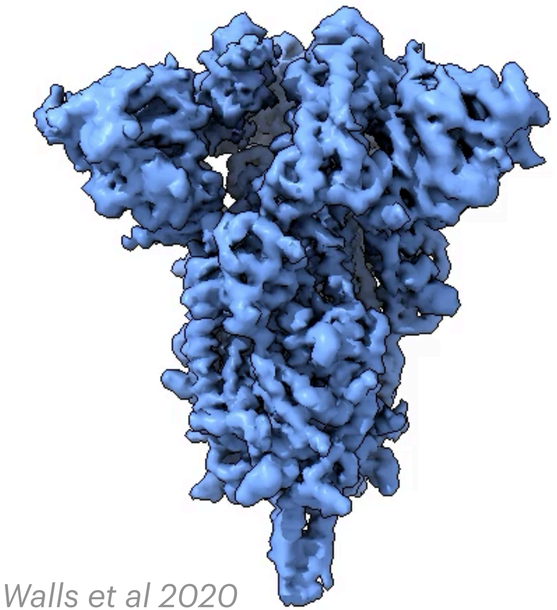
Continuous heterogeneity: See Lederman & Singer 2017

Ghanim et al. Nature 2021

CryoDRGN ❄️🐉: Deep Reconstructing Generative Networks

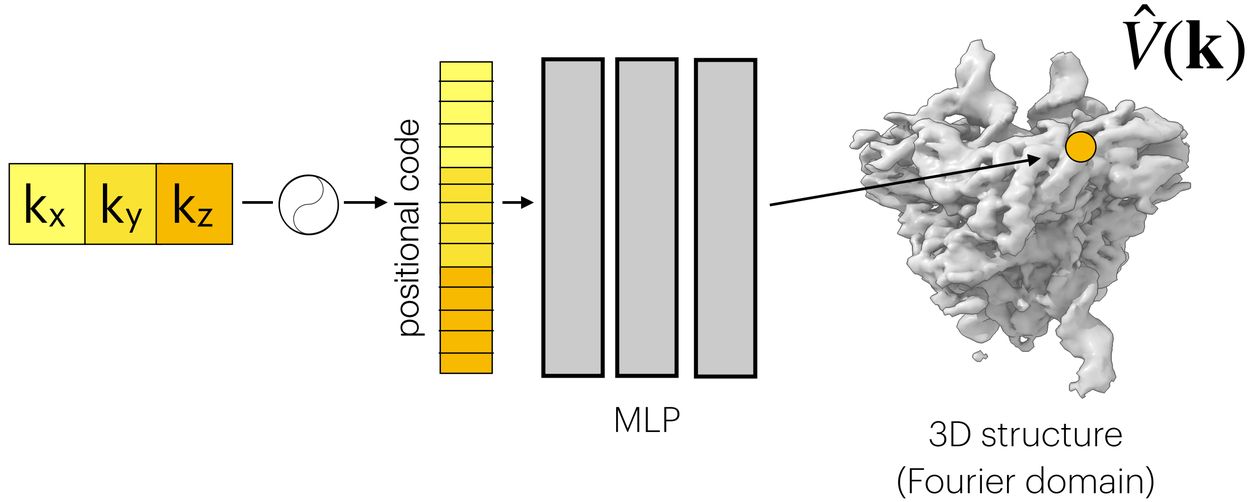
Task: 3D reconstruction from unlabeled 2D images





Dataset: Walls et al 2020

Contribution:

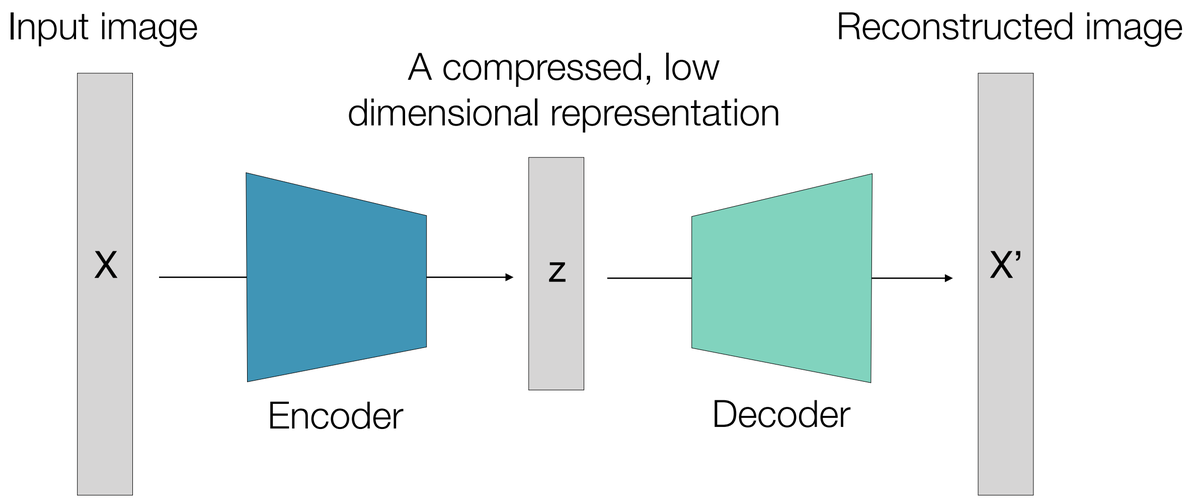


- A new paradigm for heterogeneous cryo-EM reconstruction based on *deep generative models*
- Addresses a major open problem in the field of reconstructing *continuous heterogeneity*
- Introduced a *neural field* representation of 3D structure that has shown broad applicability in computer vision (e.g. NeRF)

Zhong, Bepler, Davis, Berger, ICLR 2020 Spotlight

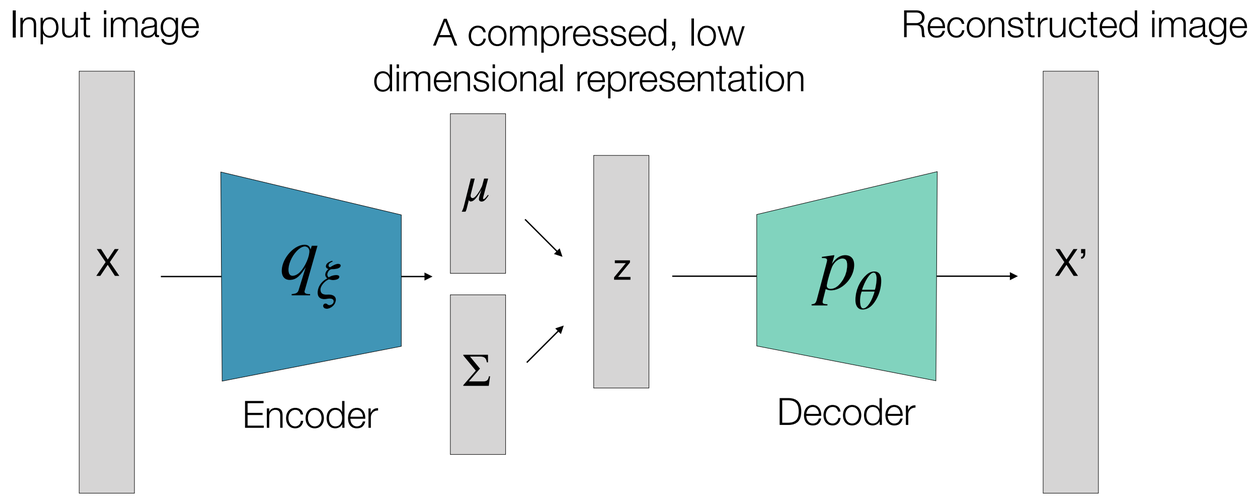
Autoencoders and Variational Autoencoders (VAEs)

- The autoencoder is a nonlinear, dimensionality reduction technique



Autoencoders and Variational Autoencoders (VAEs)

- The autoencoder is a nonlinear, dimensionality reduction technique



- The VAE extends the AE as inference of a probabilistic model — *"a regularized autoencoder"*

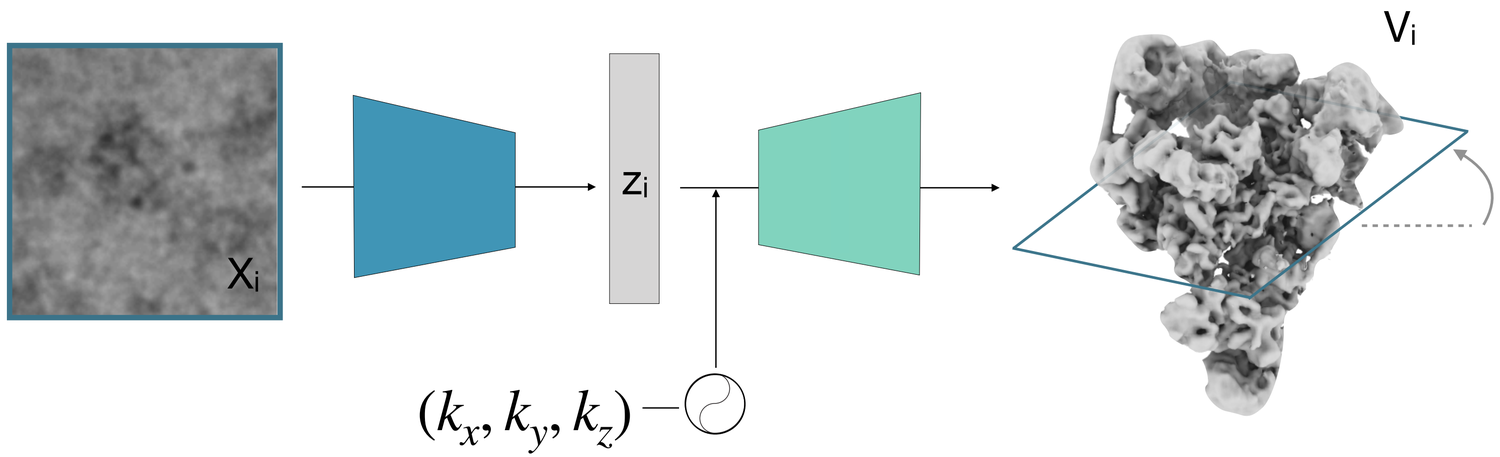
$$\mathscr{L}_{VAE}(X;\theta,\xi) = \mathbb{E}_{q_\xi(z|X)}[\log p_\theta(X|z)] - KL(q_\xi(z|X)||p(z))$$

Reconstruction error — Regularization

CryoDRGN ❄️🐉: Deep reconstructing generative networks

Unsupervised learning of a **deep generative model** of 3D biomolecular structures from 2D cryo-EM images

1. We develop **coordinate-based neural networks** to directly approximate the 3D structure
2. Fourier space **image encoder-volume decoder** architecture based on the variational autoencoder (VAE)
3. Exact inference for pose and variational inference for heterogeneity

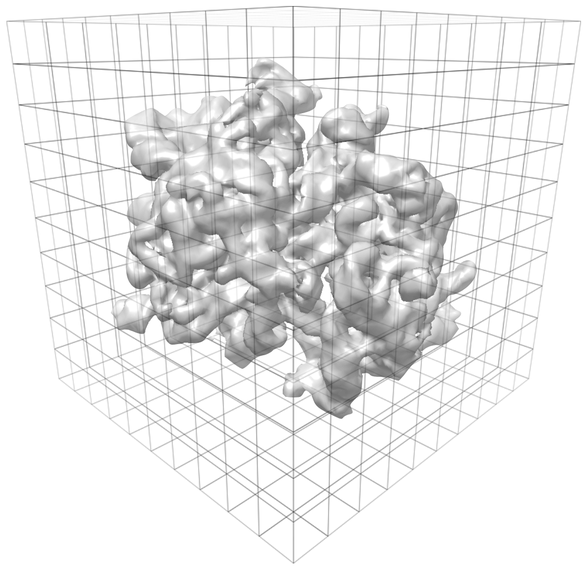


Zhong et al. ICLR 2020. Spotlight

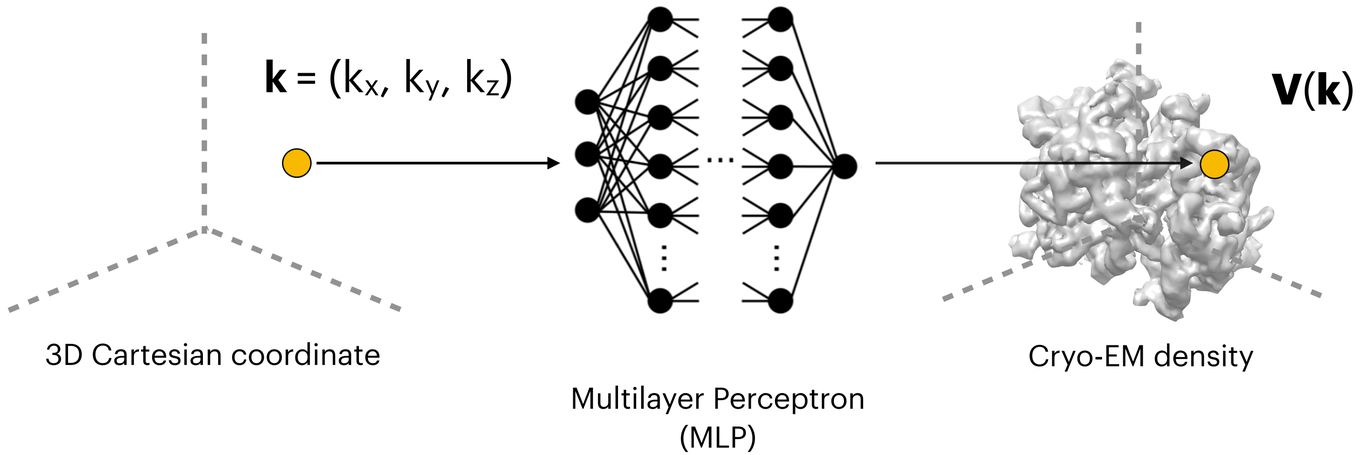
Coordinate-based neural networks for 3D volumes

- **Key idea:** Instead of representing the structure as discrete points on a 3D lattice, learn a *continuous* function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$

Traditional algorithms



cryoDRGN

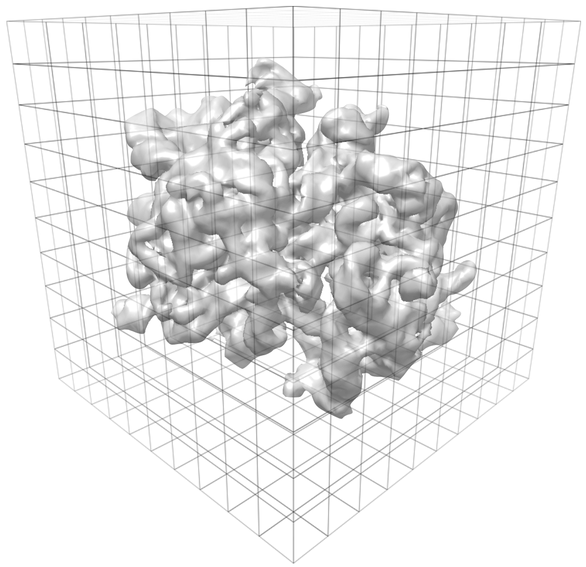


Zhong et al. ICLR 2020. Spotlight

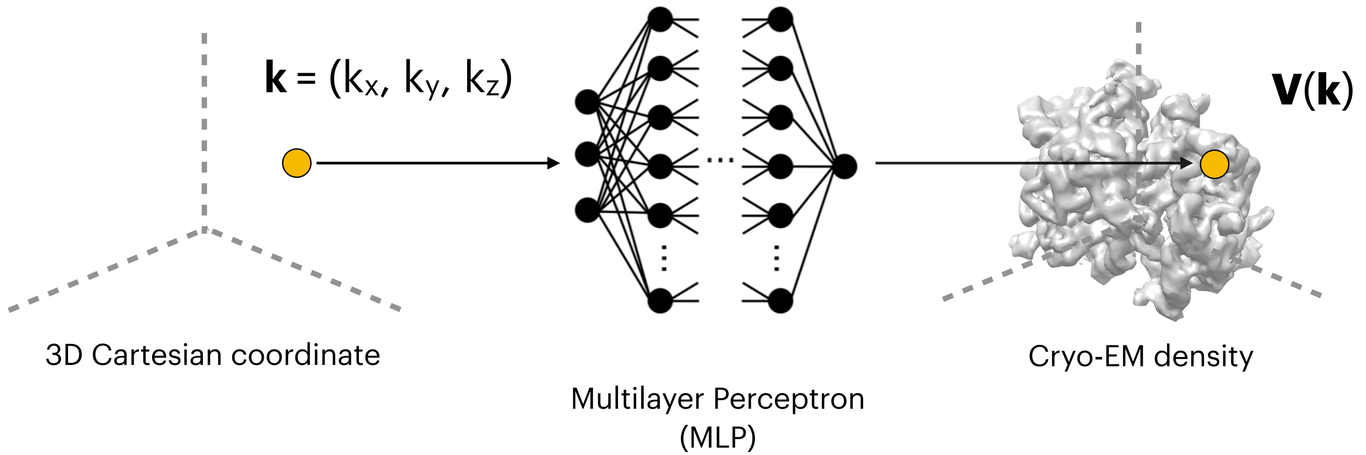
Coordinate-based neural networks for 3D volumes

- **Key idea:** Instead of representing the structure as discrete points on a 3D lattice, learn a *continuous* function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$

Traditional algorithms



cryoDRGN



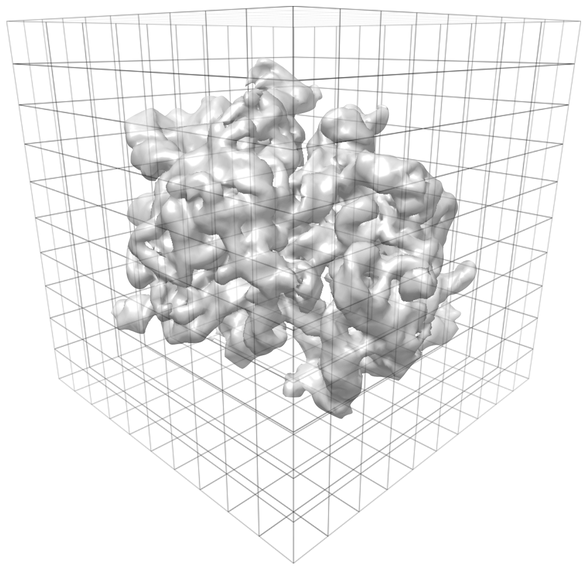
CryoDRGN structures are parameterized as a **neural network** instead of a voxel array

Zhong et al. ICLR 2020. Spotlight

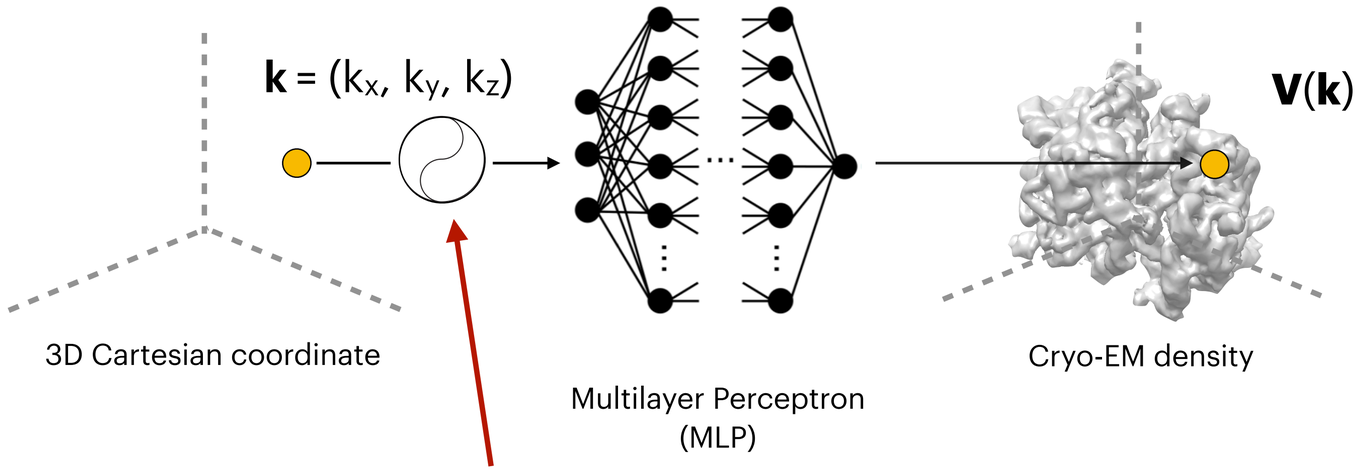
Coordinate-based neural networks for 3D volumes

- **Key idea:** Instead of representing the structure as discrete points on a 3D lattice, learn a *continuous* function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$

Traditional algorithms



cryoDRGN



Positional encoding function

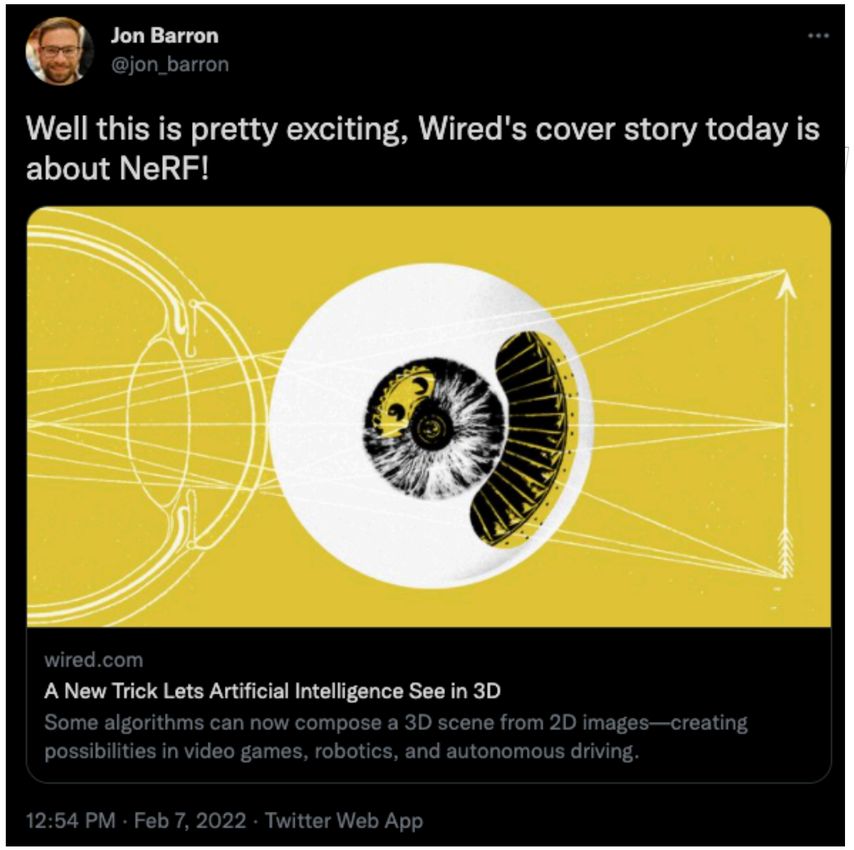
$$pe^{(2i)}(k_j) = sin(k_j D\pi(2/D)^{2i/D}),\ i = 1, ..., D/2; k_j \in k$$

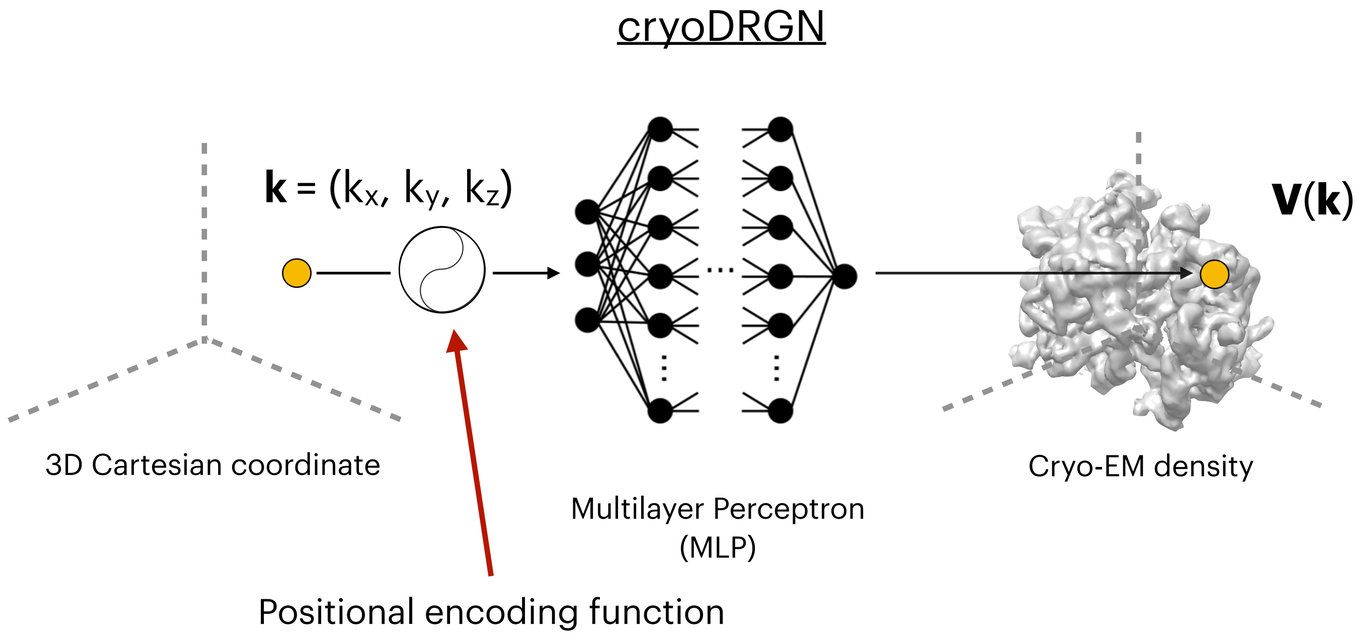
$$pe^{(2i+1)}(k_j) = cos(k_j D\pi(2/D)^{2i/D}),\ i = 1, ..., D/2; k_j \in k$$

Zhong et al. ICLR 2020. Spotlight

Coordinate-based neural networks for 3D volumes

- **Key idea:** Instead of representing the structure as discrete points on a 3D lattice, learn a *continuous* function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$





$$pe^{(2i)}(k_j) = sin(k_j D\pi(2/D)^{2i/D}),\ i = 1, ..., D/2; k_j \in k$$

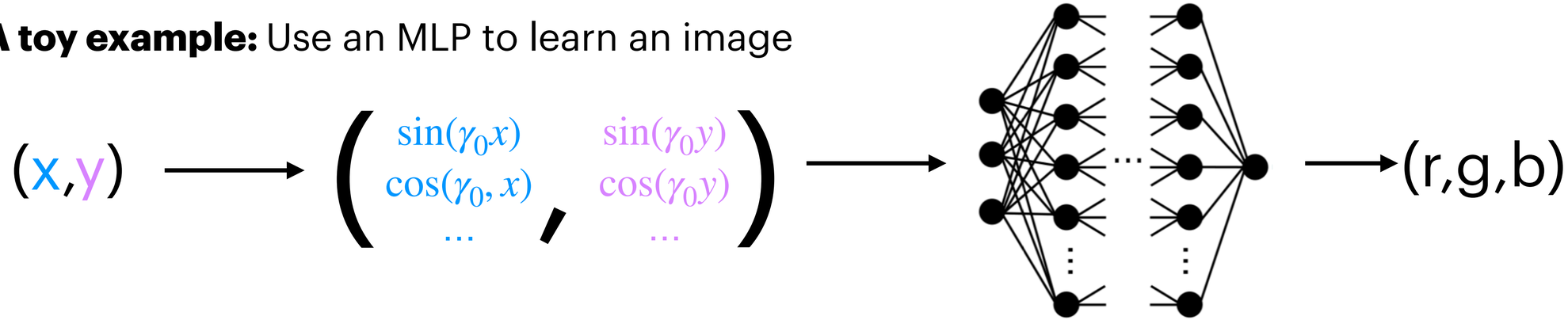
$$pe^{(2i+1)}(k_j) = cos(k_j D\pi(2/D)^{2i/D}),\ i = 1, ..., D/2; k_j \in k$$

Also see:
NeRF, Mildenhall et al. ECCV 2020

Zhong et al. ICLR 2020. Spotlight

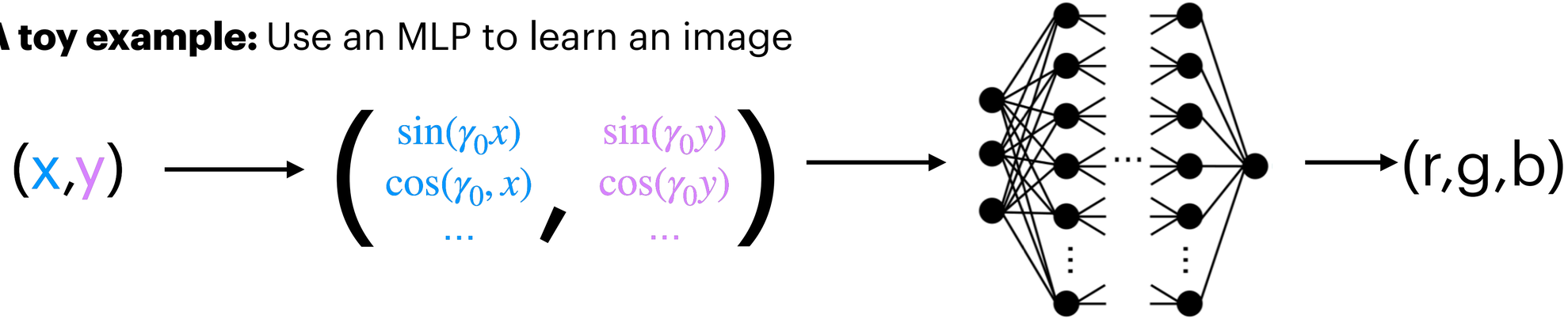
A sinusoidal encoding to featurize input coordinates

