RECONSTRUCTING CONTINUOUS DISTRIBUTIONS OF 3D PROTEIN STRUCTURE FROM CRYO-