A toy example: Use an MLP to learn an image



A sinusoidal encoding to featurize input coordinates

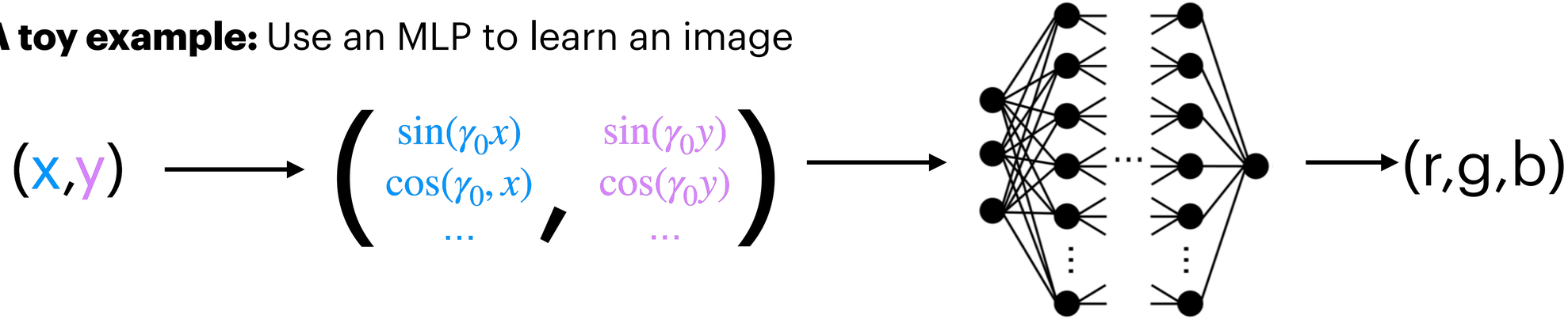
A toy example: Use an MLP to learn an image



Ground truth

A sinusoidal encoding to featurize input coordinates

A toy example: Use an MLP to learn an image

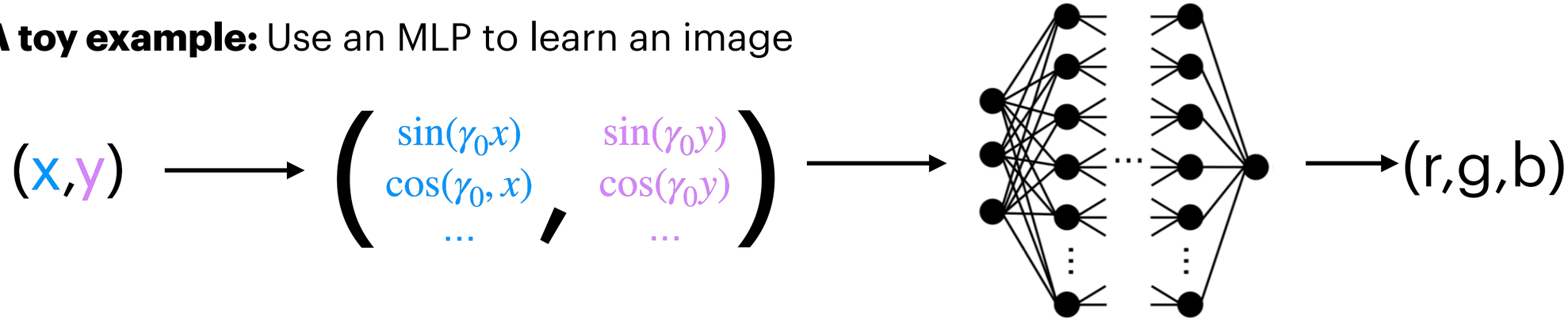


Ground truth

No sinusoidal encoding

A sinusoidal encoding to featurize input coordinates

A toy example: Use an MLP to learn an image



Ground truth

No sinusoidal encoding

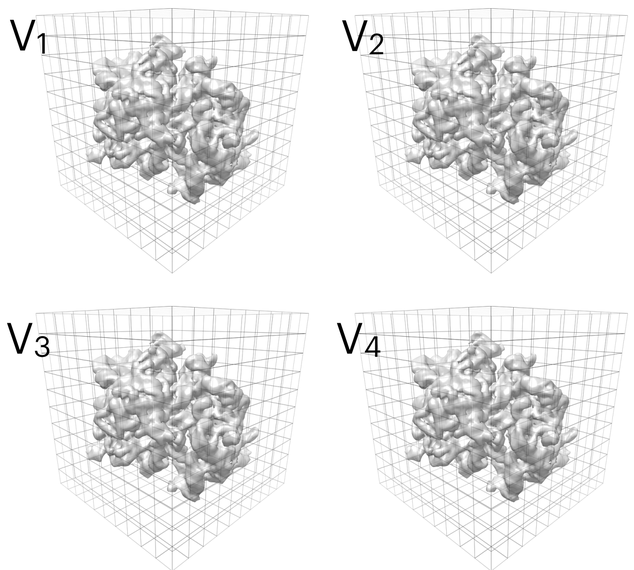
With sinusoidal encoding

Also see: Tancik et al, NeurIPS 2020

Latent variable models for heterogeneous structures

Multiclass refinement

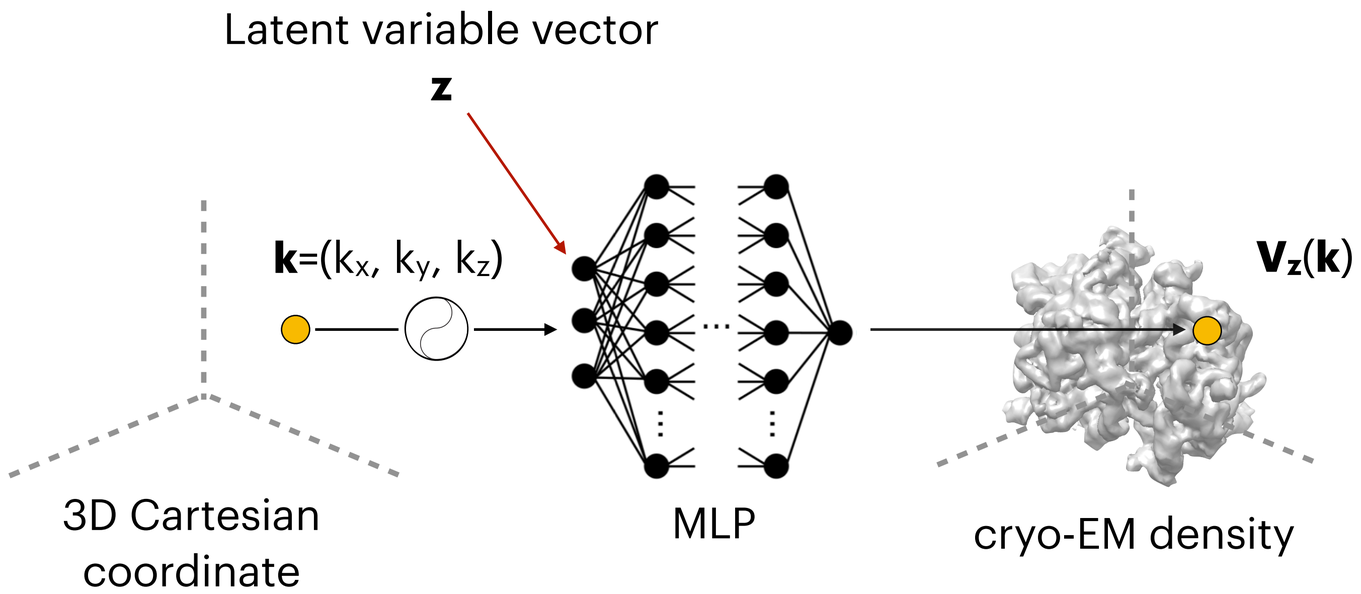
V_z, where z in {1,2,3...,K}



...

- Manual selection of K and initial volumes
- Typically, K < 10

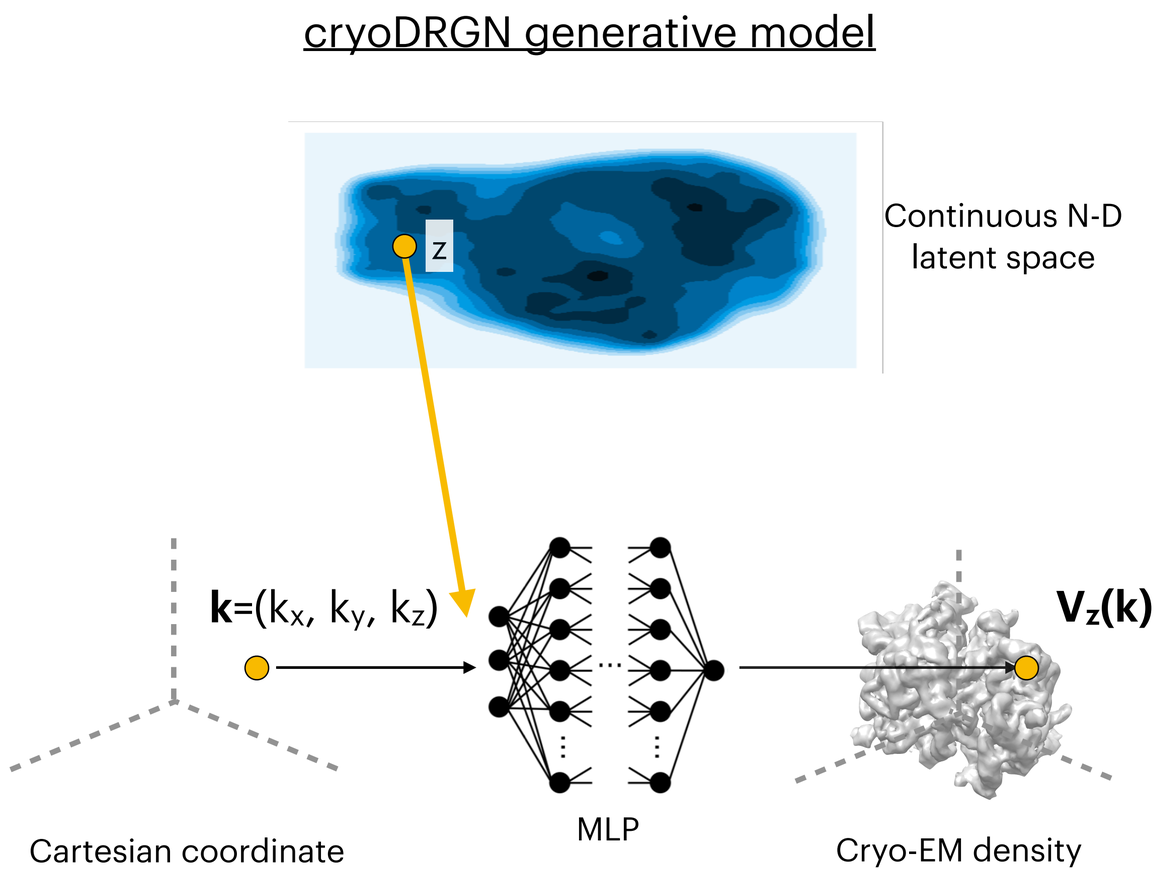
cryoDRGN



Zhong et al. ICLR 2020. Spotlight

CryoDRGN's continuous latent variable model

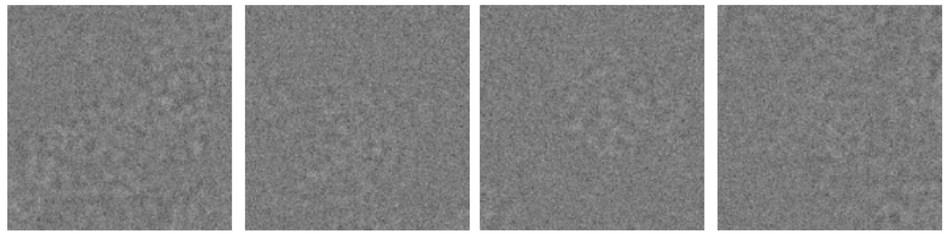
- Extend the neural representation of volume with a conditional latent variable model
- Encodes a N-dimensional *continuous* distribution over strutures

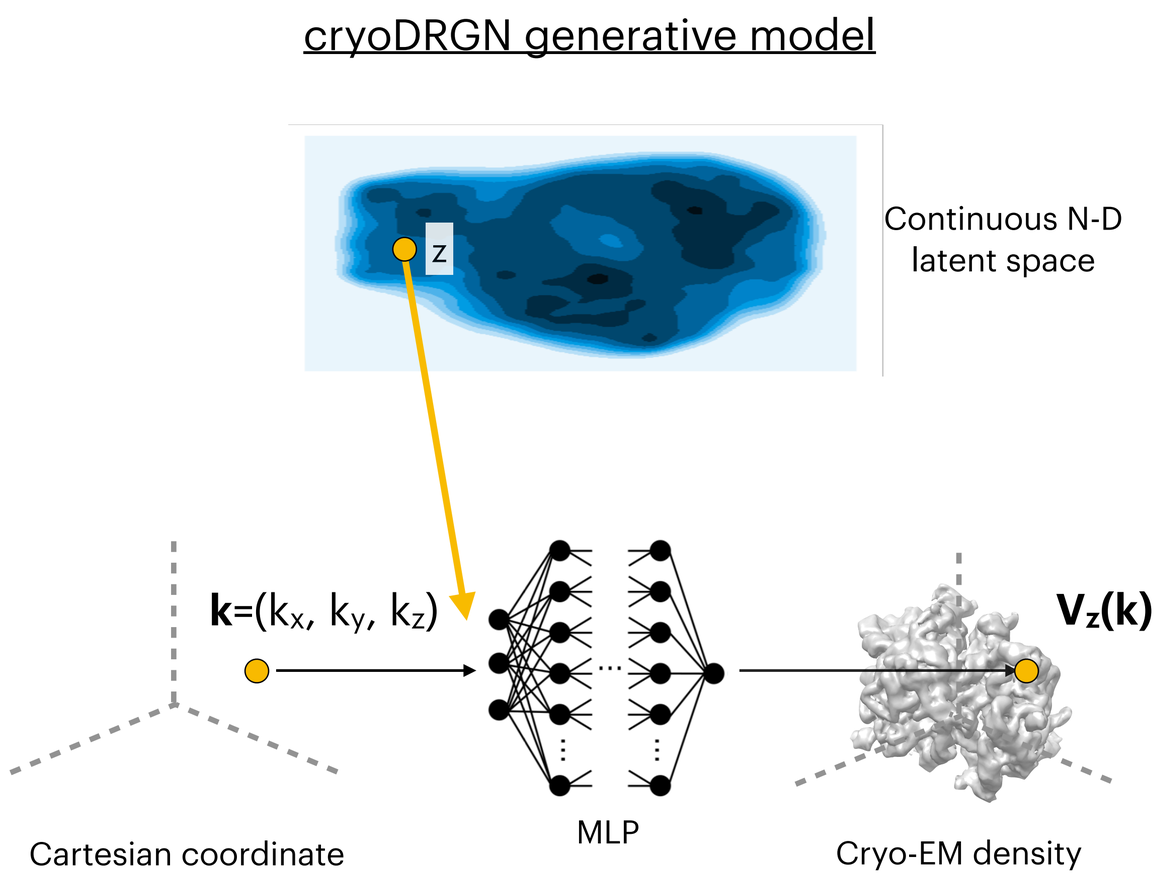


Zhong et al. ICLR 2020. Spotlight

CryoDRGN's continuous latent variable model

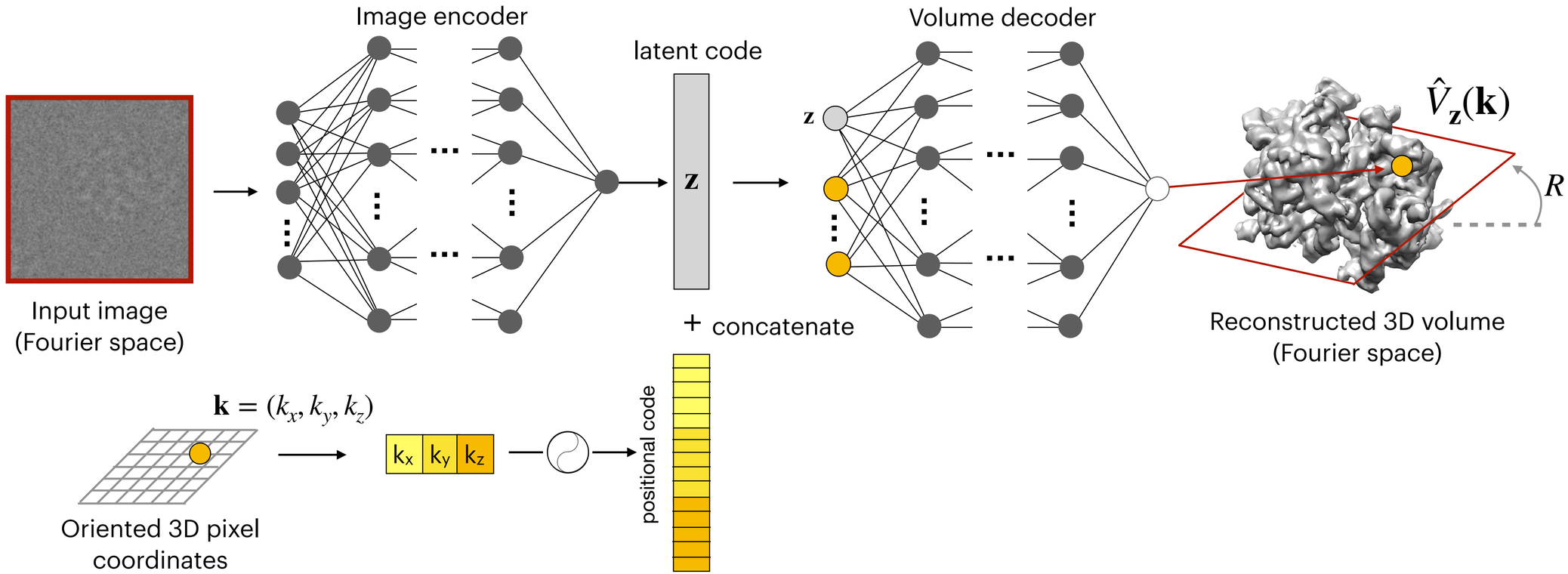
- Extend the neural representation of volume with a conditional latent variable model
- Encodes a N-dimensional *continuous* distribution over strutures
- How to learn such a model from data?





CryoDRGN's overall architecture

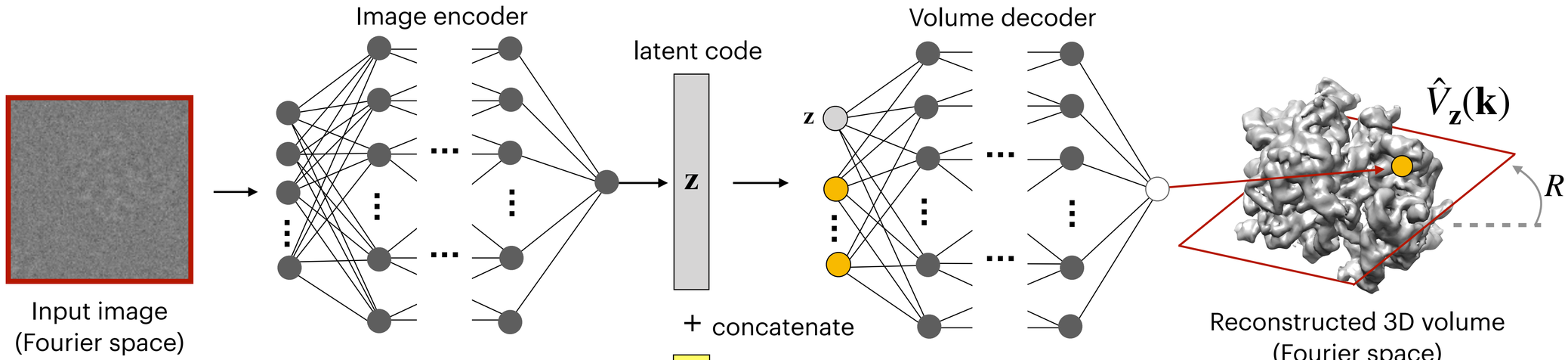
- We propose a Fourier domain **image encoder - volume decoder** architecture based on the VAE



- The decoder reconstructs an image pixel-by-pixel given **z** and the 3D coordinates of the pixels
- Coordinate-based volume architecture enforces geometric consistency between 2D views (Fourier slice theorem)

CryoDRGN's overall architecture

- We propose a Fourier domain **image encoder - volume decoder** architecture based on the VAE



- To obtain oriented 3D pixel coordinates, a coordinate lattice on the x-y plane is rotated by R
- For each image, we need to approximate its pose $\phi = (R, t)$. How?

Possible paradigms for pose inference

- Amortized variational inference [1]

$$\phi_i \sim q_\xi(\phi \mid X)$$

- Gradient descent [2]

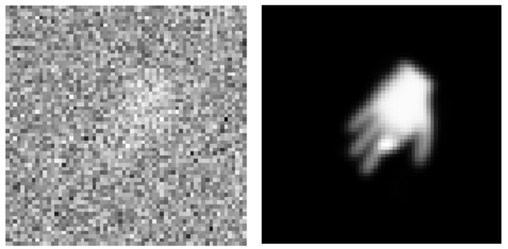
$$\phi^{(n+1)} = \phi^{(n)} - \alpha \nabla_\phi \mathcal{L}(\phi)$$

- Distribution matching/GANs [3]

$$\underset{V}{argmin}\ D(p_{sim}(X \mid V), p_{data}(X))$$

[1] Spatial-VAE Bepler et al. NeurIPS 2019; Rosenbaum et al, 2021; CryoPoseNet Nashed et al, 2021; **CryoAI** *Levy et al 2022;* **CryoFIRE** *Levy et al 2022*
[2] NeRF-- Wang et al, arXiv 2021; [3] CryoGAN Gupta et al, 2021

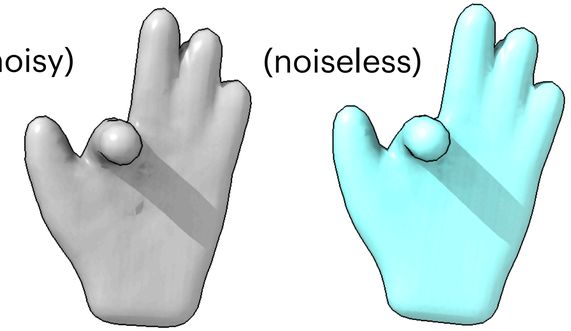
Spurious local minima in the training objective



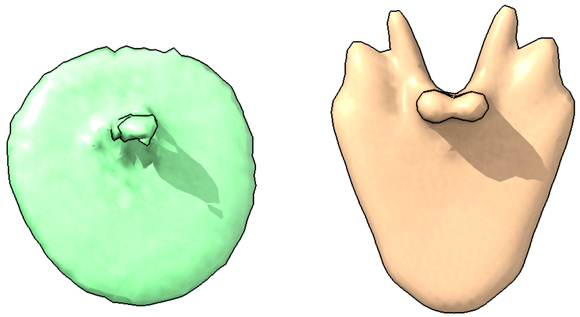
Example image

Ground truth poses

(noisy) (noiseless)

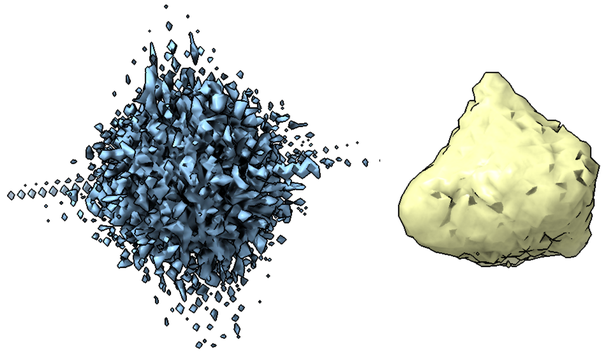


Amortized Variational Inference



$$\phi_i \sim q_\xi(\phi \mid X)$$

Pose SGD



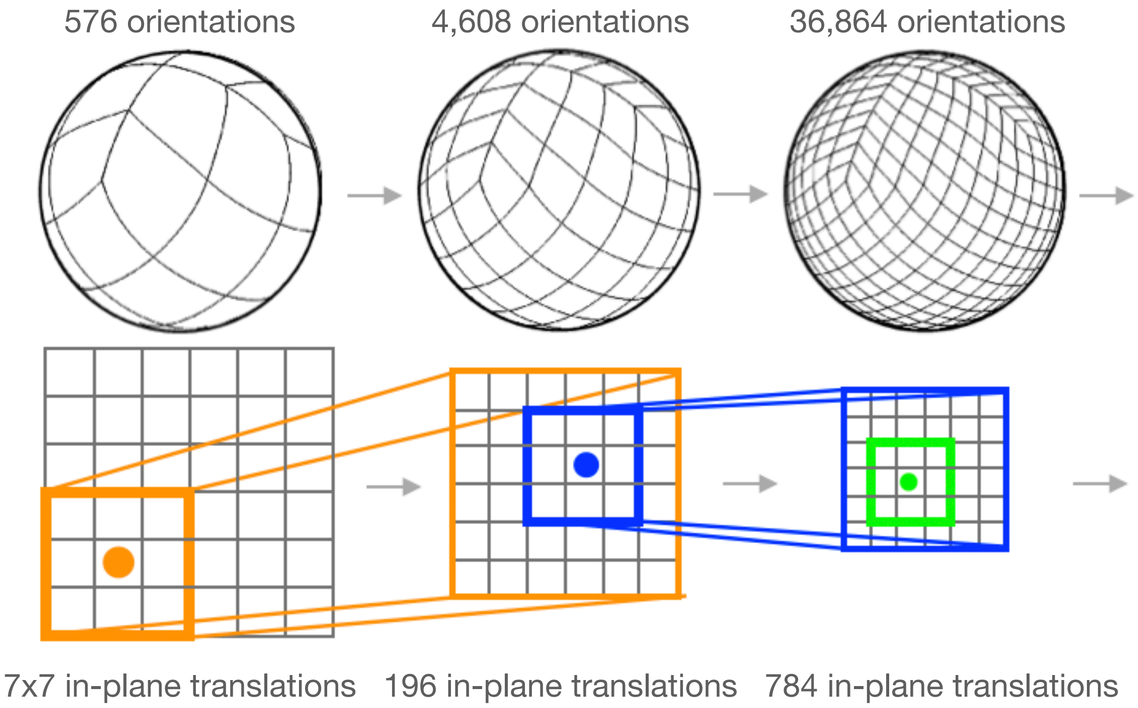
$$\phi^{(n+1)} = \phi^{(n)} - \alpha \nabla_\phi \mathcal{L}(\phi)$$

Search algorithms for inference of image pose

- Instead we perform a global search over a discretization of $SO(3) \times \mathbb{R}^2$ for the MLE pose for each image X_i given the current decoder V_θ

$$argmax_{\phi_i}\ p(X_i \mid V_\theta)$$

- A hierarchical search procedure:
  - Start with an exhaustive search over a discretization of the 5D space of poses
    - A uniform discretization of SO(3) with the Hopf fibration, regular 2D grid for in-plane translations
  - Iteratively refine the poses by keeping the top K poses that minimize the reconstruction loss
    - Choose K via a branch-and-bound procedure[1]



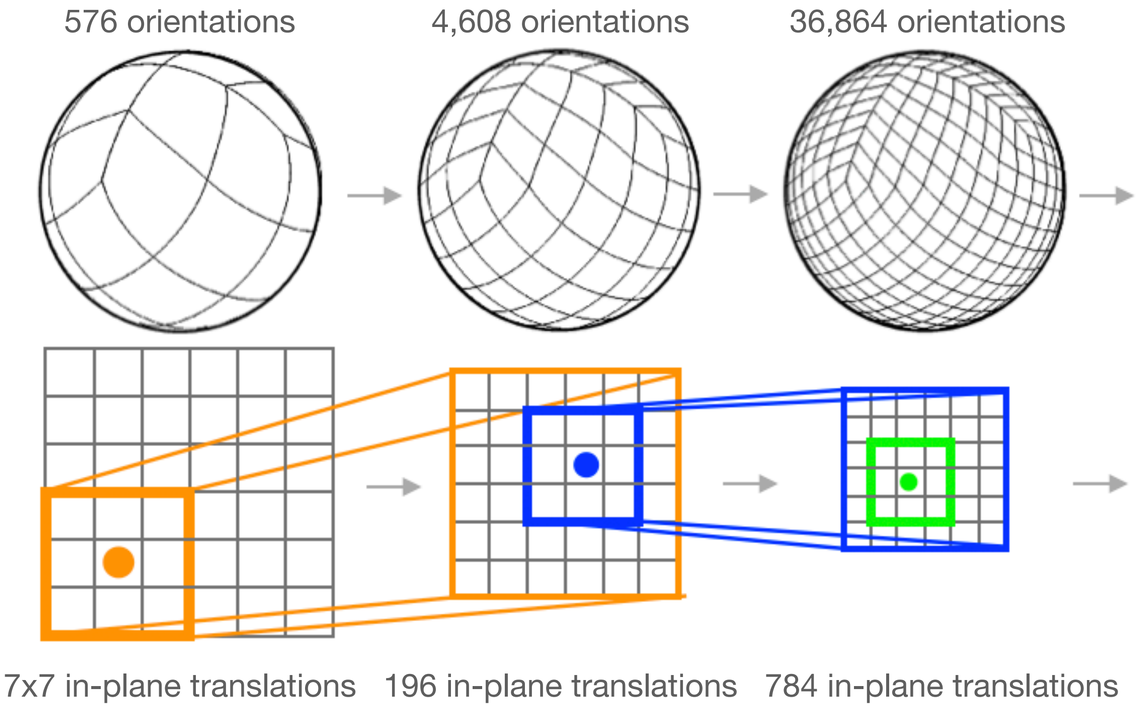
[1] cryoSPARC; Punjani et al 2017

Search algorithms for inference of image pose

- Instead we perform a global search over a discretization of $SO(3) \times \mathbb{R}^2$ for the MLE pose for each image X_i given the current decoder V_θ

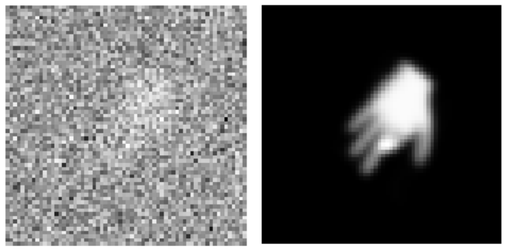
$$argmax_{\phi_i}\ p(X_i \mid V_\theta)$$

- Frequency marching[1]:
  - Band limit the loss function to low frequency components
  - Benefit 1: Computational efficiency
  - Benefit 2: Prevent overfitting



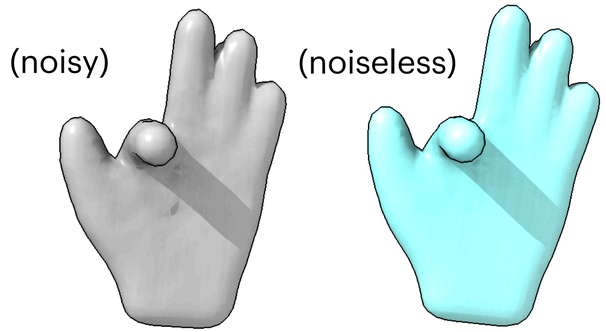
[1] Barnett et al. 2016

Spurious local minima in the training objective

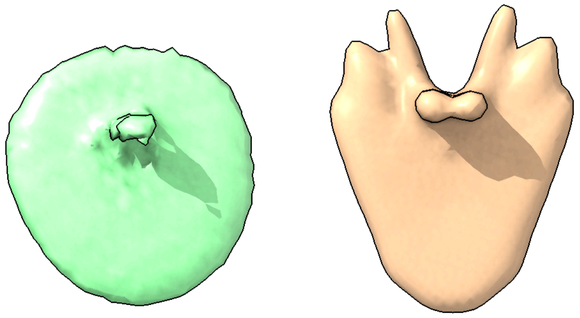


Example image

Ground truth poses

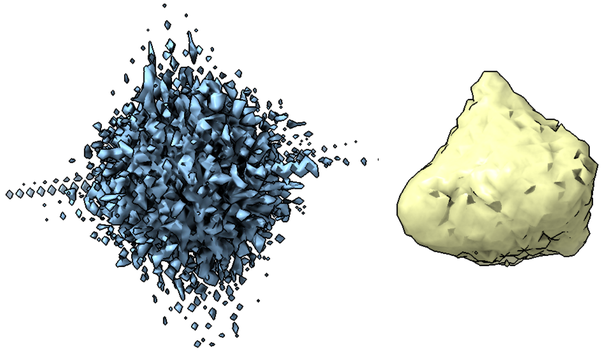


Amortized variational inference



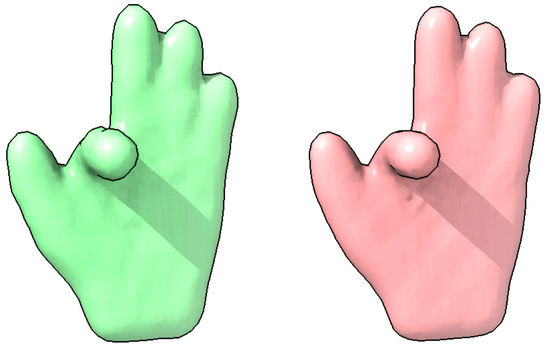
$$\phi_i \sim q_\xi(\phi \mid X)$$

Pose SGD



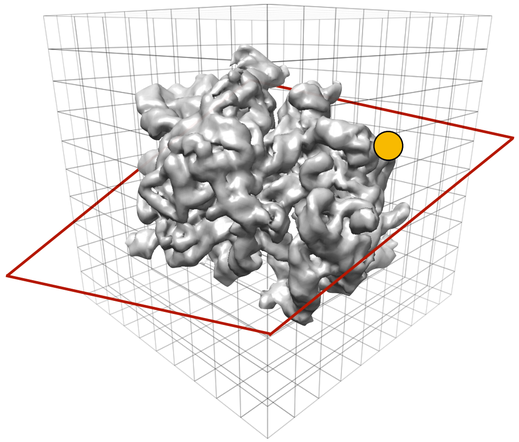
$$\phi^{(n+1)} = \phi^{(n)} - \alpha \nabla_\phi \mathcal{L}(\phi)$$

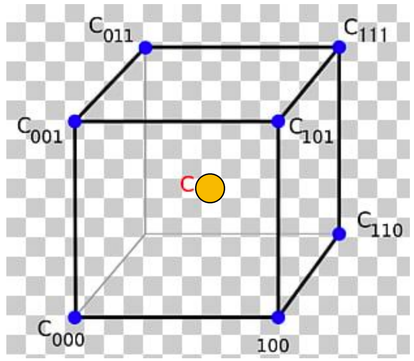
Hierarchical search



Pose search: Traditional vs. neural

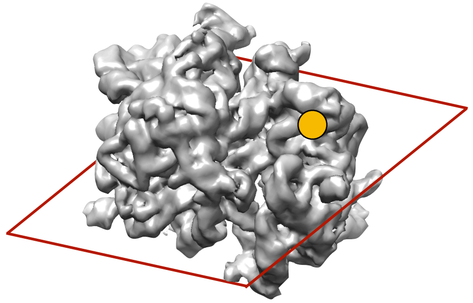
Traditional

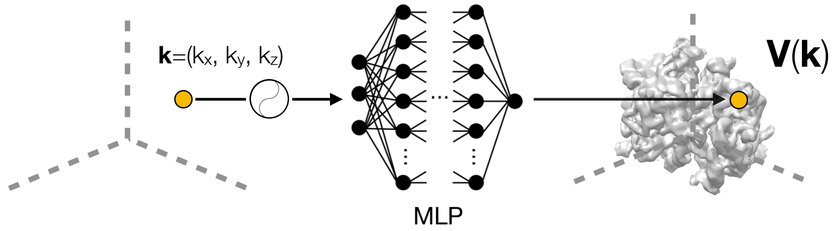




Each off-voxel point is computed as the weighted average of its 8 spatially closest neighbors

cryoDRGN

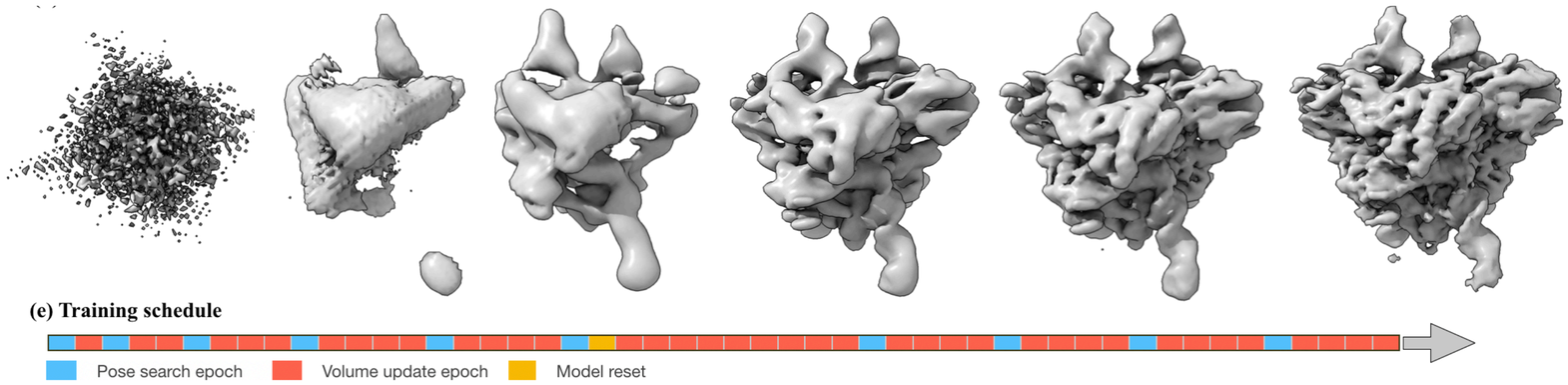


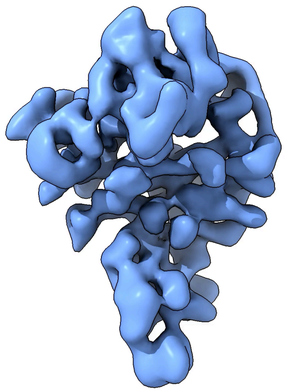
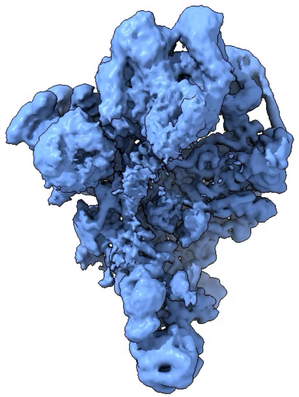
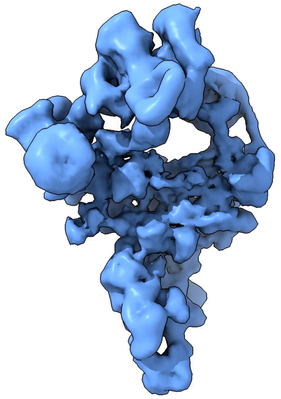


Each off-voxel point is computed by evaluating the MLP

cryoDRGN models are much more expensive to evaluate

CryoDRGN2: *Ab initio* heterogeneous reconstruction of real data



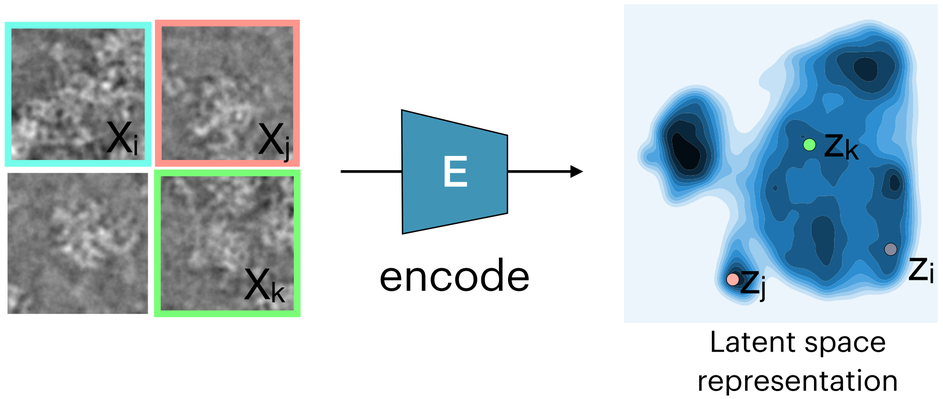
	cryoDRGN1	cryoDRGN, pose supervision	**cryoDRGN2**
			
Training time:	38.3 hours (1 GPU)	3.8 h	11.8 h

Zhong, Lerer, Davis, Berger. ICCV 2021

Analyzing the generative model

CryoDRGN at test time:

- Use the encoder network to evaluate the latent embedding **z** for each image

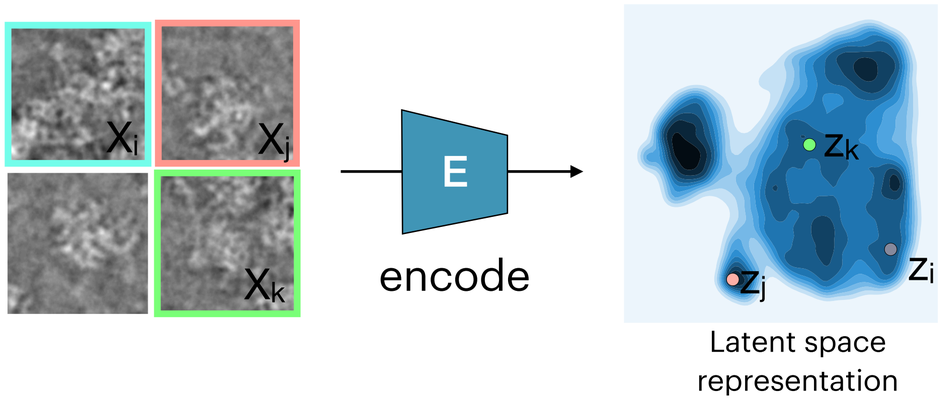


View empirical data distribution in latent space

Analyzing the generative model

CryoDRGN at test time:

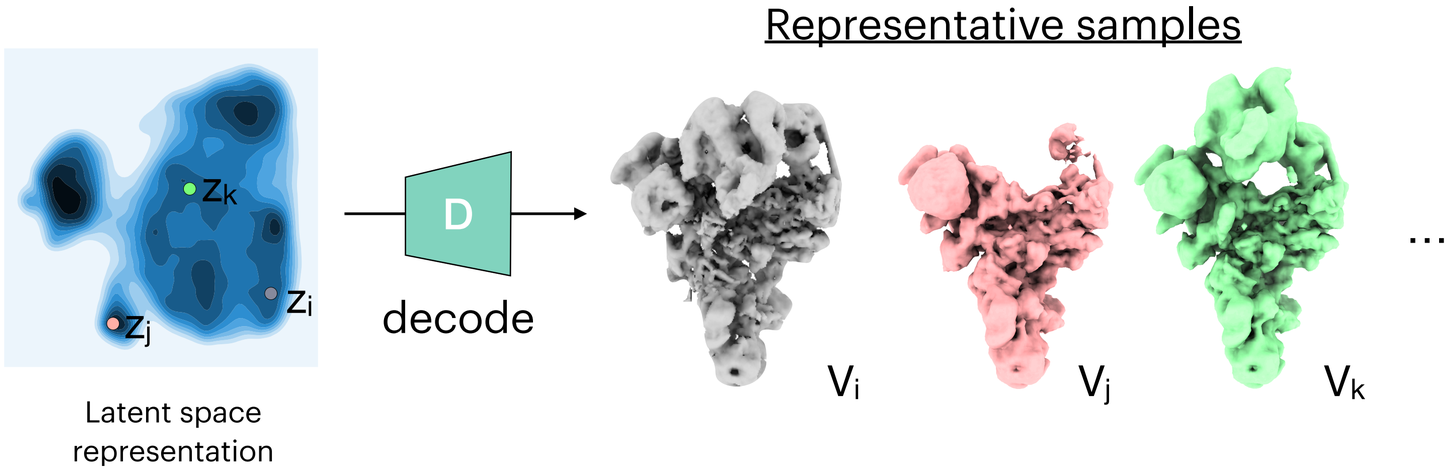
- Use the encoder network to evaluate the latent embedding **z** for each image



View empirical data distribution in latent space

- Use the decoder network to generate **V** at different values of **z**

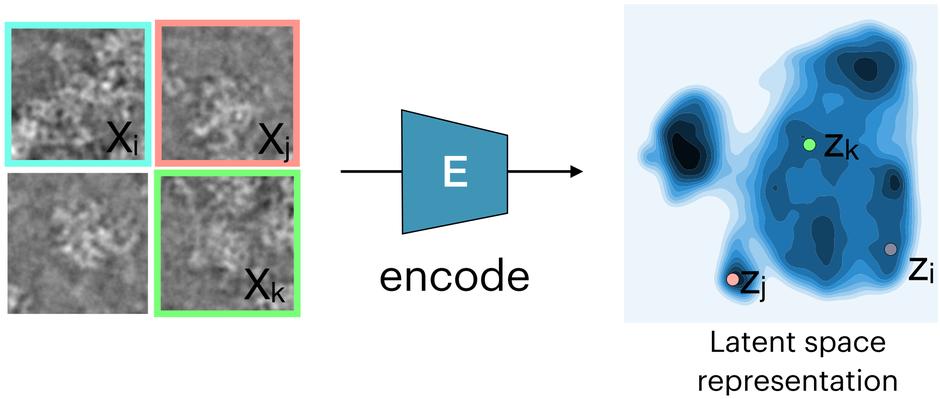
View the structural ensemble and generate movies from trajectories in latent space



Analyzing the generative model

CryoDRGN at test time:

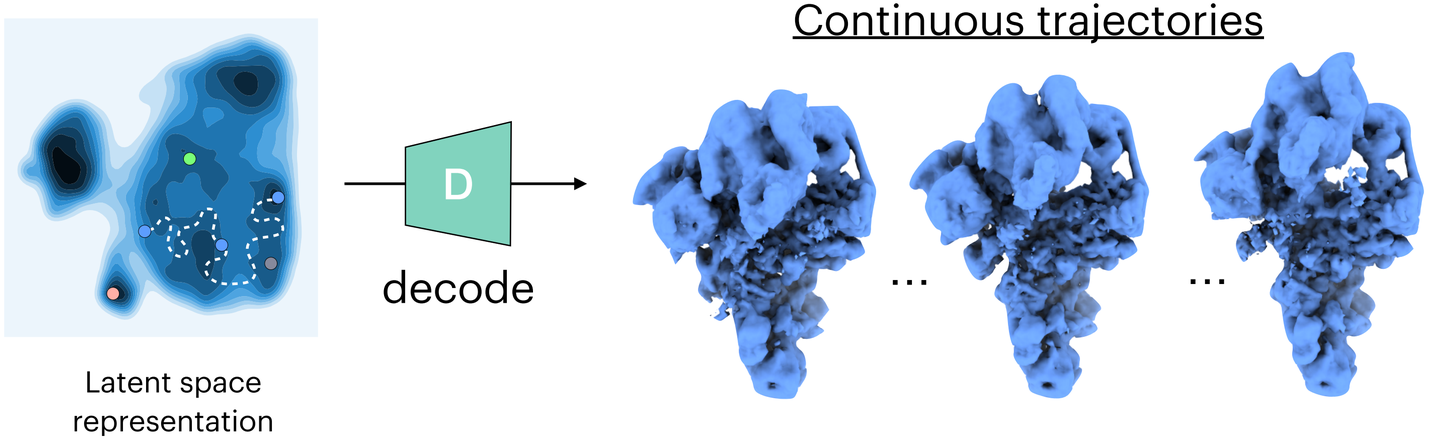
- Use the encoder network to evaluate the latent embedding **z** for each image



View empirical data distribution in latent space

- Use the decoder network to generate **V** at different values of **z**

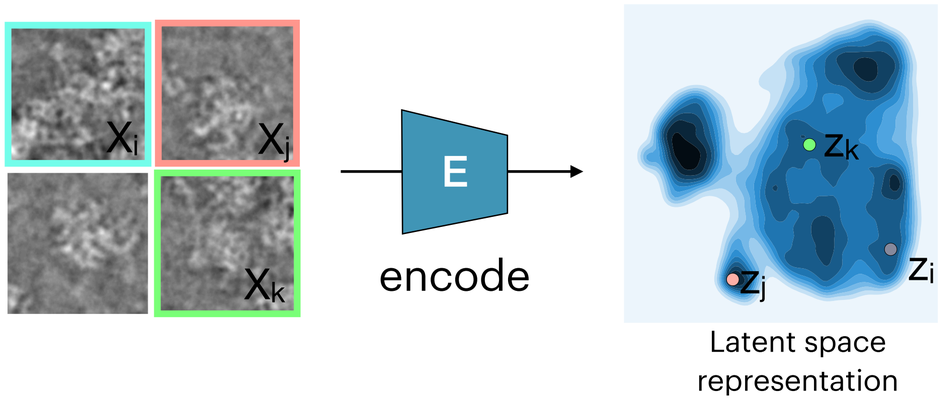
View the structural ensemble and generate movies from trajectories in latent space



Analyzing the generative model

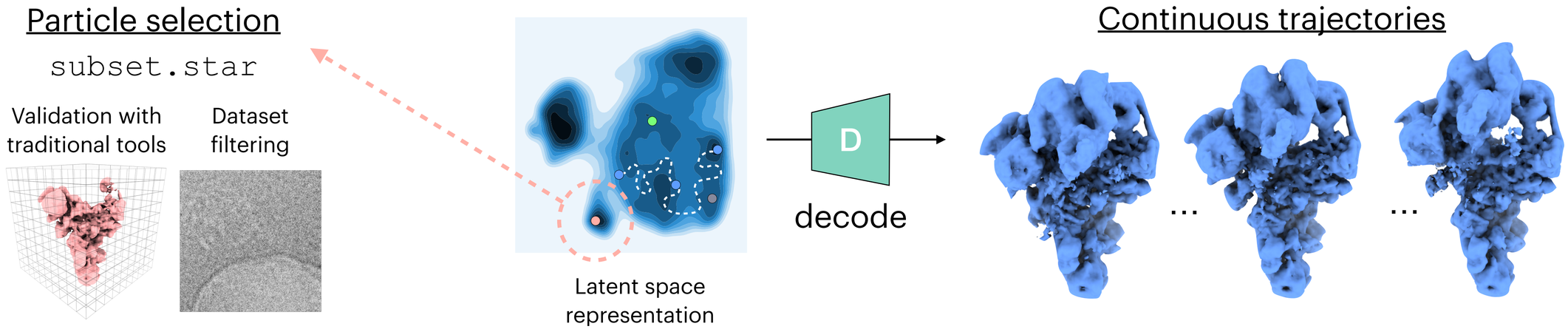
CryoDRGN at test time:

- Use the encoder network to evaluate the latent embedding **z** for each image



View empirical data distribution in latent space

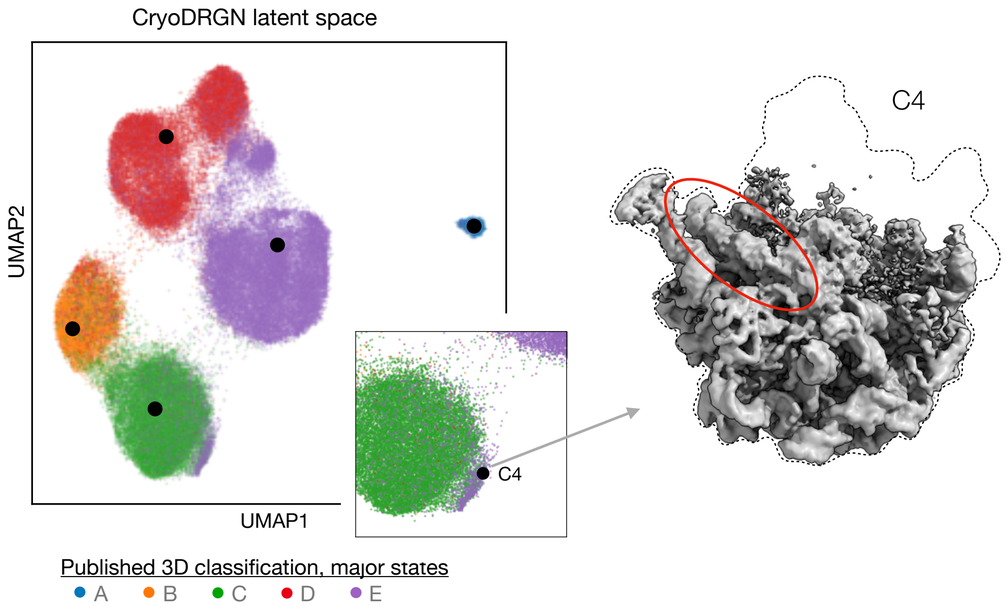
- Use the decoder network to generate **V** at different values of **z**



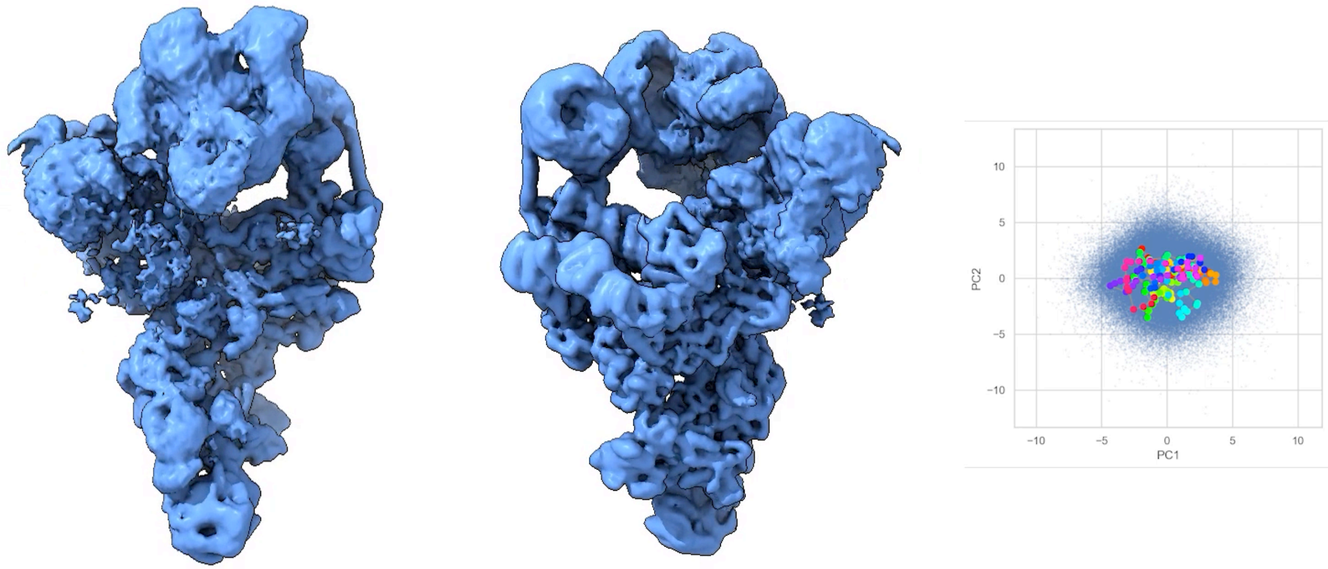
CryoDRGN: Applications and software

- Released as an open source software tool for training cryoDRGN models and interpreting results

Discovery of new structures



Visualization of continuous dynamics



Zhong, Bepler, Berger, Davis, Nature Methods 2021

CryoDRGN software

Software Pipeline

1. Preprocess inputs

```
$ cryodrgn downsample -h
$ cryodrgn parse_ctf_star -h
$ cryodrgn parse_pose_star -h
```

2. Training

```
$ cryodrgn train_vae -h
```

3. Analysis

```
# Analysis pipeline
$ cryodrgn analyze -h

# Making movies
$ cryodrgn pc_traversal -h
$ cryodrgn graph_traversal -h
```

Tutorial

cryoDRGN EMPIAR-10076 tutorial

Now described in: Kinman, Powell, Zhong, Berger, Davis. Nature Protocols 2022.

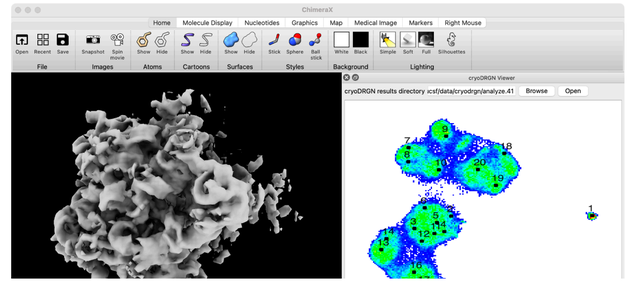
Enhancements and integrations

Suggestions wanted for cryoDRGN map and plot visualization in ChimeraX #134

Open · **tomgoddard** opened this issue 24 days ago · 6 comments

tomgoddard commented 24 days ago

I'm making a cryoDRGN visualization tool in ChimeraX and am interested in any suggestions users have about what it should do. So far it shows the umap plot and the maps computed by "cryodrgn analyze" on as points that plot and you can click on the points to see the map in the 3D view. You can cycle through the precomputed maps with a slider. I think it will be nice to allow computing new maps by clicking a point on the plot and morph between pairs of precomputed maps, and maybe make movies along paths drawn on the plot. Please add comments if you have other suggestions. Thanks!



Assignees
No one—assign your

Labels
None yet

Projects
None yet

Milestone
No milestone

Development
Create a branch for t

Notifications
You're receiving noti
watching this reposit

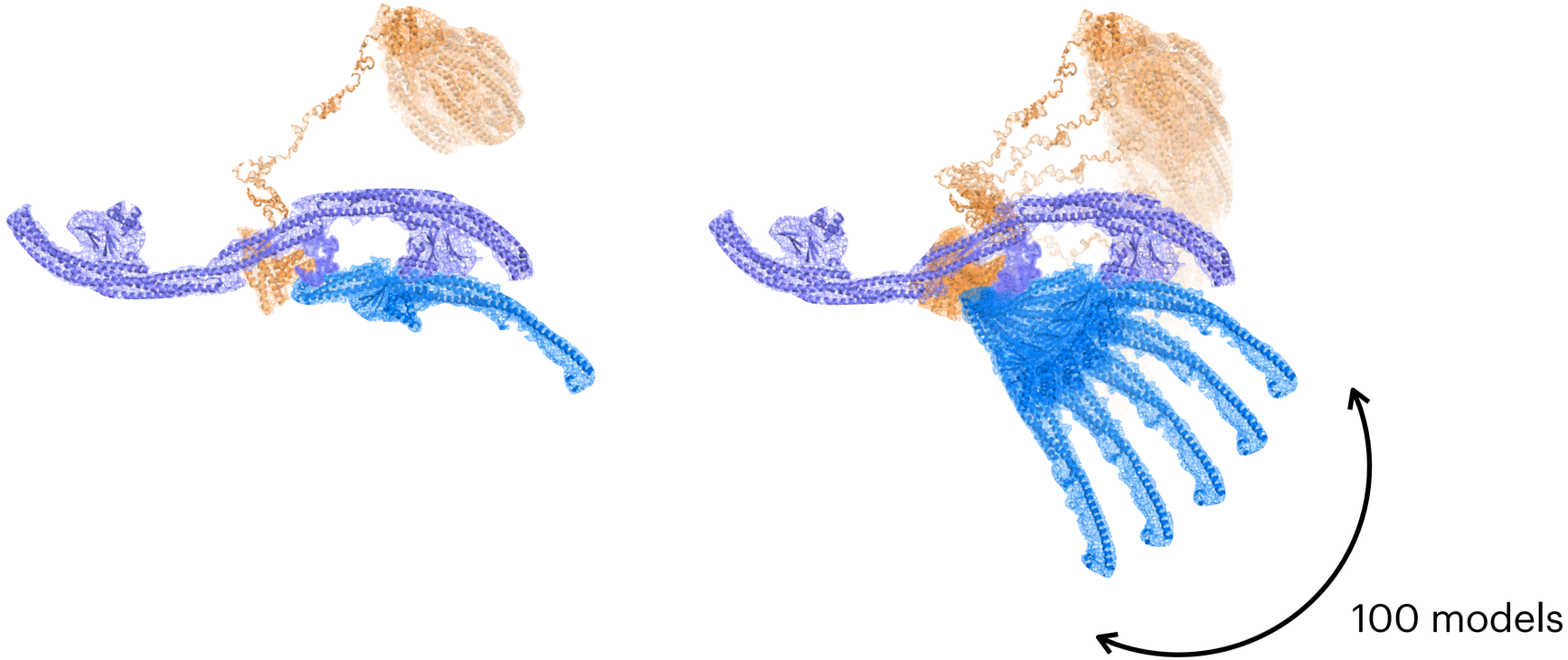
Vineet Bansal
Michal Grzadkowski
Princeton Research Computing

Roadmap

- Motivation and background
- CryoDRGN: Deep Reconstructing Generative Networks
- **Validation on synthetic benchmarks**
- CryoDRGN reconstructions of real data
- Future vision

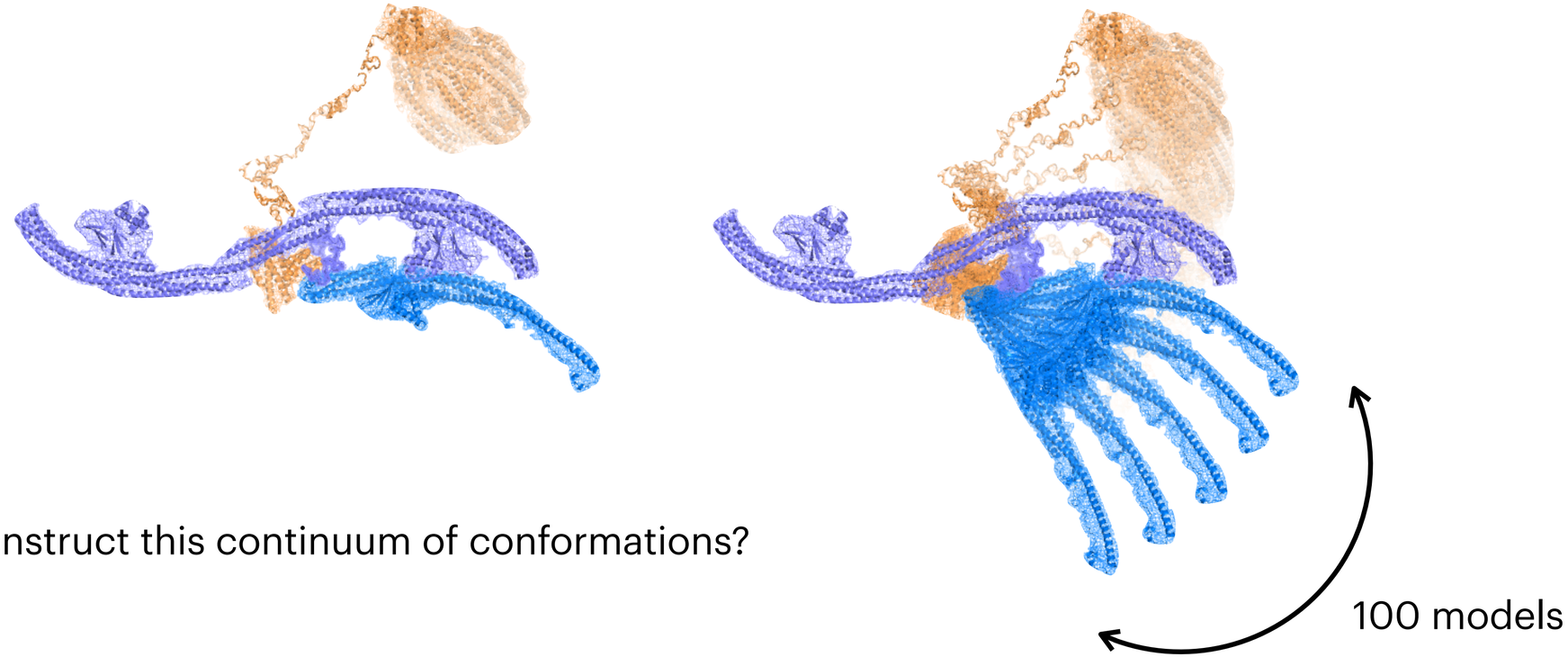
Heterogeneous reconstruction of a model protein complex containing 1 degree of freedom

- We generate a model protein complex containing one continuous degree of freedom
  - 100 atomic models varying one dihedral angle
  - 500 randomly oriented projections of each model, yielding a total of 50k projections



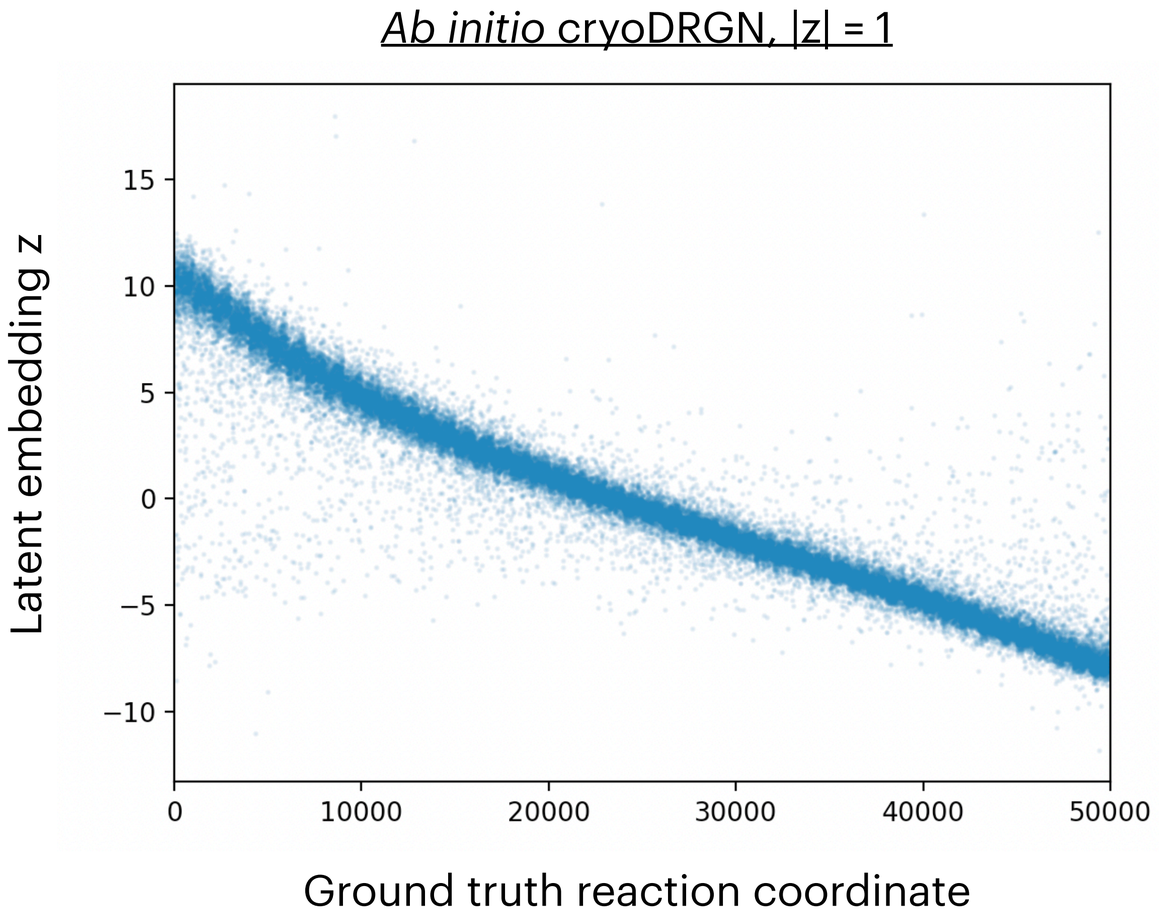
Heterogeneous reconstruction of a model protein complex containing 1 degree of freedom

- We generate a model protein complex containing one continuous degree of freedom
  - 100 atomic models varying one dihedral angle
  - 500 randomly oriented projections of each model, yielding a total of 50k projections



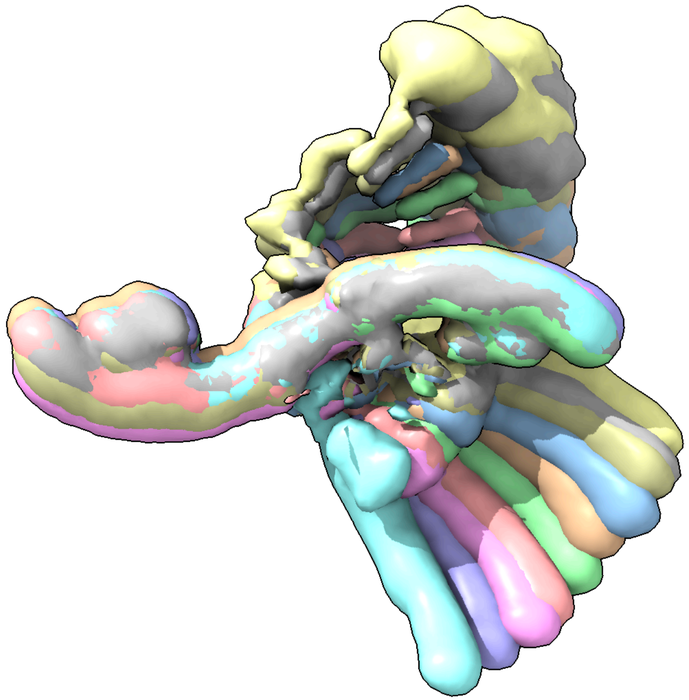
Can we reconstruct this continuum of conformations?

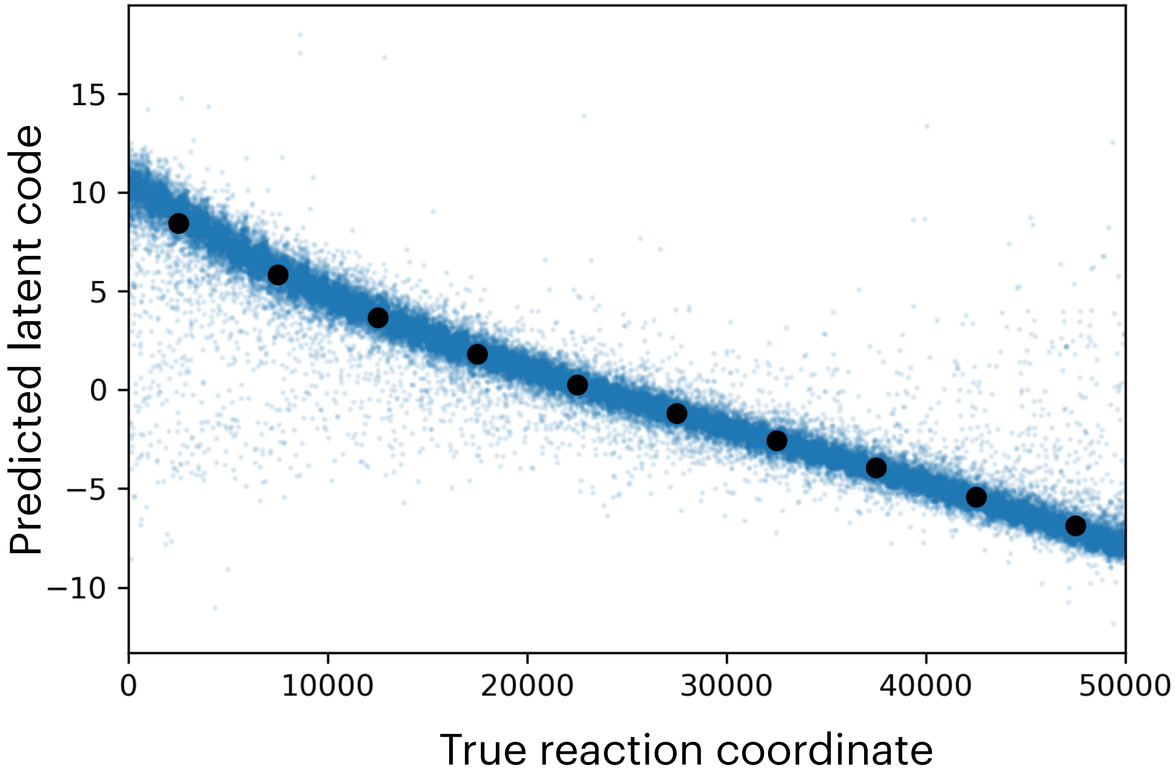
The predicted latent code correlates with the true reaction coordinate



CryoDRGN can reconstruct a continuum of structures along the true reaction coordinate

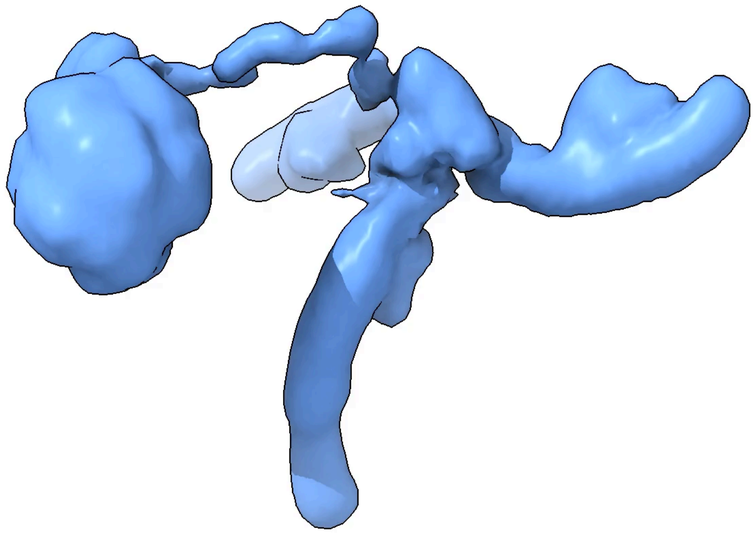
Ab initio cryoDRGN
10 structures generated from latent representation

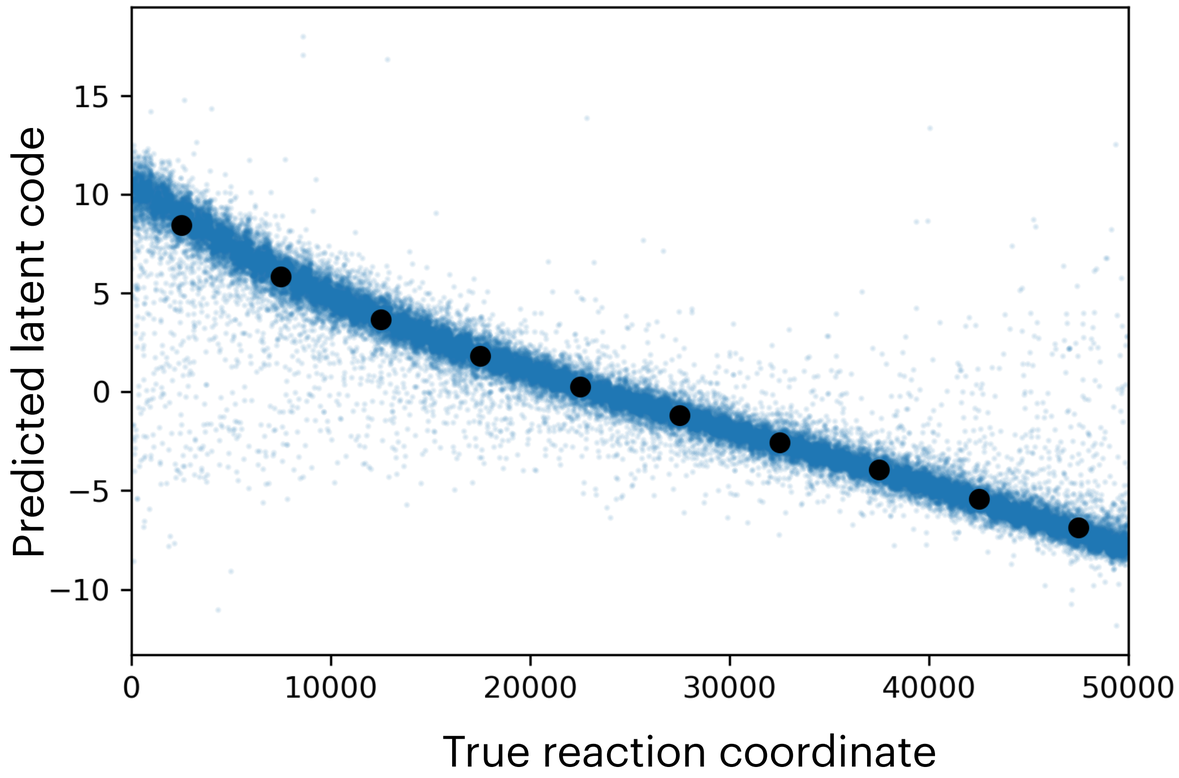




CryoDRGN can reconstruct a continuum of structures along the true reaction coordinate

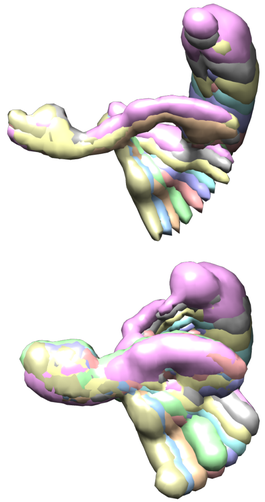
Ab initio cryoDRGN
10 structures generated from latent representation

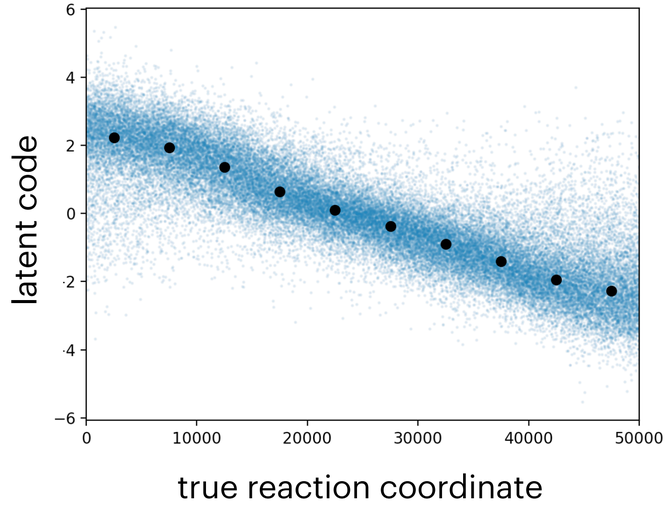




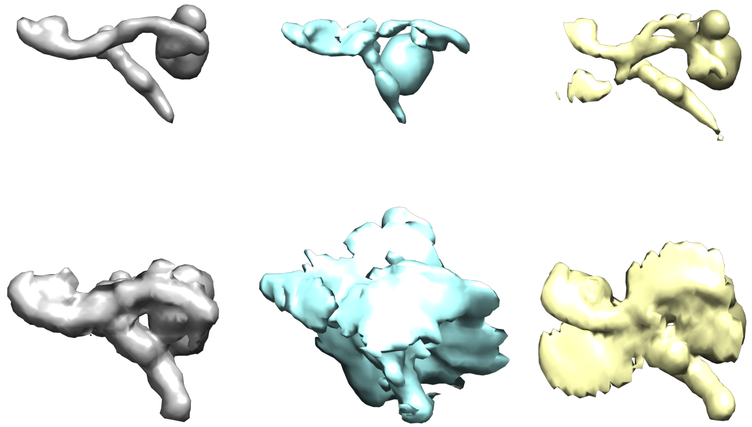
CryoDRGN can reconstruct a continuum of structures along the true reaction coordinate

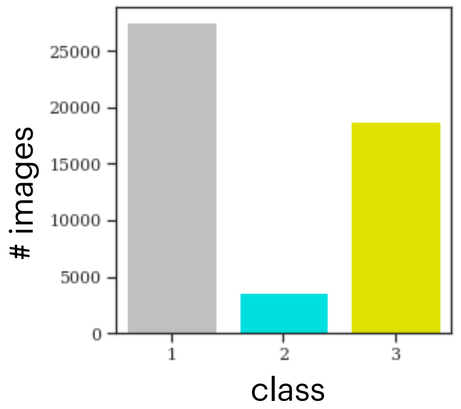
Ab initio cryoDRGN
10 structures generated from latent representation





cryoSPARC discrete multiclass reconstruction
K=3 classes

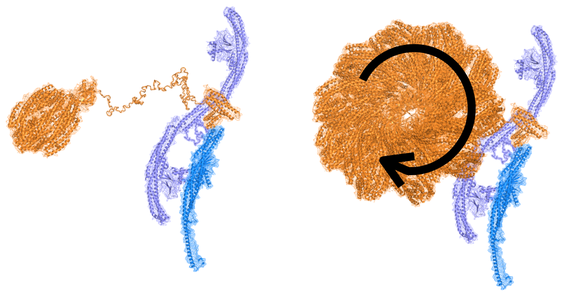




Zhong et al. ICLR 2020 Spotlight

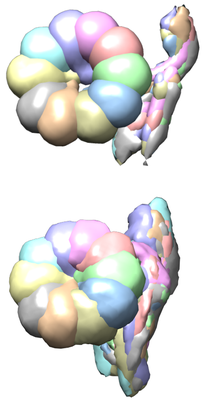
Heterogeneous reconstruction of a model protein complex with continuous motions

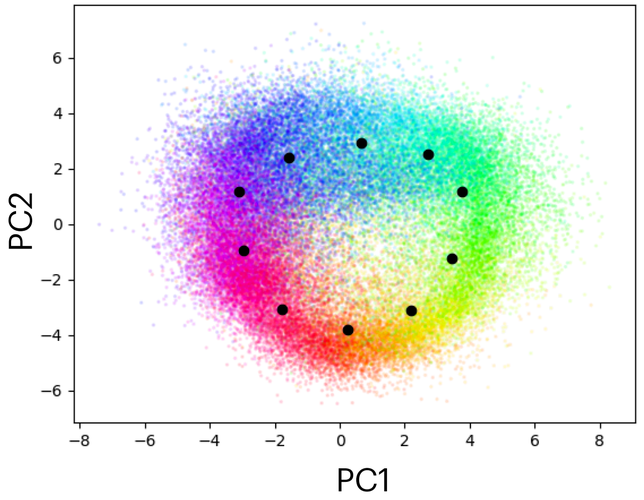
Ground truth



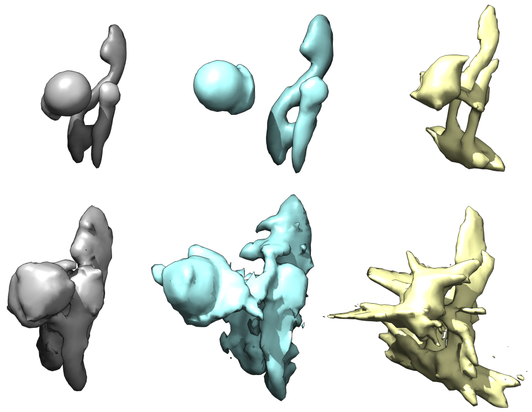
- 100 atomic models varying one dihedral angle
- 500 randomly oriented projections of each model
- 50k projection images

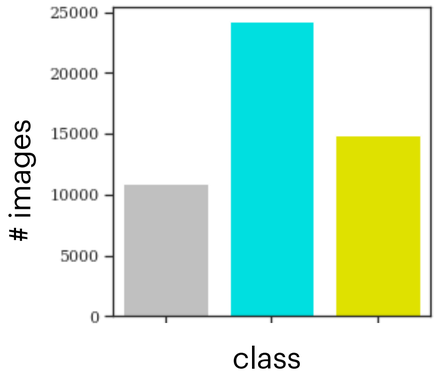
Ab initio cryoDRGN, |z| = 10
10 structures sampled along latent space





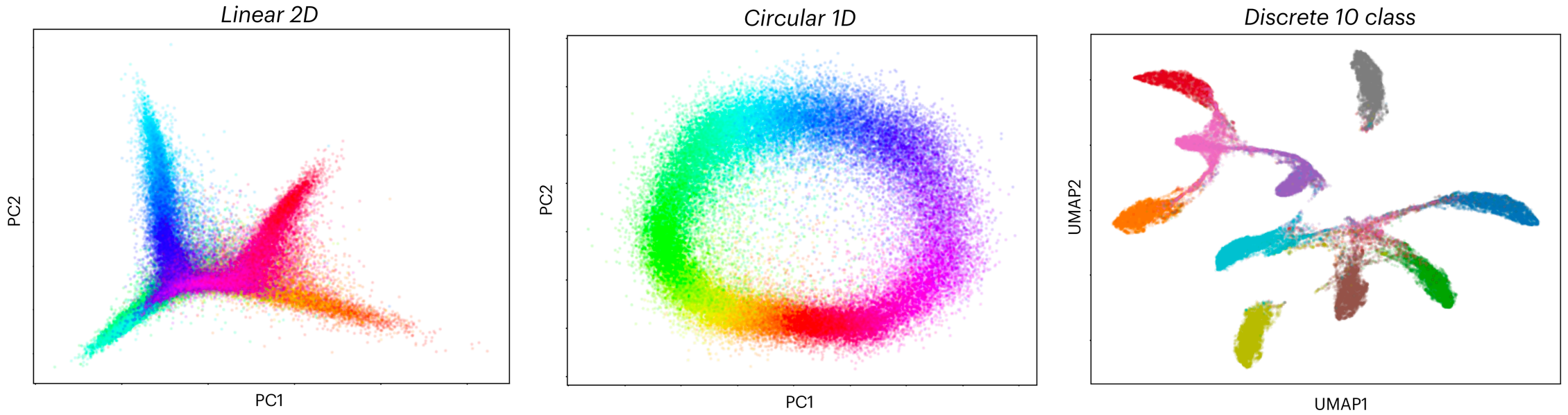
cryoSPARC discrete multiclass reconstruction
K=3 classes





Zhong et al. ICLR 2020 Spotlight

Additional datasets with more complex latent structure



Reconstruction accuracy quantified by an FSC=0.5 resolution metric between predicted and ground truth volume
(Lower is better; best possible is 2 pixels)

Dataset	*cryoDRGN*	*cryoDRGN + tilt pairs*	*cryoSPARC*
Linear 1D motion	**2.50**	**2.35**	3.60
Linear 2D motion	**4.44**	**2.93**	6.90
Circular 1D motion	**3.86**	**2.63**	4.87
Discrete 10 class	**4.95**	**2.58**	5.69

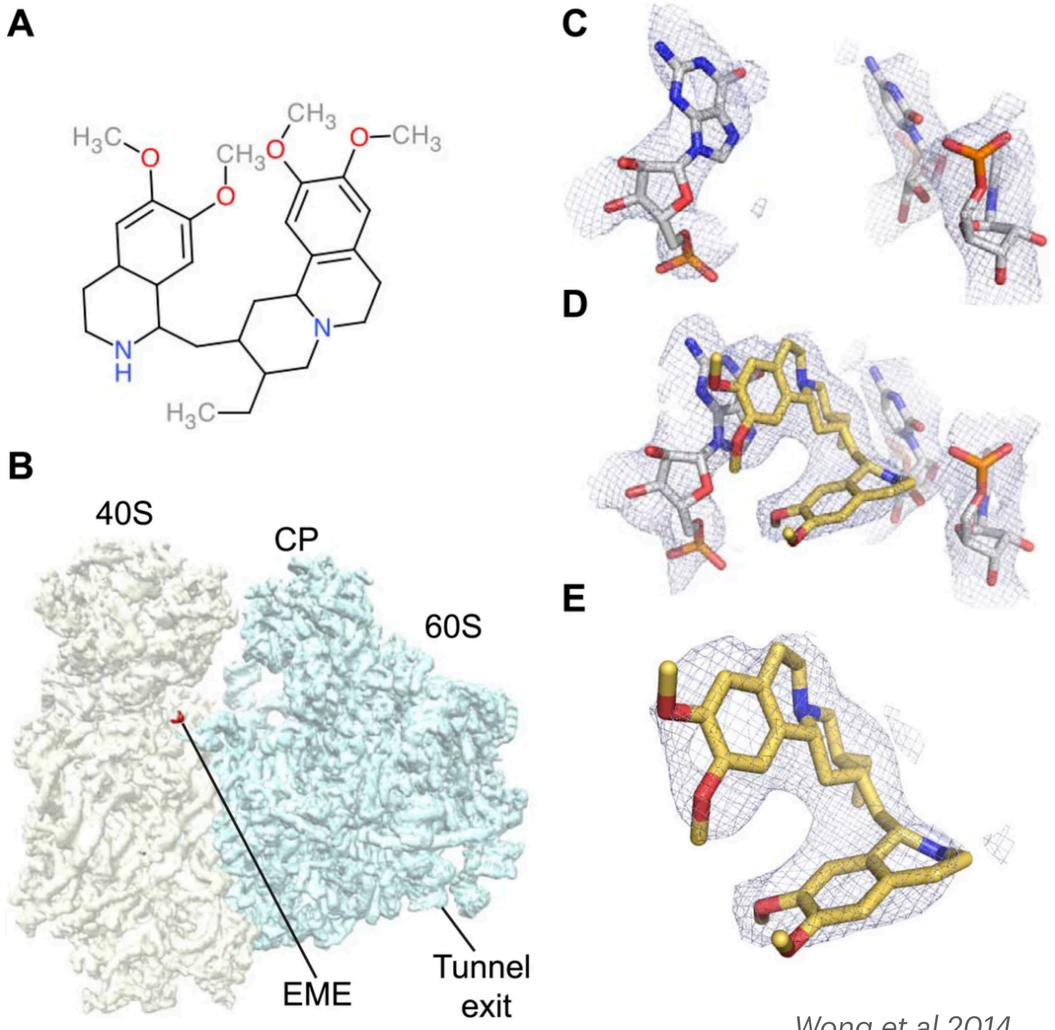
Zhong et al. ICLR 2020 Spotlight

Roadmap

- Motivation and background
- CryoDRGN: Deep Reconstructing Generative Networks
- Validation on synthetic benchmarks
- **CryoDRGN reconstructions of real data**
  - **Uncovering residual heterogeneity in high resolution “homogeneous” datasets**
  - Discovering new states of the assembling ribosome
  - Reconstructing continuous motions of the pre-catalytic spliceosome
- Future vision

Homogeneous reconstruction of the *Pf*80S ribosome bound to the anti-protozoan drug emetine

- 3.2 Å homogeneous reconstruction from 105k cryo-EM images (EMPIAR-10028)
- Difference map between structures with and without EME identified its binding site
- Lower local resolution in head group of small subunit and peripheral regions



eLIFE
elifesciences.org

RESEARCH ARTICLE

Cryo-EM structure of the *Plasmodium falciparum* 80S ribosome bound to the anti-protozoan drug emetine

Wilson Wong[1,2†], Xiao-chen Bai[3†], Alan Brown[3†], Israel S Fernandez[3], Eric Hanssen[4], Melanie Condron[1,2], Yan Hong Tan[1,2], Jake Baum[1,2*‡], Sjors HW Scheres[3*]

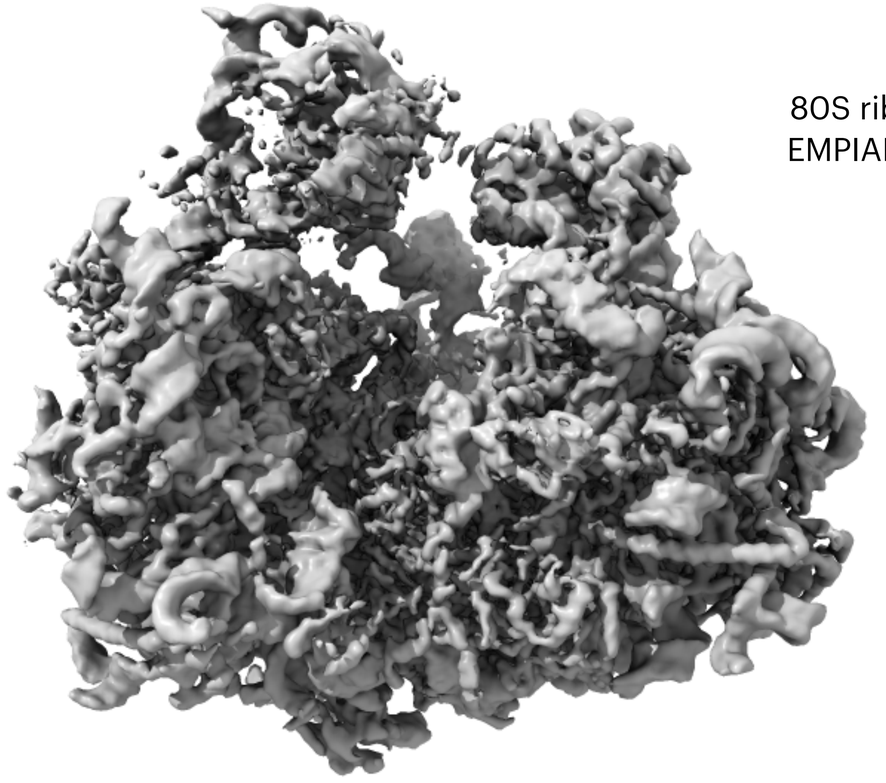
Wong et al 2014

CryoDRGN's neural model can learn high resolution cryo-EM density maps

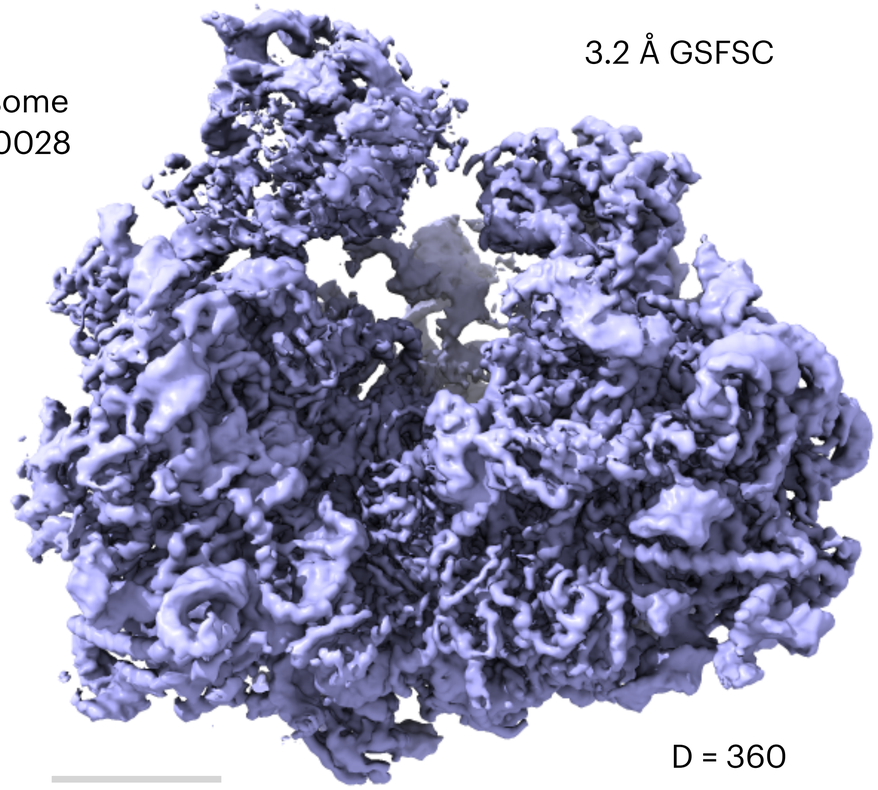
- Train the cryoDRGN decoder (with no latent variable input) on images from EMPIAR-10028

Neural network representation

Voxel-based representation



80S ribosome
EMPIAR 10028



3.2 Å GSFSC

D = 360

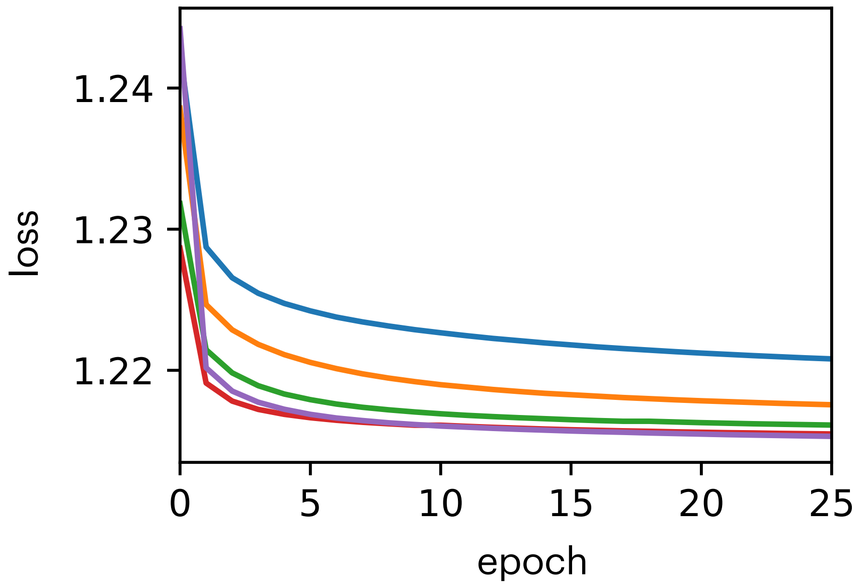
50 Å

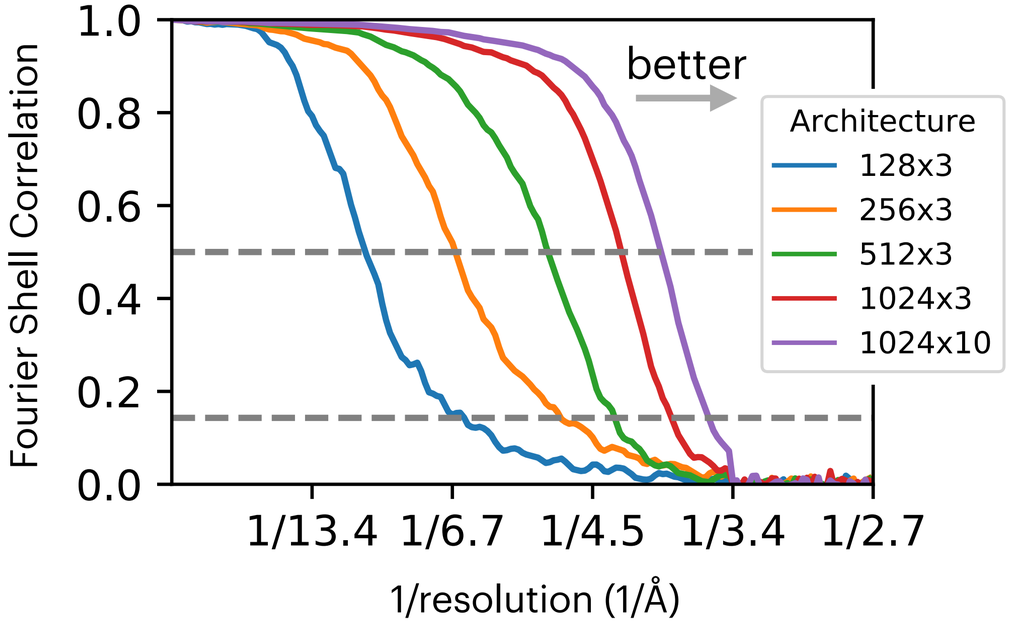
1024x10 architecture
50 epochs

Zhong et al., Nature Methods 2021

Achievable resolution is bounded by image size and model capacity

- The representation capacity of a cryoDRGN model is affected by:
  - Architecture
  - Fourier featurization, see Tancik et al. NeurIPS 2020
  - Latent variable dimension (for heterogeneous reconstruction)
- Inverse tradeoff between architecture size and training speed



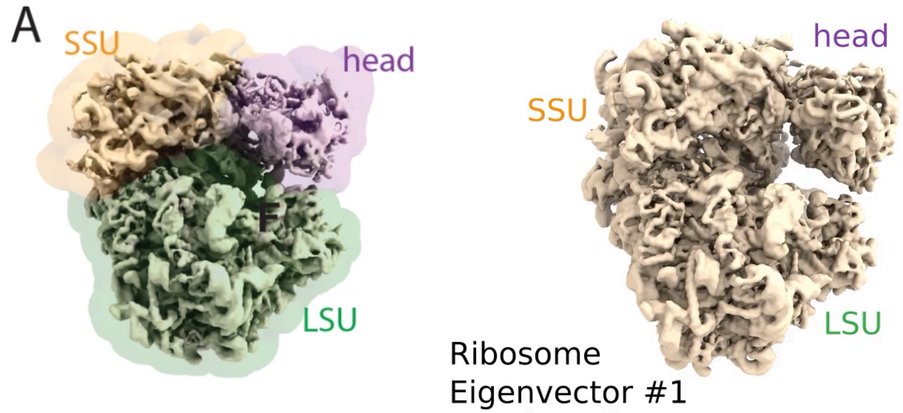


Zhong et al., Nature Methods 2021

Advanced methods for heterogeneity analysis

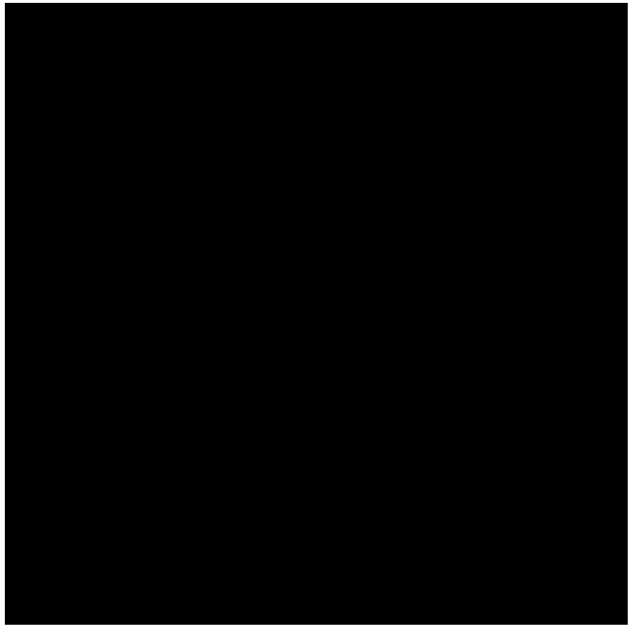
- Multi-body analysis[1]: Motions between *B* rigid bodies

$$Image\ X_i = \mathrm{CTF}_i \left(\sum_{b=1}^{B} \mathbf{P}_{\phi_b} V_b \right) + N_i \,,$$



- cryoSPARC 3DVA[2]: Linear interpolations between “eigenvolumes"

$$\begin{aligned} Image\ X_i &= \alpha_i C_i P(\phi_i) \mathcal{V}(z_i) + \eta \\ &= \alpha_i C_i P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) + \eta \end{aligned}$$

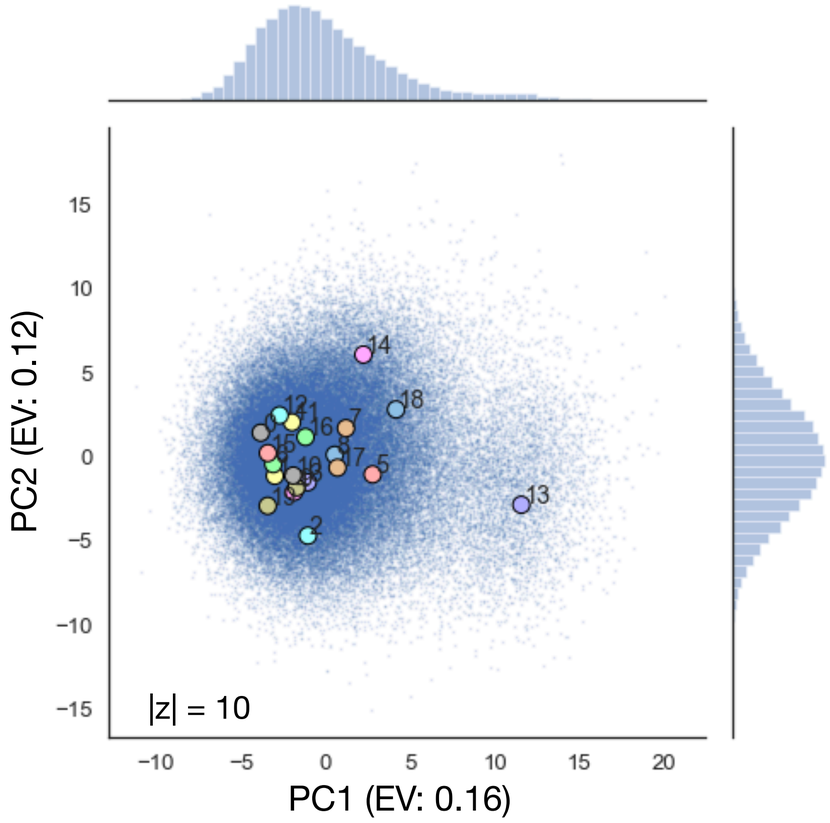


https://cryosparc.com/docs/tutorials/3d-variability-analysis/

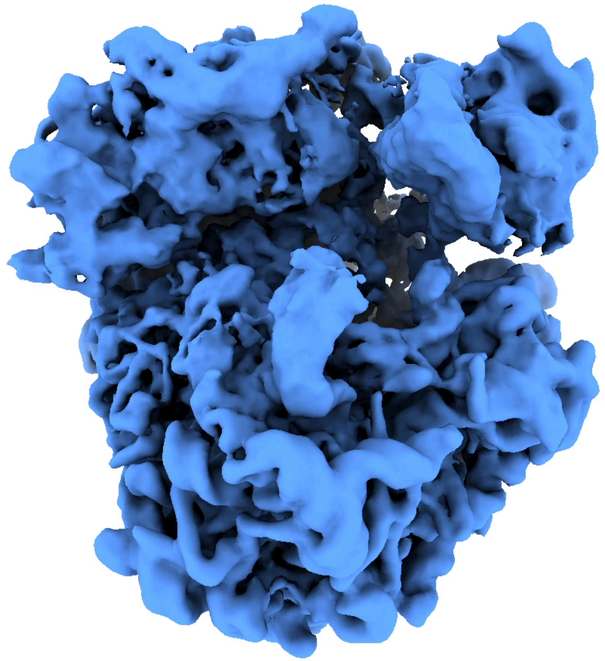
1 Nakane et al. eLife 2018; 2 Punjani & Fleet, JSB 2021

Discovering residual heterogeneity of the *Pf*80S ribosome [EMPIAR-10028]

CryoDRGN latent space:



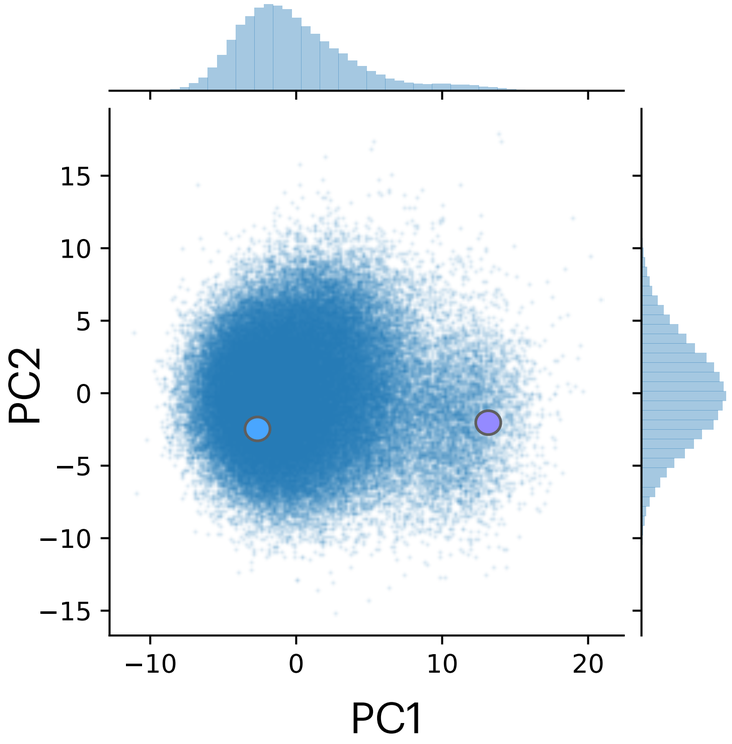
20 sampled structures:



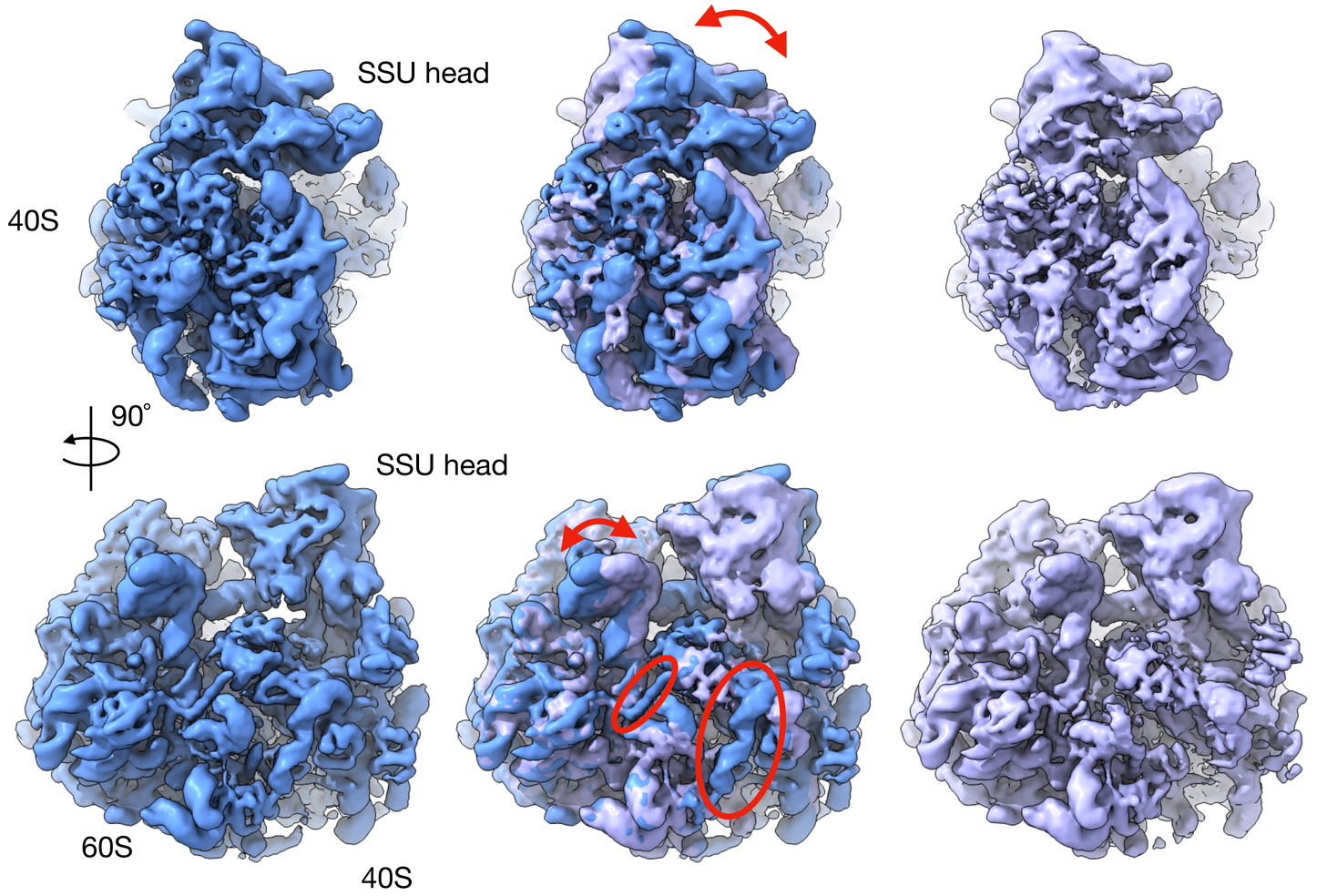
- Subset of images separated by PC1 correspond to the 40S subunit in a rotated state
- Many heterogeneous elements in the large and small subunit

Zhong et al., Nature Methods 2021

Variation in 40S SSU is consistent with other methods for heterogeneity analysis



CryoDRGN, comparison of 2 volumes



Roadmap

- Motivation and background
- CryoDRGN: Deep Reconstructing Generative Networks
- Validation on synthetic benchmarks
- **CryoDRGN reconstructions of real data**
  - Uncovering residual heterogeneity in high resolution “homogeneous” datasets
  - **Discovering new states of the assembling ribosome**
  - Reconstructing continuous motions of the pre-catalytic spliceosome
- Future vision

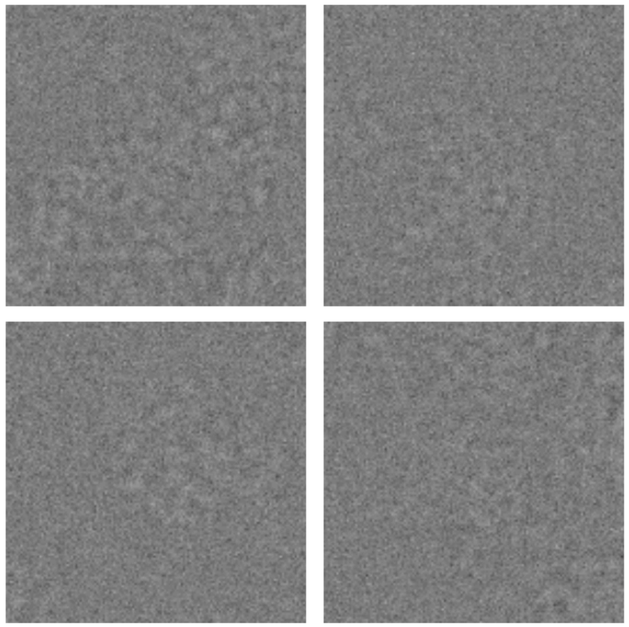
Learning ribosome assembly landscapes [EMPIAR-10076]

Modular assembly of the bacterial large ribosomal subunit (LSU)

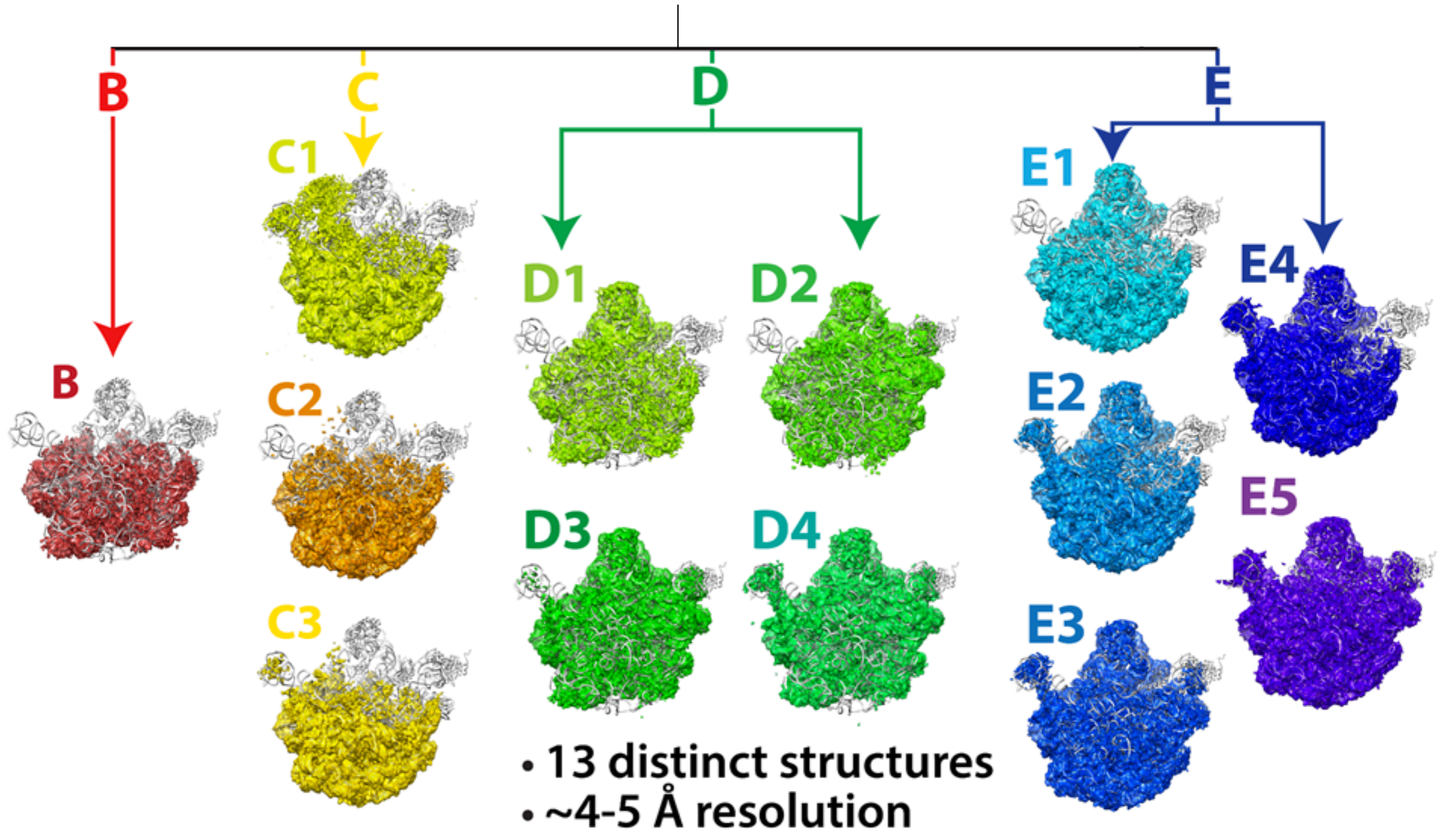
Dataset: 131k cryo-EM images of a mixture of LSU assembly intermediates

4 major and 13 minor states of the LSU identified from hierarchical multiclass reconstruction

Example images



[EMPIAR-10076]



Davis et al. Cell 2016

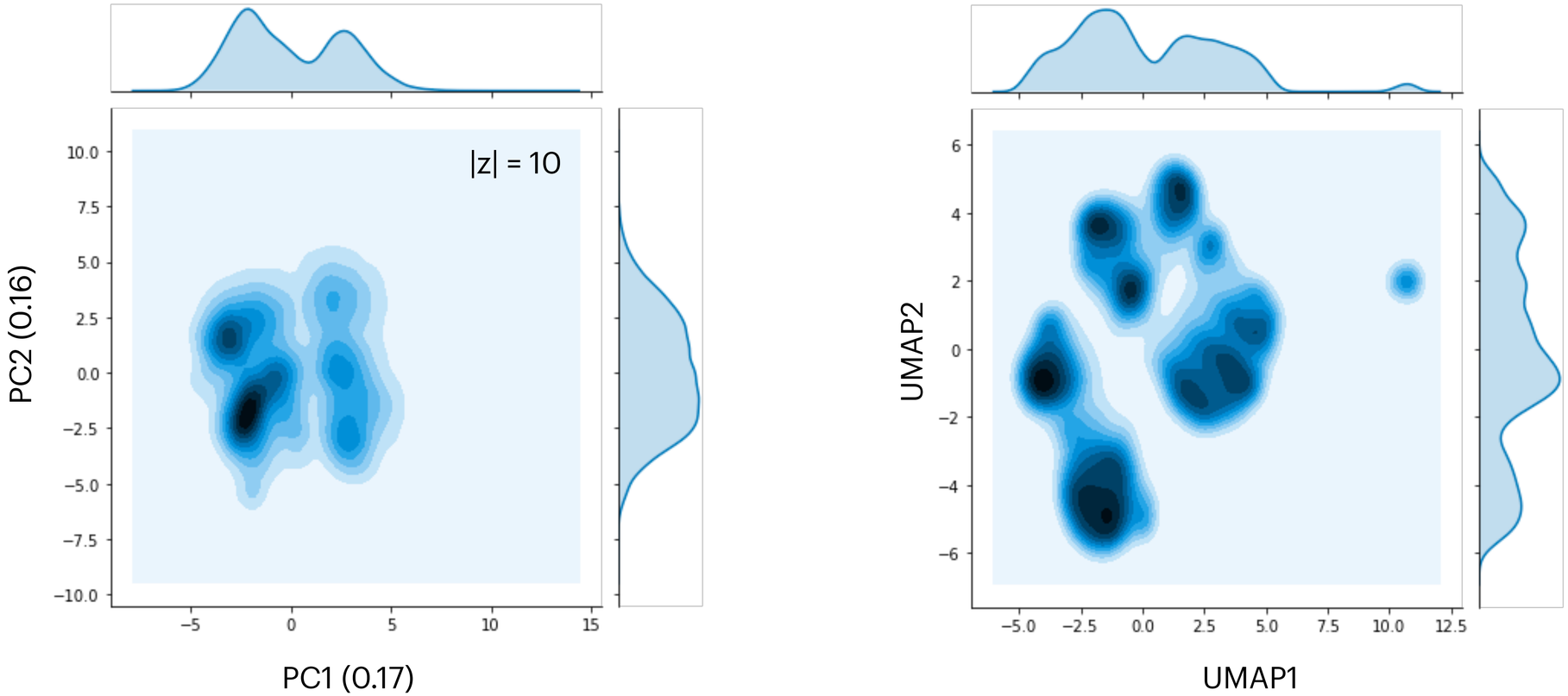
Learning ribosome assembly landscapes [EMPIAR-10076]

Modular assembly of the bacterial large ribosomal subunit (LSU)

Dataset: 131k cryo-EM images of a mixture of LSU assembly intermediates

4 major and 13 minor states of the LSU identified from hierarchical multiclass reconstruction

Latent embeddings from cryoDRGN

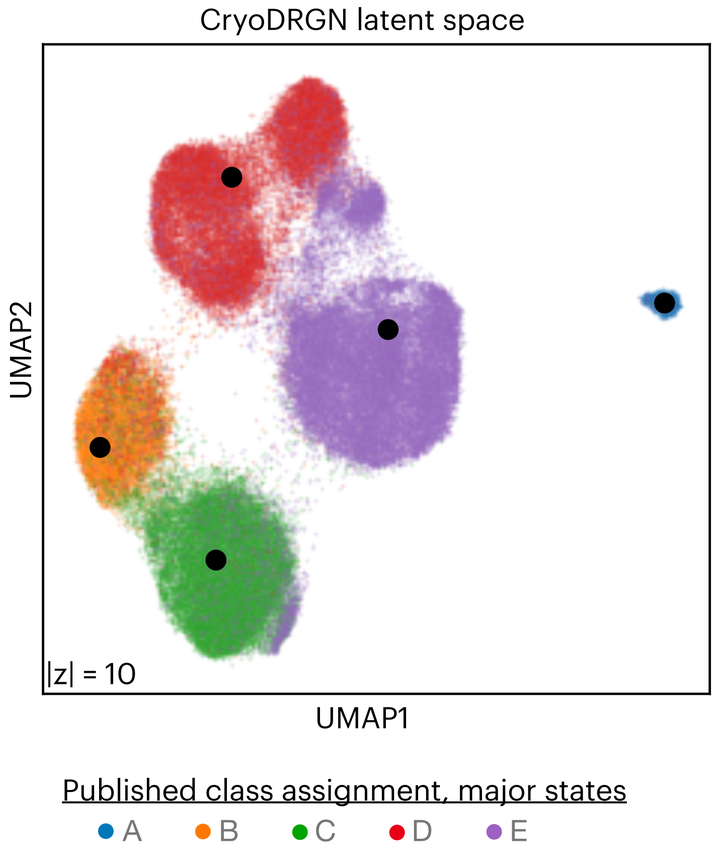


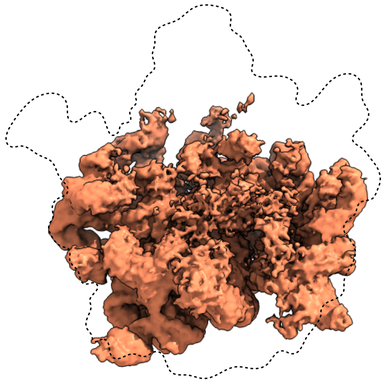
Learning ribosome assembly landscapes [EMPIAR-10076]

Modular assembly of the bacterial large ribosomal subunit (LSU)

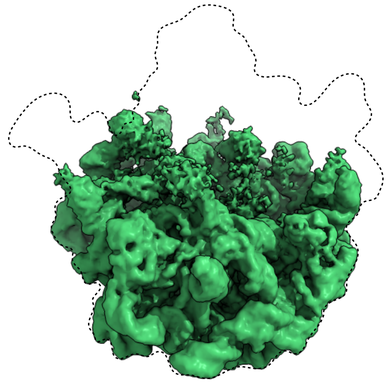
Dataset: 131k cryo-EM images of a mixture of LSU assembly intermediates

4 major and 13 minor states of the LSU identified from hierarchical multiclass reconstruction

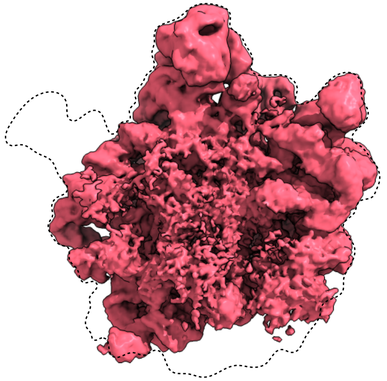




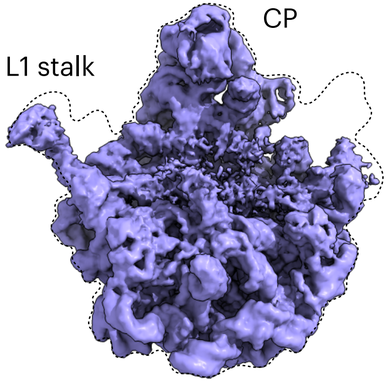
LSU assembly class B



LSU assembly class C

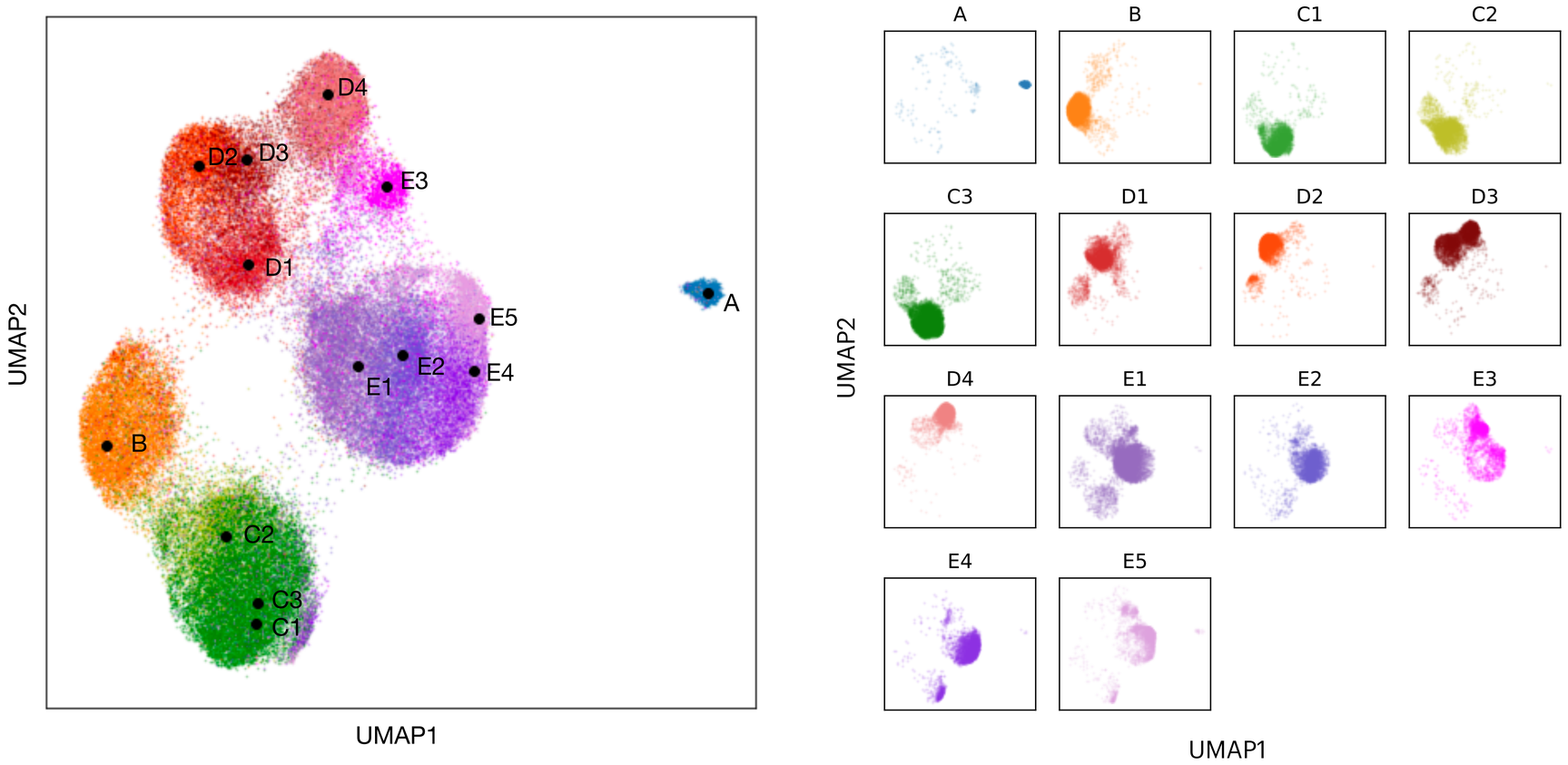


LSU assembly class D



LSU assembly class E

Clusters in the latent space vs. expert-driven hierarchical classification

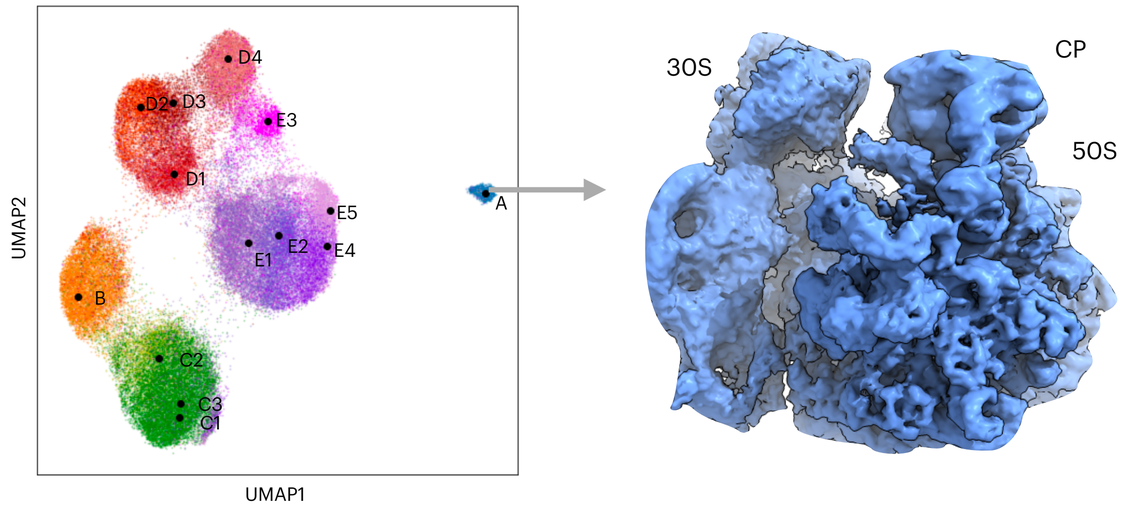


Published 3D classification, minor states

A B C1 C2 C3 D1 D2

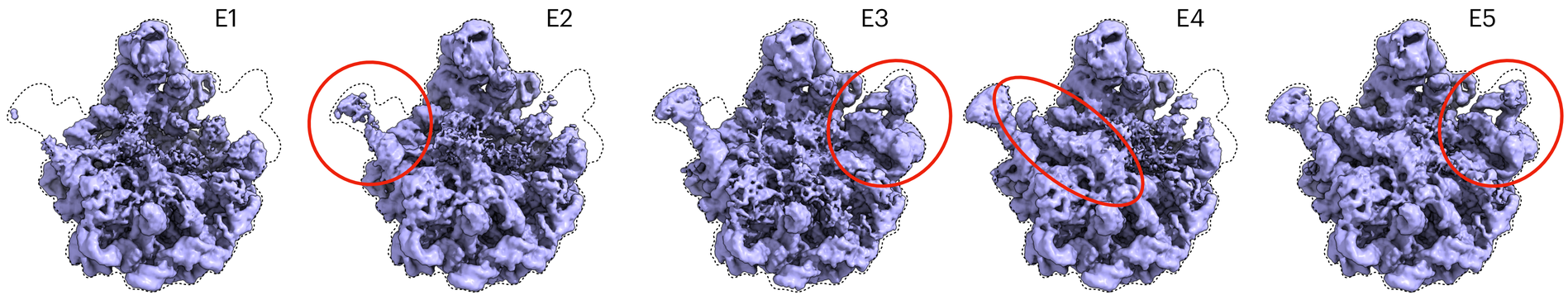
D3 D4 E1 E2 E3 E4 E5

Additional samples from the latent space

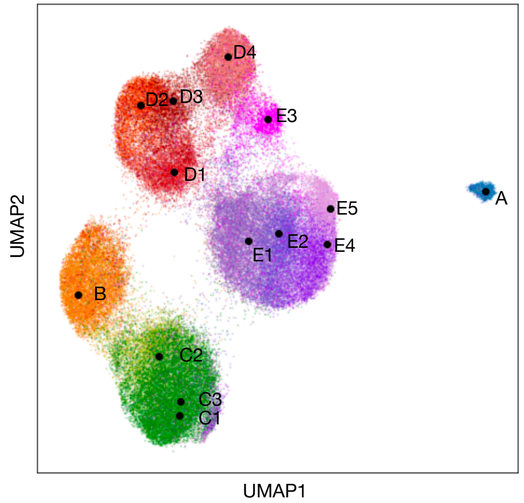


the 70S ribosome

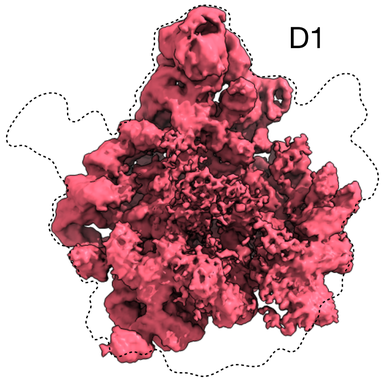
E class minor assembly states

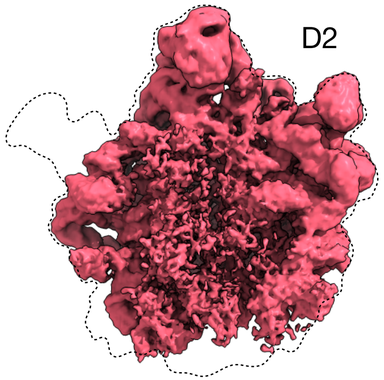


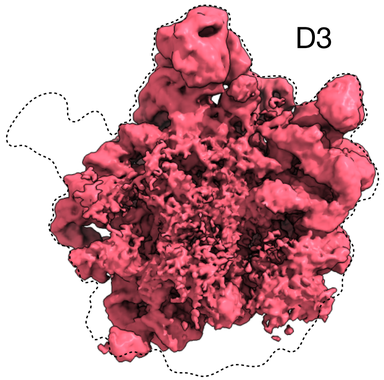
Additional samples from the latent space

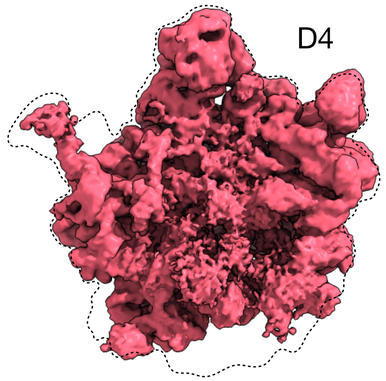


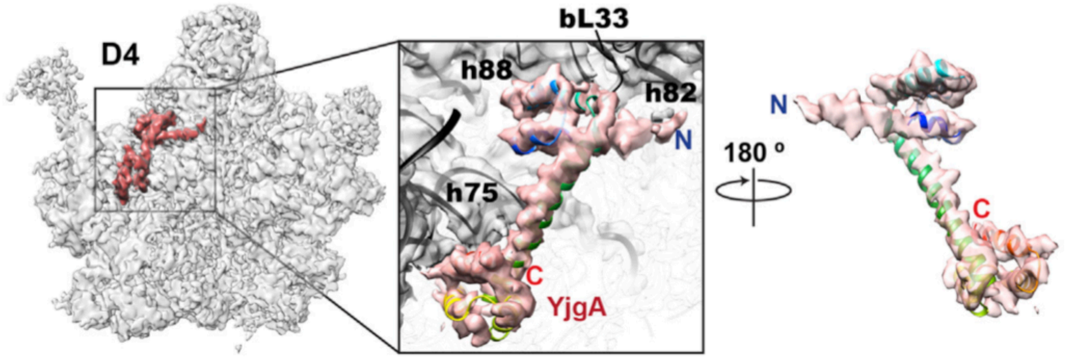
D class minor assembly states





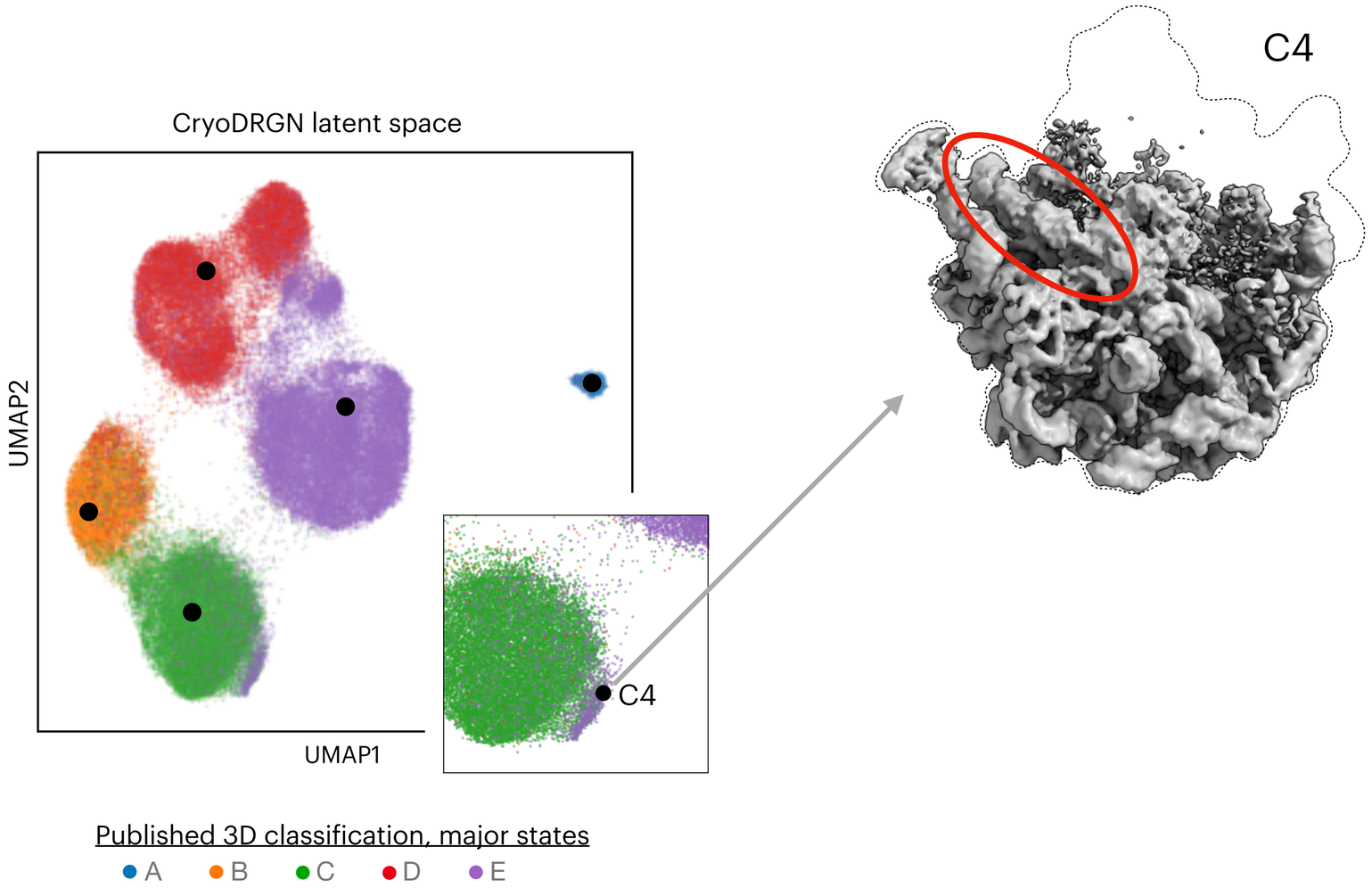




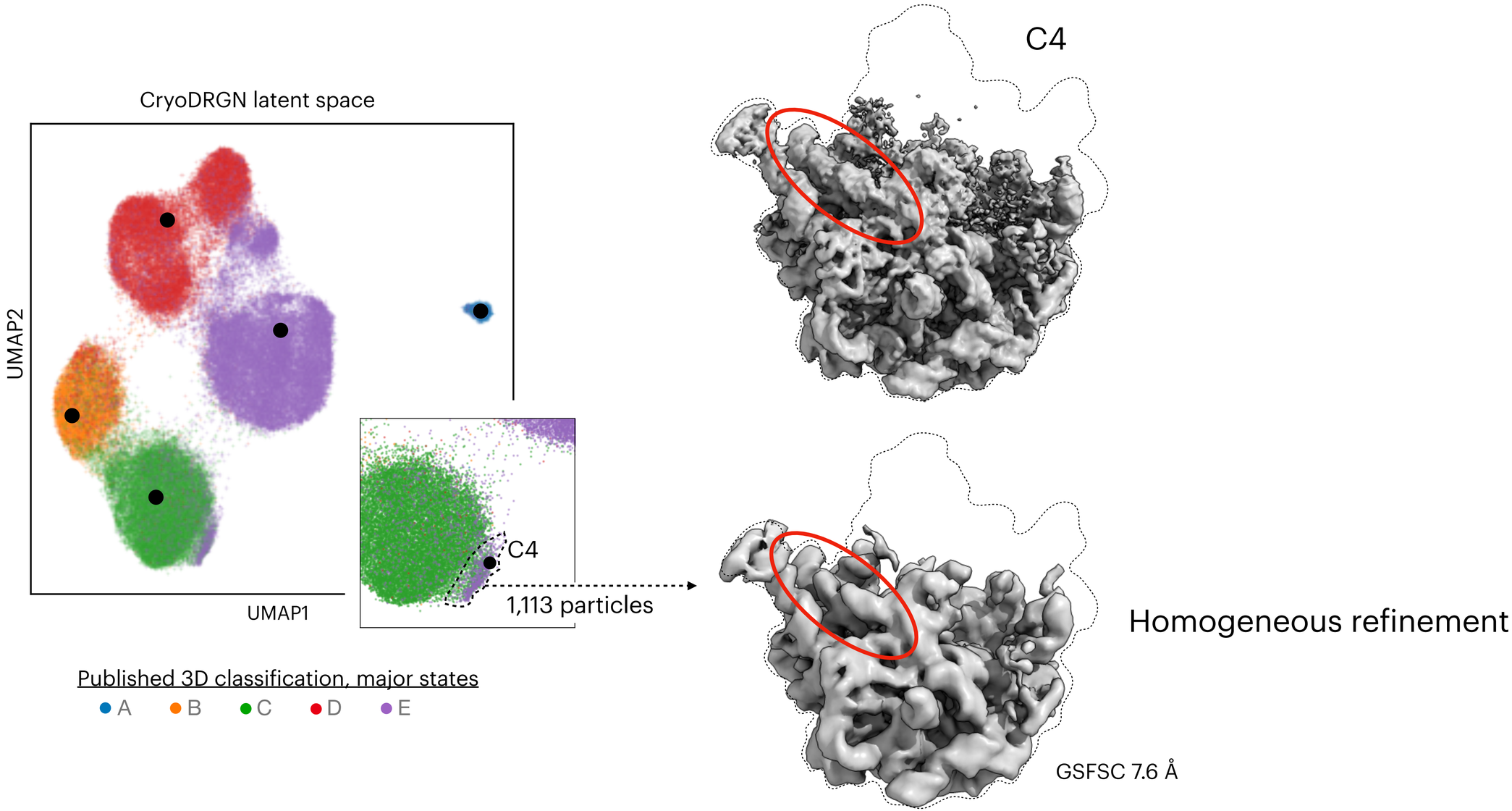


Adapted from Figure 5, *Davis et al 2016*

Discovery of a new assembly state, C4



Discovery of a new assembly state, C4

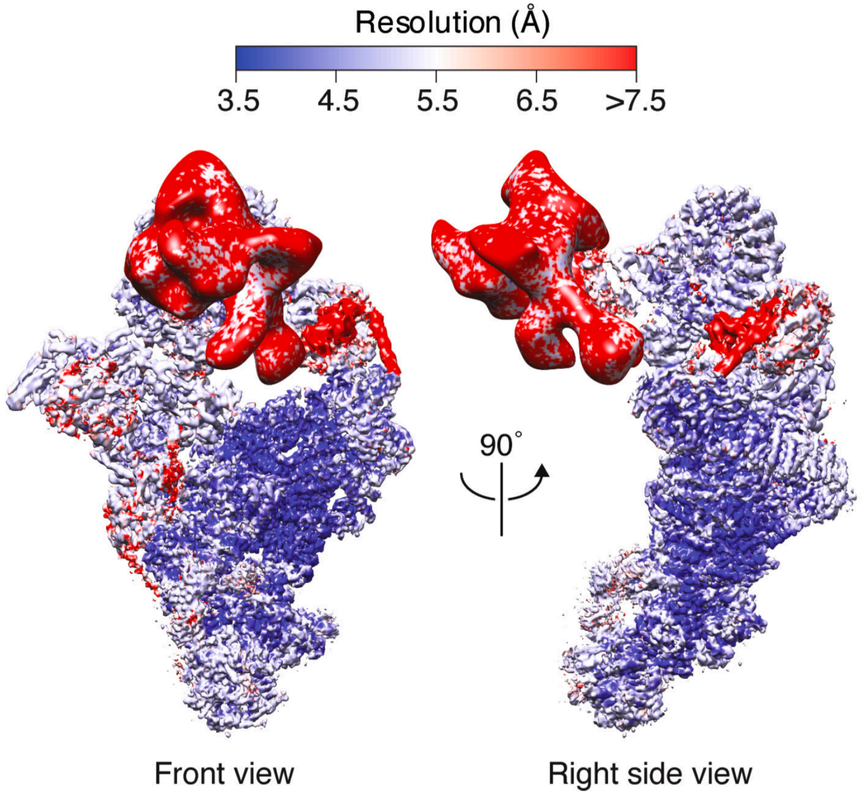


Roadmap

- Motivation and background
- CryoDRGN: Deep Reconstructing Generative Networks
- Validation on synthetic benchmarks
- **CryoDRGN reconstructions of real data**
  - Uncovering residual heterogeneity in high resolution “homogeneous” datasets
  - Discovering new states of the assembling ribosome
  - **Reconstructing continuous motions of the pre-catalytic spliceosome**
- Future vision

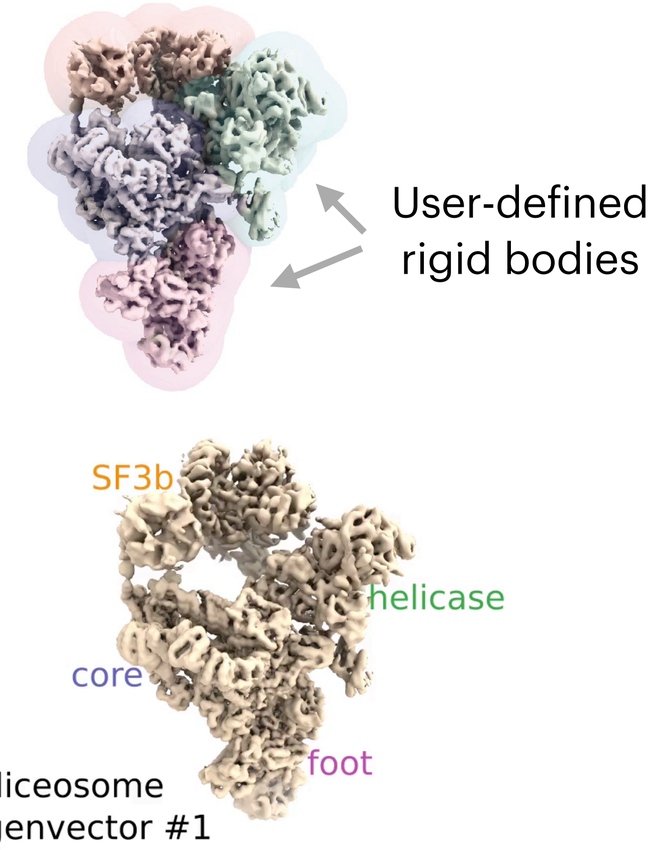
Structure of the pre-catalytic spliceosome

Sub-complexes resolved separately through many rounds of focused classification



Plaschka, Lin, & Nagai. Nature 2017

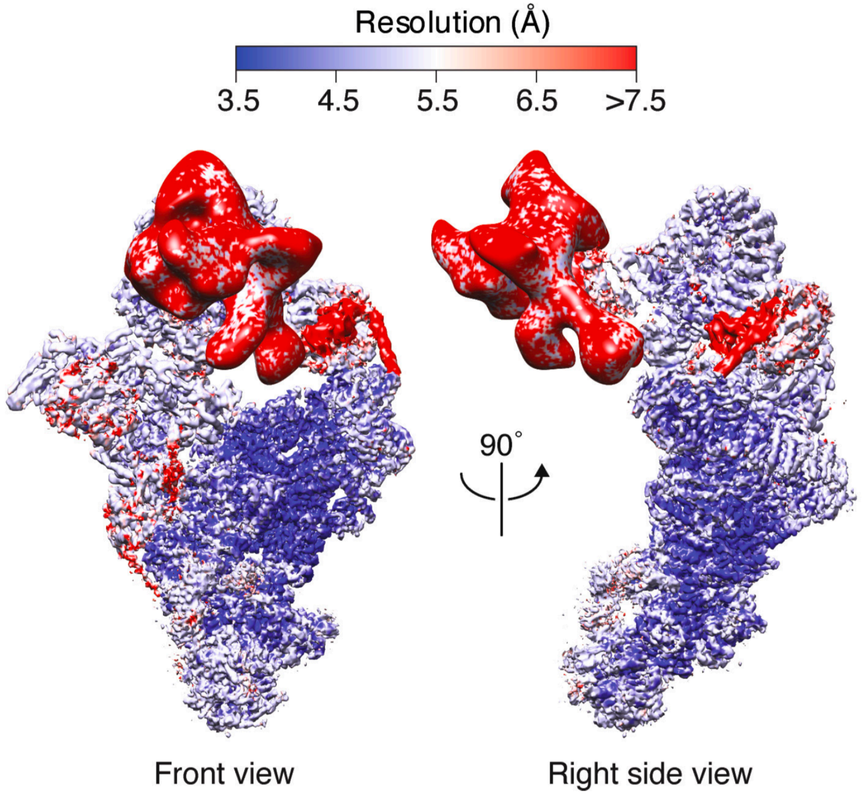
Multibody refinement



Nakane et al. eLife 2018

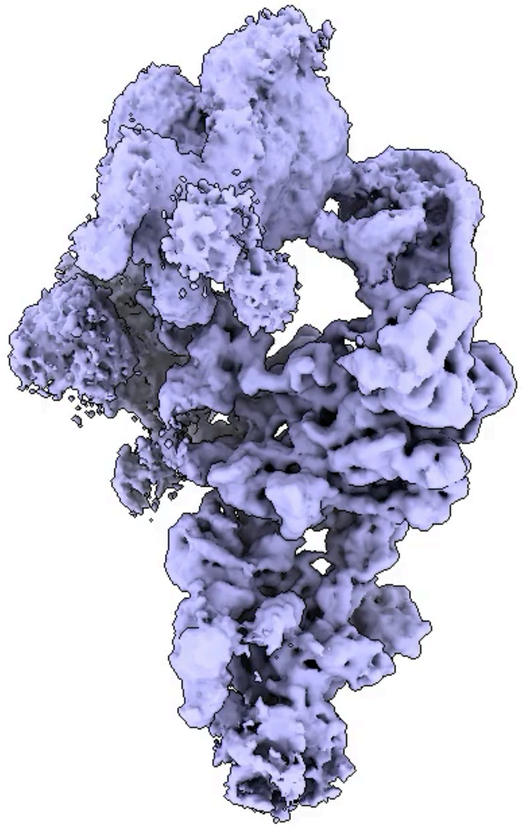
Structure of the pre-catalytic spliceosome

Sub-complexes resolved separately through many rounds of focused classification



Plaschka, Lin, & Nagai. Nature 2017

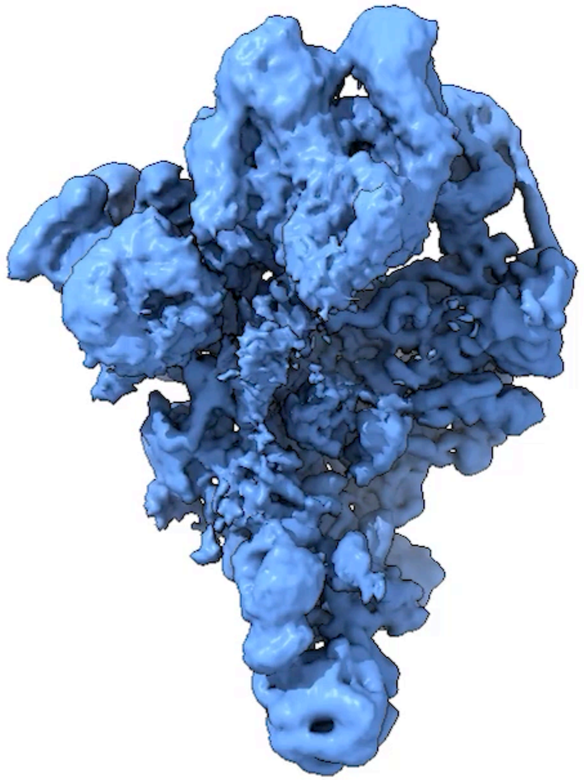
cryoSPARC 3DVA

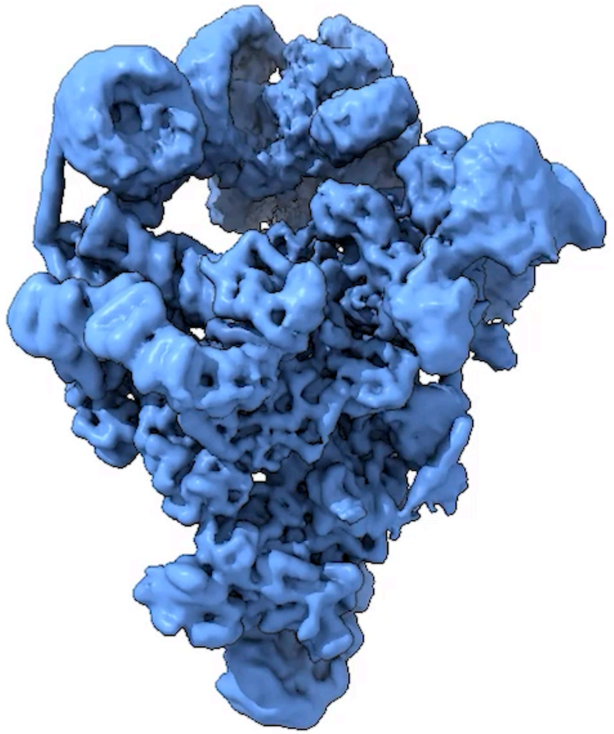


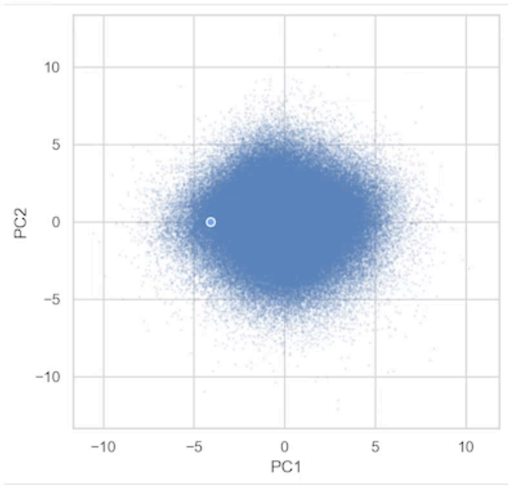
Punjani & Fleet. JSB 2021

Reconstructing continuous motions of the pre-catalytic spliceosome [EMPIAR-10180]

Trajectories along principle component axis of the latent space show variability within dataset



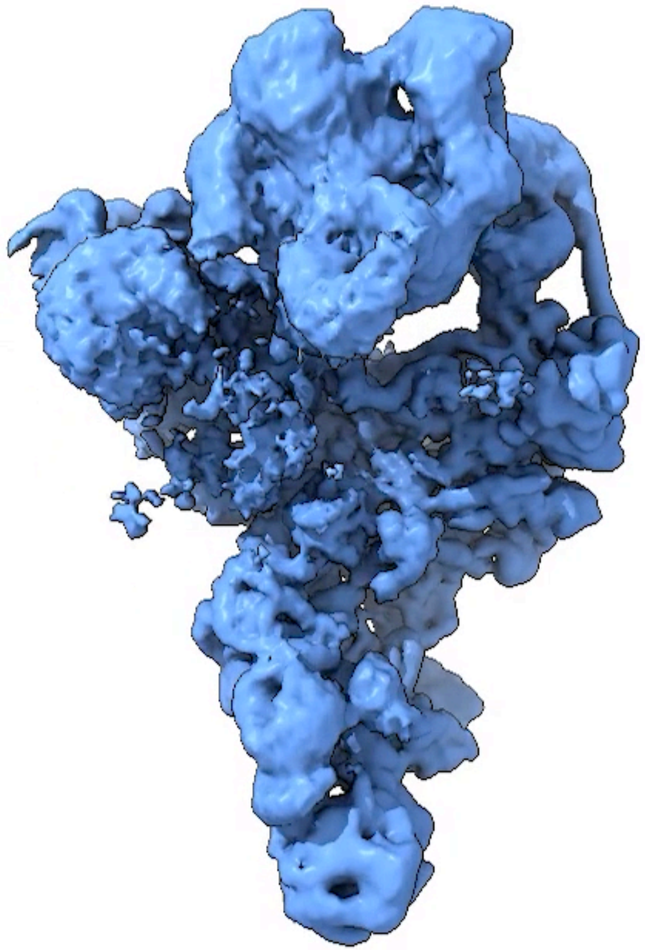


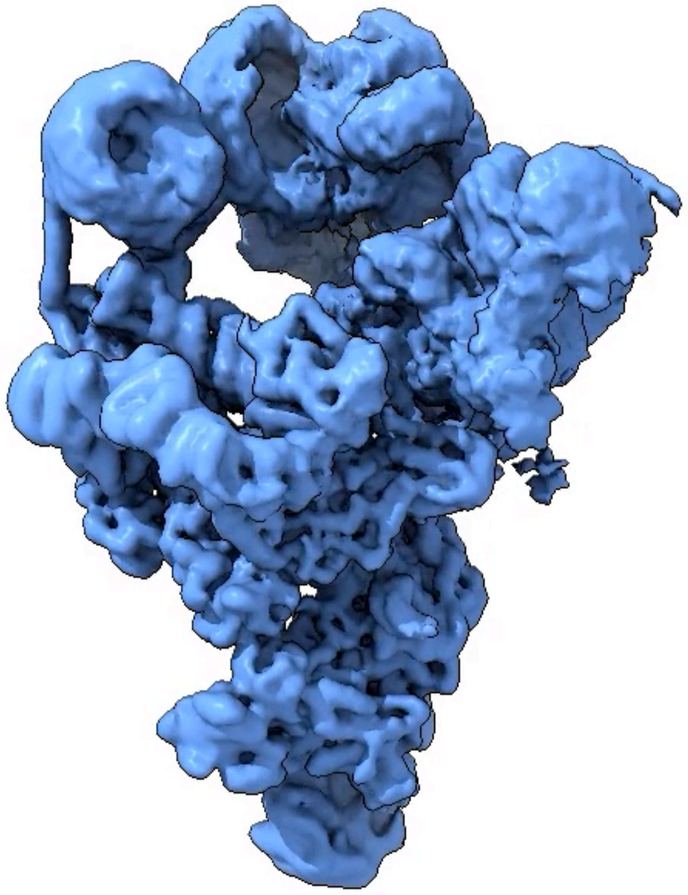


`cryodrgn pc_traversal`

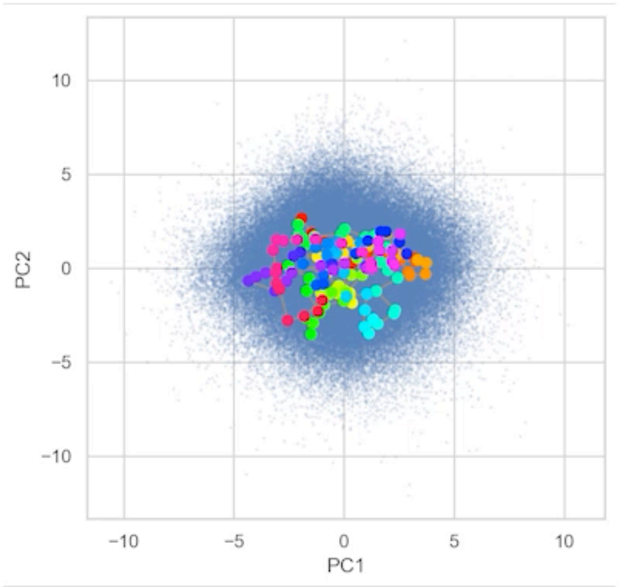
Caveat: Interpolation along PCs can produce nonphysical motions e.g. under compositional heterogeneity and in general when the data distribution is not supported along the interpolation path

Generating trajectories with a graph traversal algorithm





```
cryodrgn graph_traversal
```

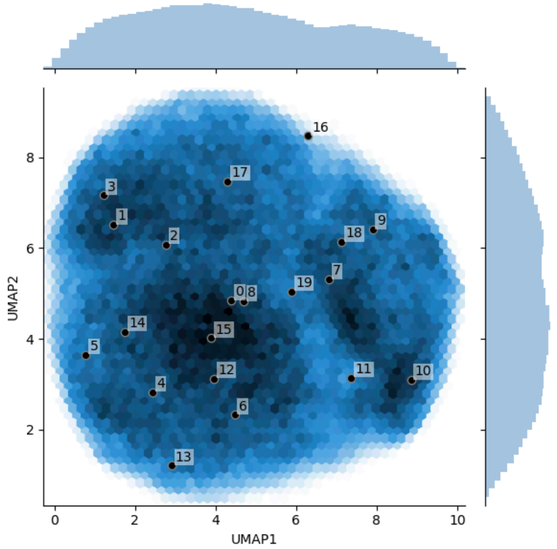


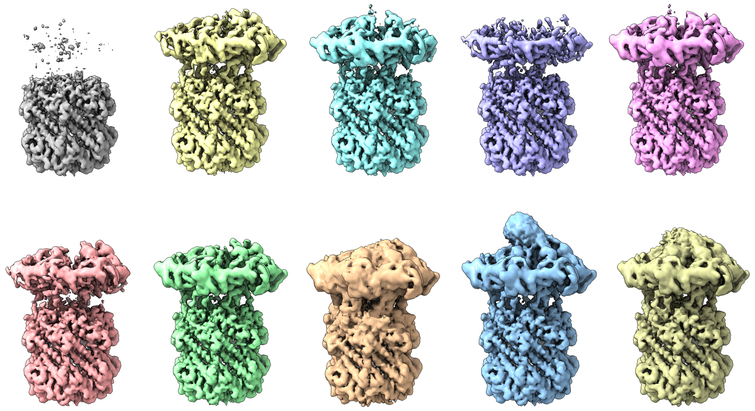
- Graph traversal algorithm along latent embedding nearest-neighbor graph
- Explore the learned distribution

CryoDRGN interactive analysis

In the generative modeling paradigm, cryoDRGN can reconstruct an arbitrary number of cryo-EM volumes.

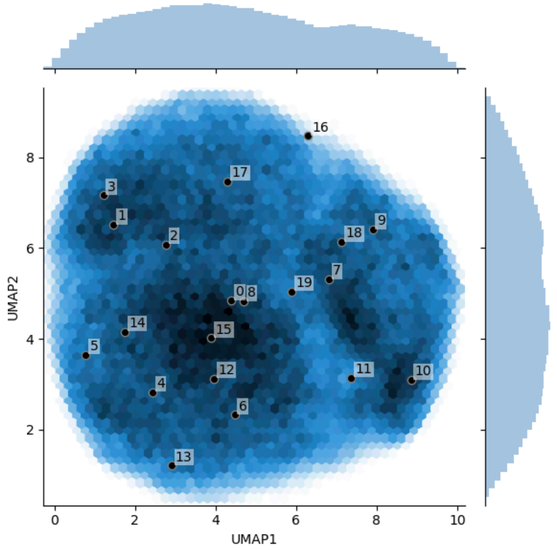
How do we analyze the resulting ensemble of structures?





CryoDRGN interactive analysis

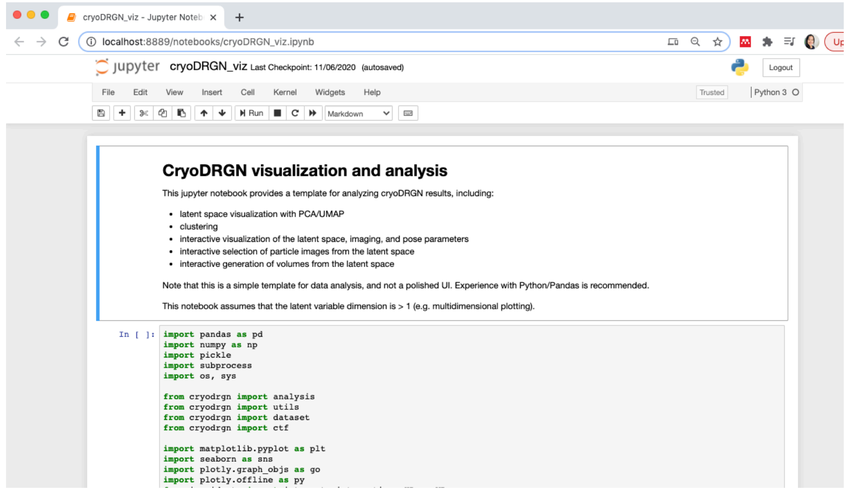
In the generative modeling paradigm, cryoDRGN can reconstruct an arbitrary number of cryo-EM volumes.



How do we analyze the resulting ensemble of structures?

The cryoDRGN **jupyter notebook** is a web application that allows **exploratory** data analysis:

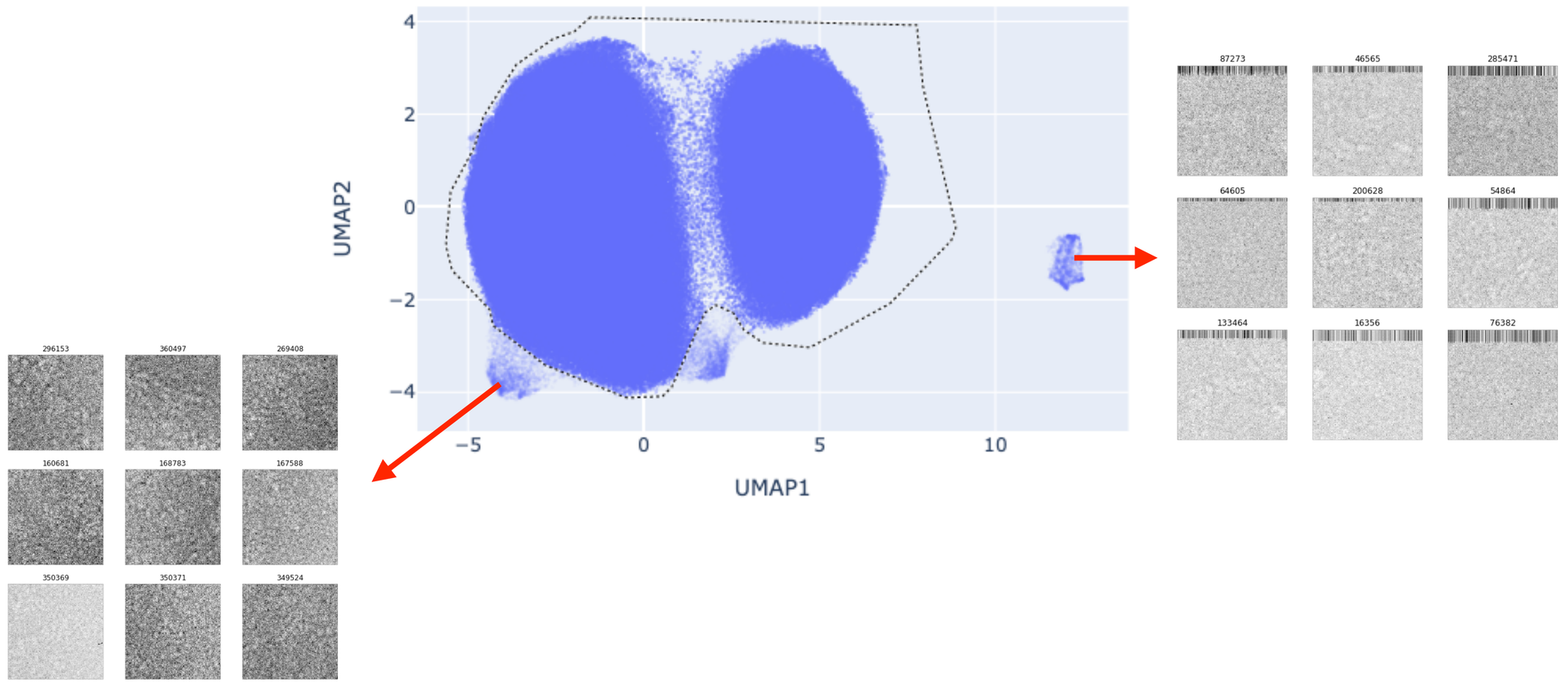
- visualization of the latent space embeddings
- visualization of images
- generation of new volumes



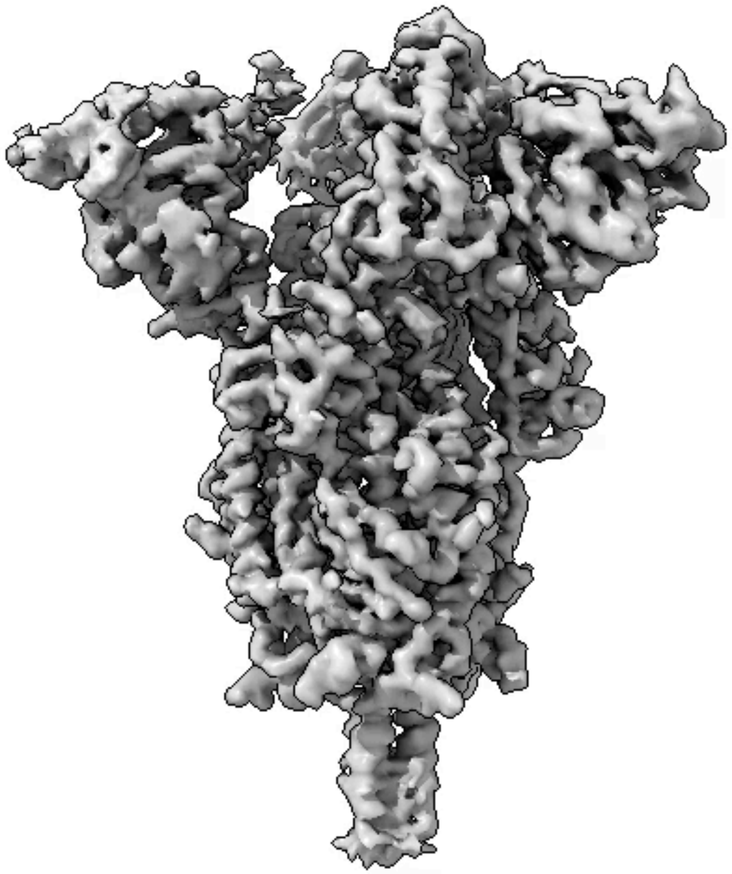
Interactive filtering of non-structural imaging variability

- Non-structural imaging variability (e.g. junk particles, ice artifacts, peripheral particles) may interfere with optimization and representation learning

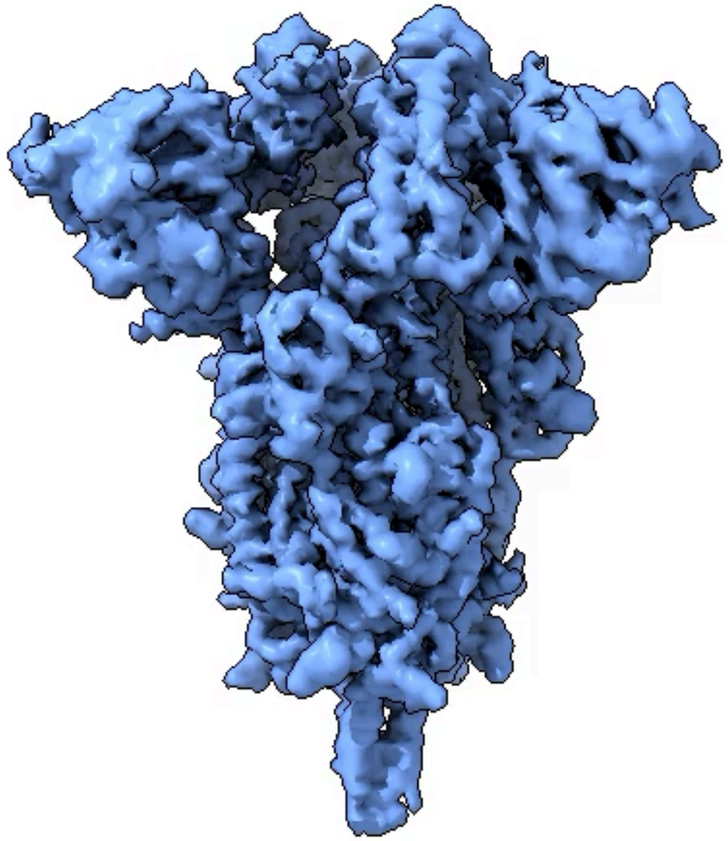
Interactive lasso tool



CryoDRGN reconstruction of the SARS CoV-2 spike protein



20 sampled structures



Graph traversal trajectory

Dataset and training details:

* Walls et al 2020
* 276k particles
* D=256, large architecture
* 25 epochs
* 4 GPU training, 10 hr total

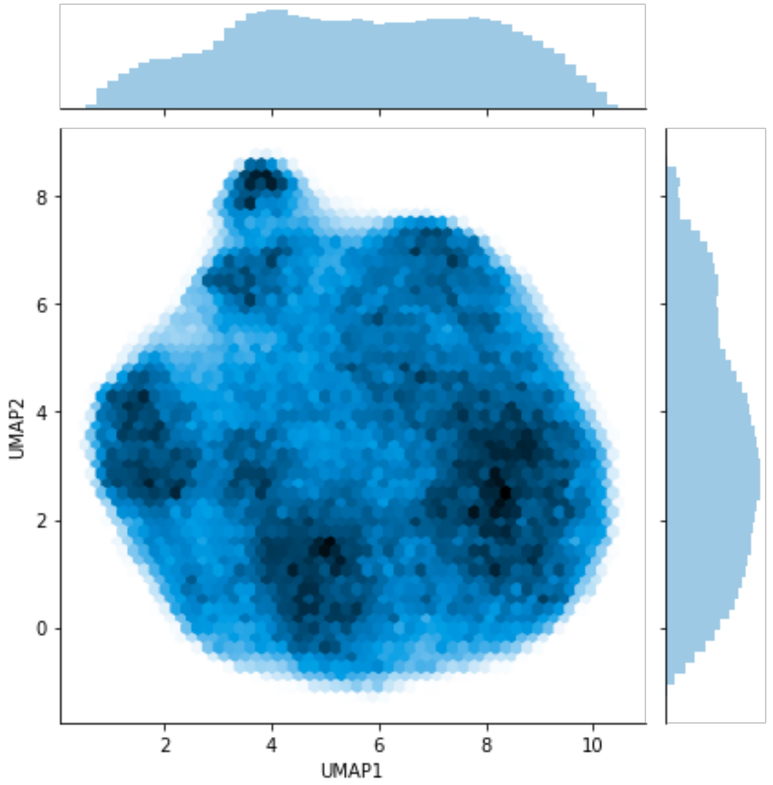
Towards automated analysis of the structure distribution

Dataset: Structural basis of ClpXP recognition and unfolding of ssrA-tagged substrates
Fei et al. 2020, eLife

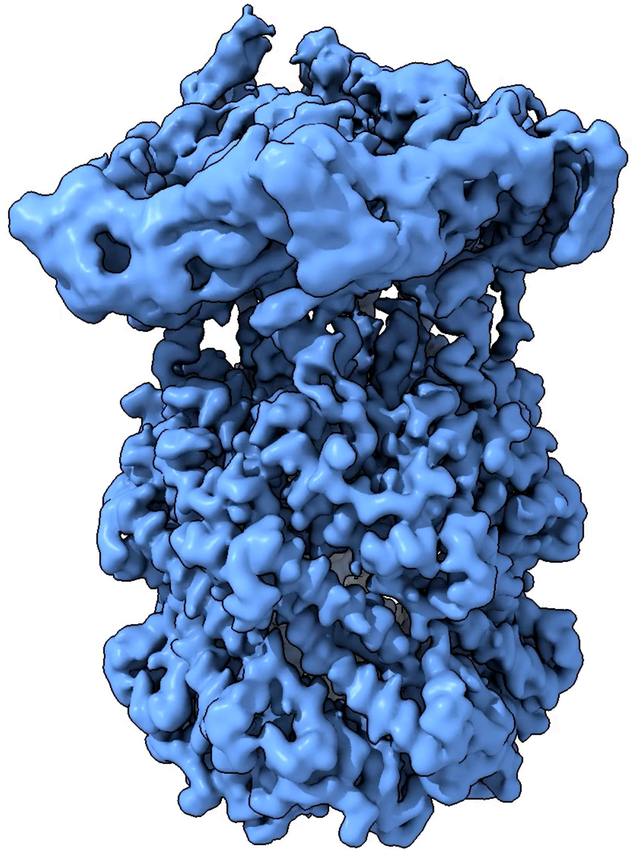

Xue Fei


Bob Sauer

cryoDRGN latent space



`cryodrgn pc_traversal`



Intermediate <-> Recognition complex

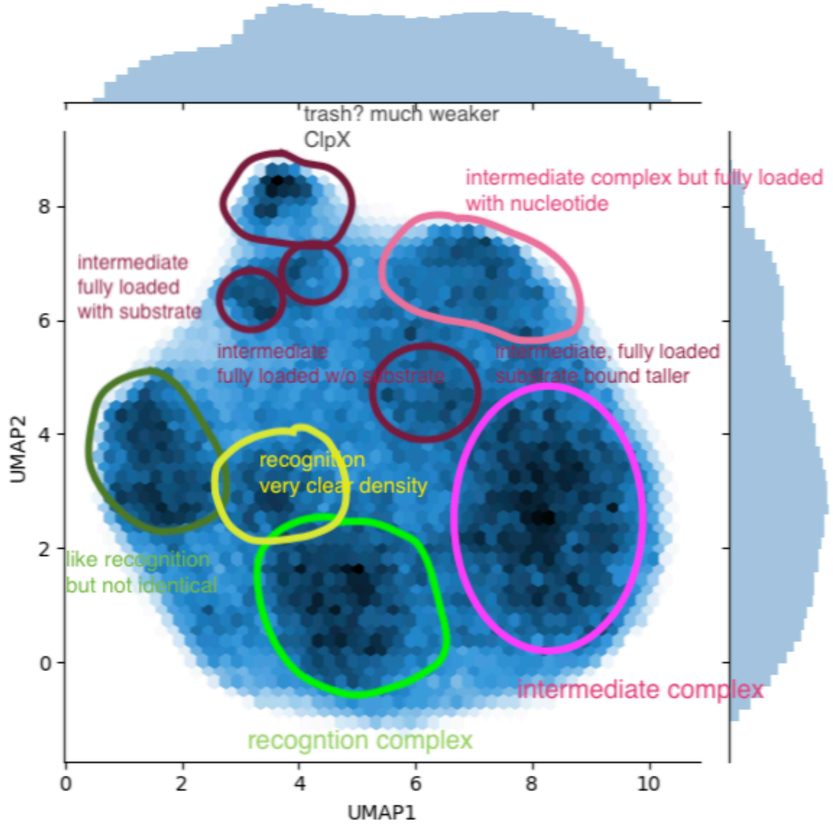
Towards automated analysis of the structure distribution

Dataset: Structural basis of ClpXP recognition and unfolding of ssrA-tagged substrates
Fei et al. 2020, eLife

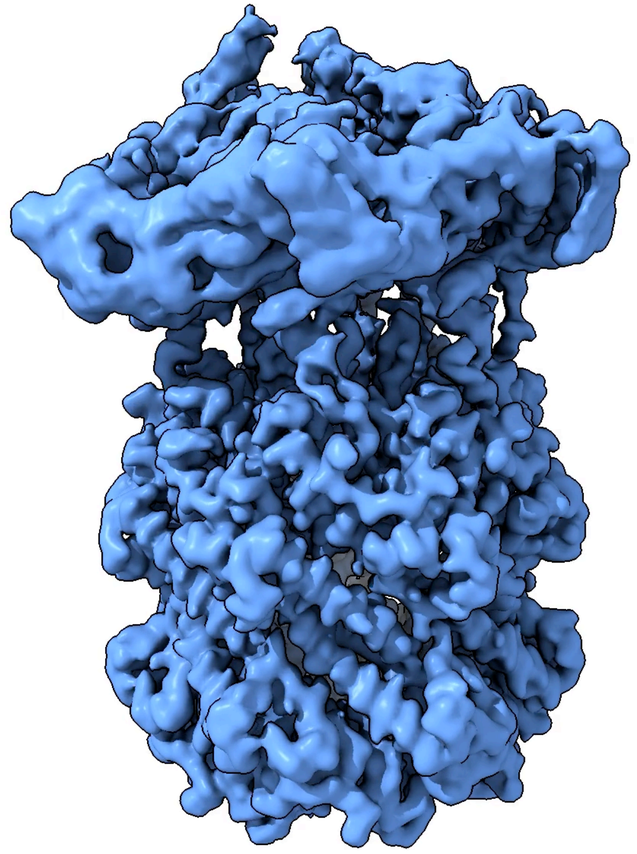

Xue Fei


Bob Sauer

cryoDRGN latent space



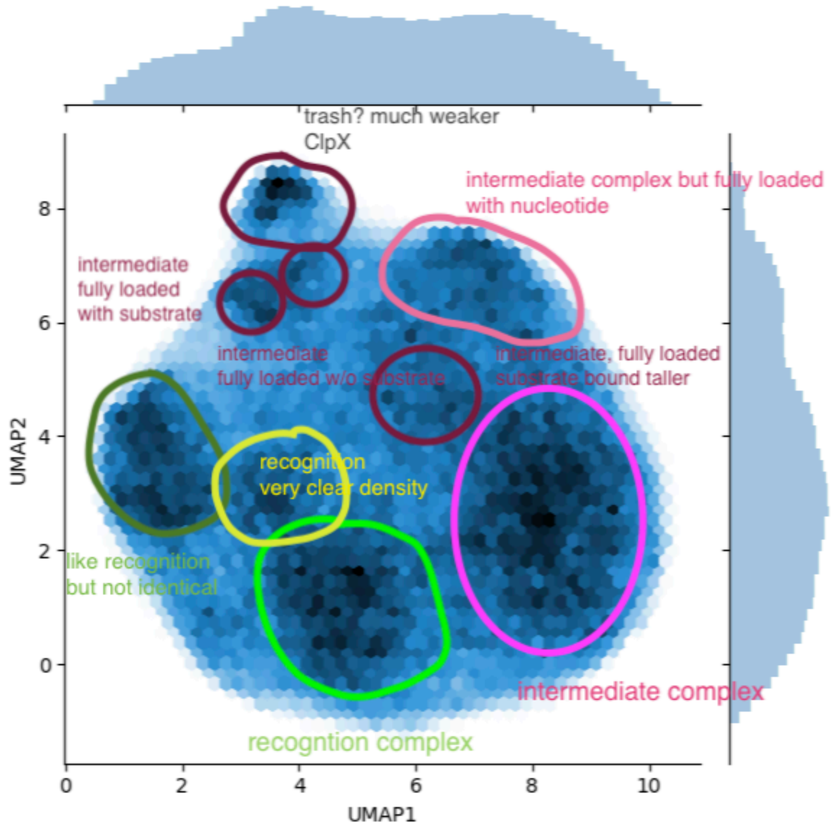
`cryodrgn pc_traversal`



Intermediate <-> Recognition complex

Extending the cryoDRGN toolkit with a scalable structural landscape analysis: `cryodrgn analyze_landscape`

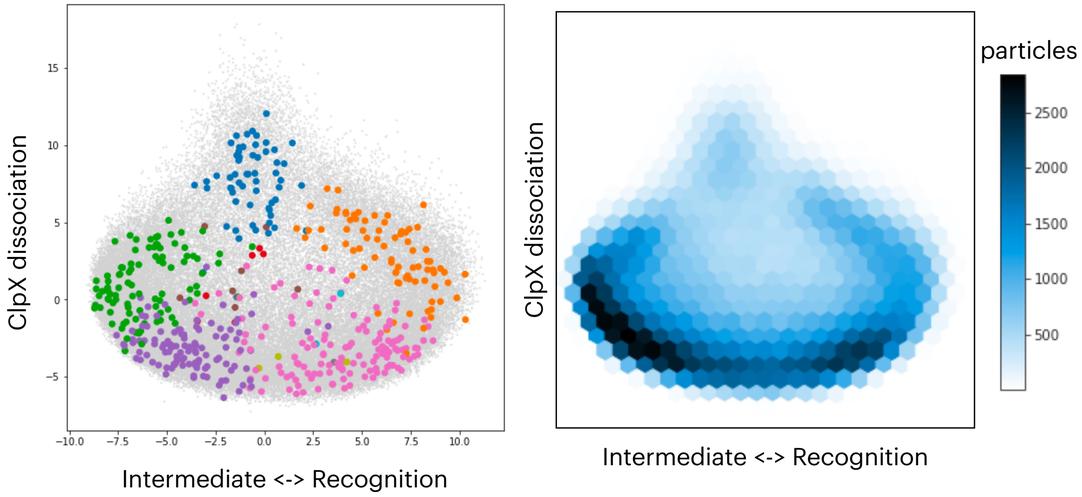
How can we gain insight from high-dimensional biological datasets?



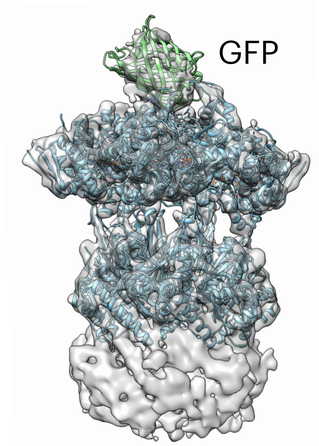
continuous

discrete

Mapping interpretable reaction coordinates



Identification of rare states

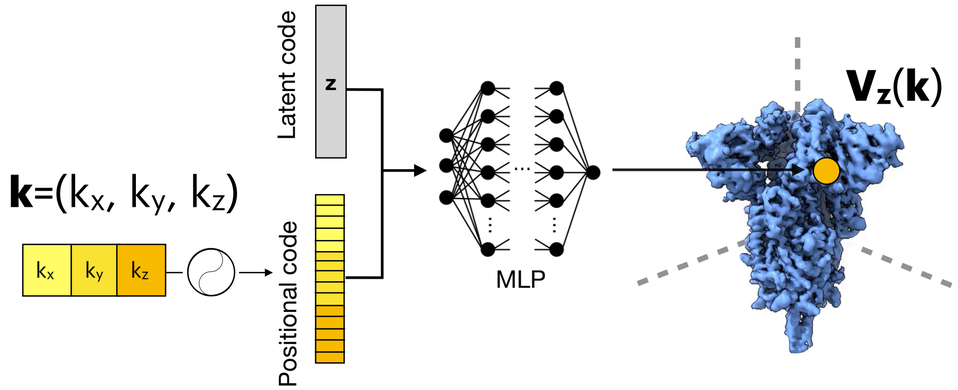


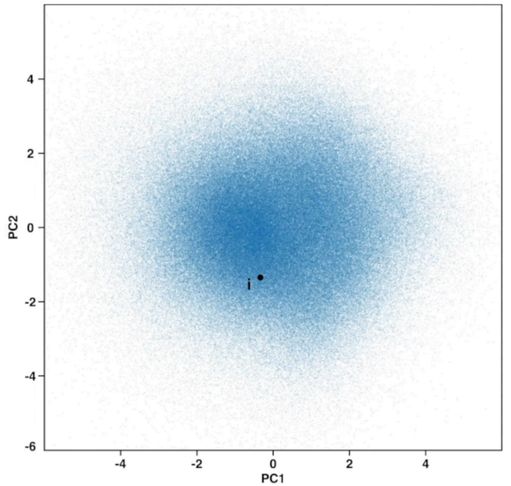
8.5 Å, 1,255 particles
0.3% of the dataset

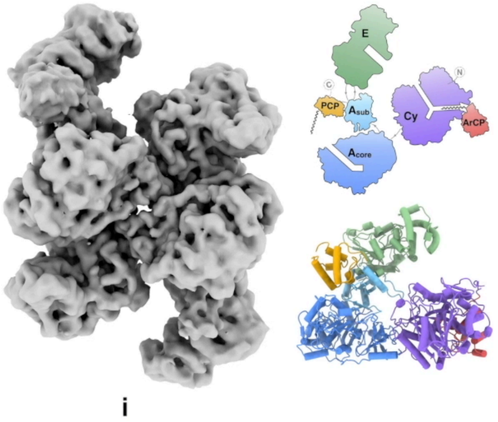
Available in cryoDRGN 1.0

Summary

- Unsupervised reconstruction of a **continuous** distribution of protein structures from cryo-EM images
- A new neural representation for modeling high-resolution density maps
- The deep generative model provides a **general, flexible framework** for modeling heterogeneity
  - Discovery of new structures
  - Visualization of continuous dynamics
- **Novel structures** and molecular motions from cryo-EM data
- **Future outlook:** A nascent area of ML for protein structure determination



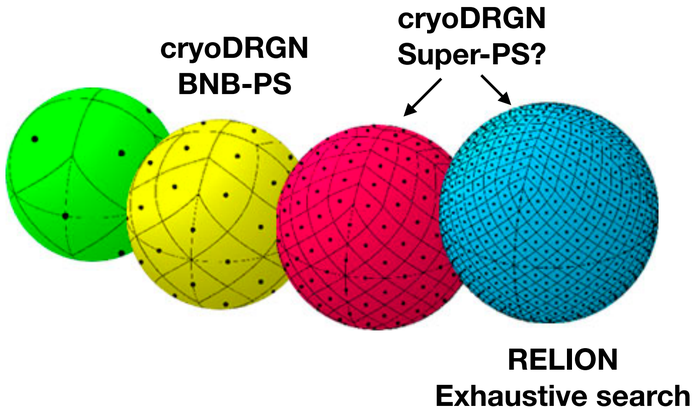




Catalytic trajectory of PchE. Wang et al 2022

A nascent area in ML algorithms for cryo-EM reconstruction

- *Ab initio* cryoDRGN on real datasets
  - Engineering, optimization, and identifiability challenges

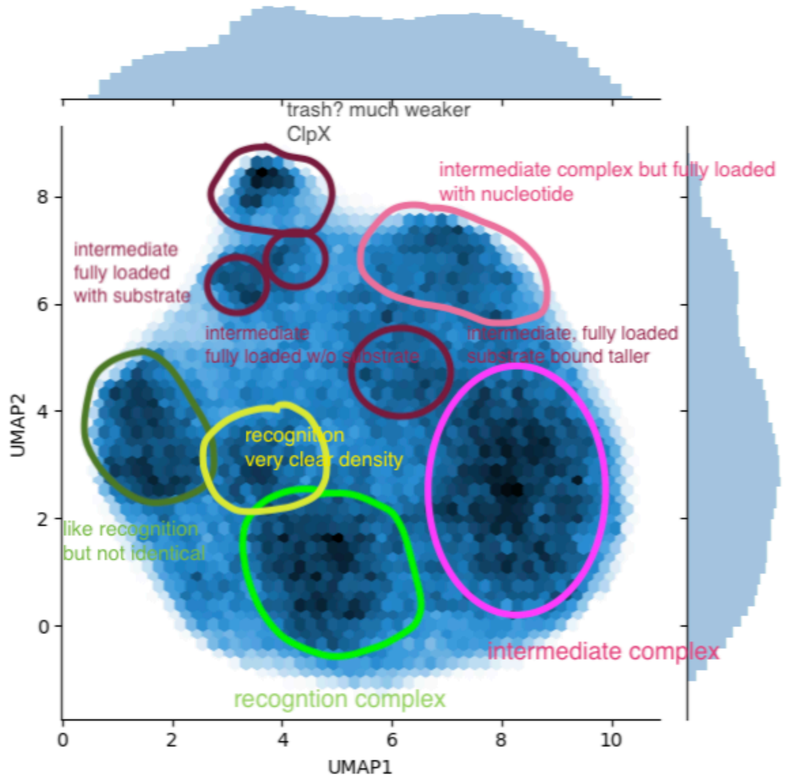


Zhong, Lerer, Davis, Berger, ICCV 2021

A nascent area in ML algorithms for cryo-EM reconstruction

- *Ab initio* cryoDRGN on real datasets
  - Engineering, optimization, and identifiability challenges
- Characterizing distributions of protein structure
  - Methods for exploratory data analysis, benchmarks, and atomic modeling

How can we gain insight from high-dimensional biological datasets?

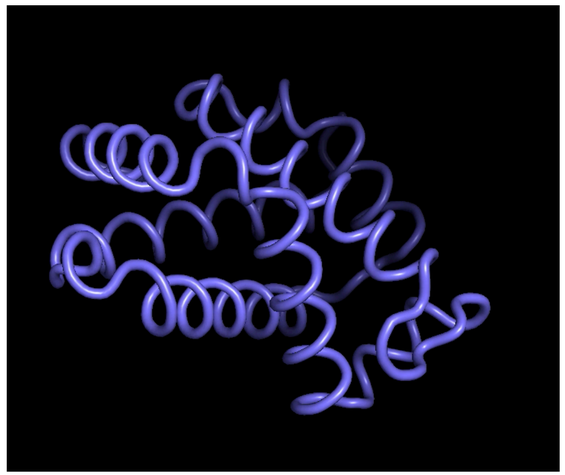


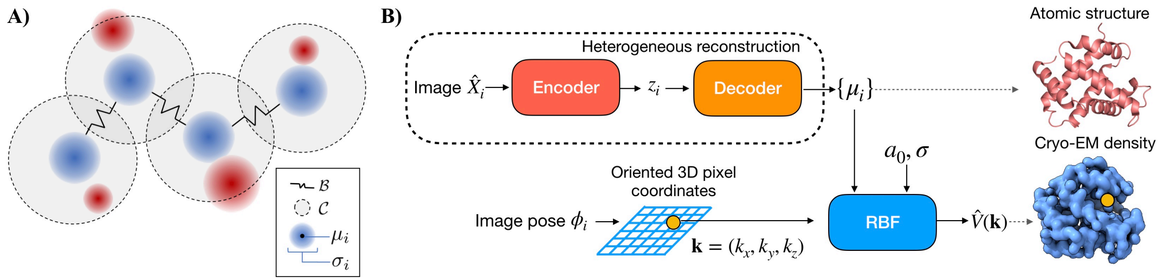
With Ashwin Narayan, Bonnie Berger, Laurel Kinman, Barrett Powell, Joey Davis, Xue Fei, Bob Sauer

A nascent area in ML algorithms for cryo-EM reconstruction

- *Ab initio* cryoDRGN on real datasets
  - Engineering, optimization, and identifiability challenges
- Characterizing distributions of protein structure
  - Methods for exploratory data analysis, benchmarks, and atomic modeling
- New representations and generative modeling paradigms
  - Better inductive biases for protein motion/dynamics; Exploiting information from structure/sequence databases

Radial basis function representation:



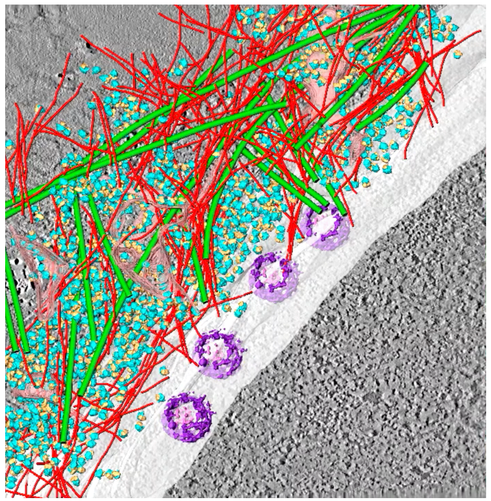


Zhong, Lerer, Davis, Berger. NeurIPS 2020 **Workshop on ML for Structural Biology**

A nascent area in ML algorithms for cryo-EM reconstruction

- *Ab initio* cryoDRGN on real datasets
  - Engineering, optimization, and identifiability challenges
- Characterizing distributions of protein structure
  - Methods for exploratory data analysis, benchmarks, and atomic modeling
- New representations and generative modeling paradigms
  - Better inductive biases for protein motion/dynamics, Exploiting information from structure/sequence databases
- *In situ* cryoDRGN
  - Towards in situ structural biology with cryo-electron tomography (cryo-ET)

https://pdb101.rcsb.org/sci-art/goodsell-gallery/escherichia-coli-bacterium

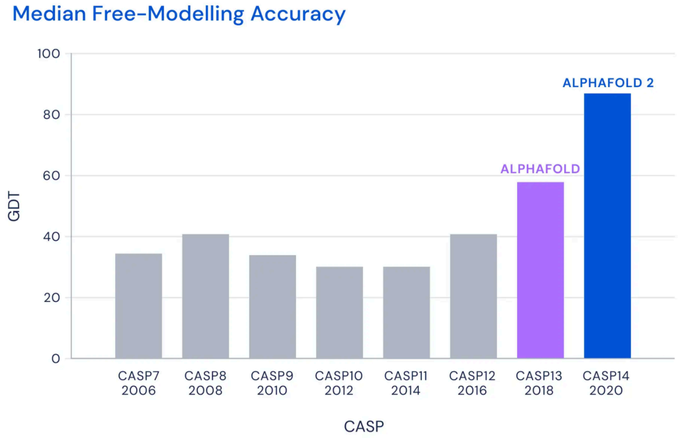


Visualizing the molecular sociology at the HeLa cell nuclear periphery
Mahamid et al, Science 2016

A nascent area in ML algorithms for cryo-EM reconstruction

- *Ab initio* cryoDRGN on real datasets
  - Engineering, optimization, and identifiability challenges
- Characterizing distributions of protein structure
  - Methods for exploratory data analysis, benchmarks, and atomic modeling
- New representations and generative modeling paradigms
  - Better inductive biases for protein motion/dynamics, Exploiting information from structure/sequence databases
- *In situ* cryoDRGN
  - Towards in situ structural biology with cryo-electron tomography (cryo-ET)

- **Outlook in the post-AlphaFold2 era?**



Thank you for listening!

E.Z. Lab

- **Ramya Rangan**
- **Axel Levy**
- Rish Raghu
- Ryan Feathers
- Vineet Bansal

Stanford

- Fred Poitevin
- Gordon Wetzstein

NVIDIA

- Tim Dockhorn
- Karsten Kreis

Flatiron Institute

- Pilar Cossio
- Sonya Hanson

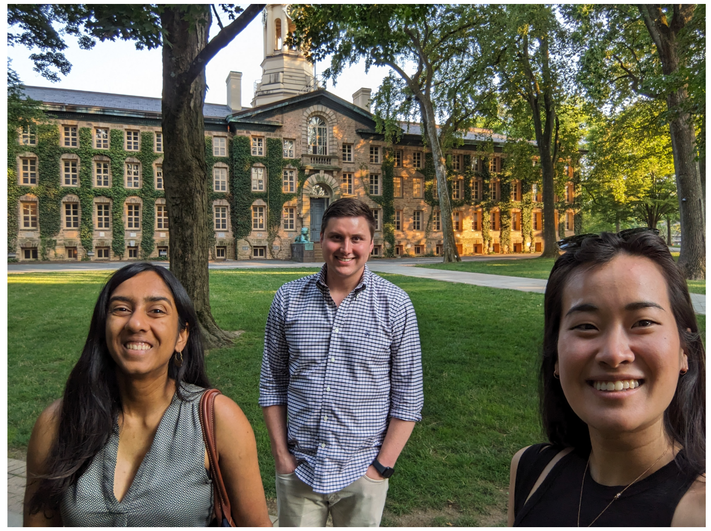
ET/Thermo Fischer

- Sagar Svenekar (MPI Biochemistry)
- Jake Johnston (Columbia/NYSBC)
- Adam Lerer (DeepMind)
- Martin Obr (TF)
- Ron Kelley (TF)
- Abhay Kotecha (TF)

CryoDRGN1

- Tristan Bepler
- Bonnie Berger (MIT Mathematics)
- Joey Davis (MIT Biology)
- Ashwin Narayan
- Barrett Powell
- Laurel Kinman

EZ Lab at NeurIPS Machine Learning for Structural Biology Workshop



PRINCETON UNIVERSITY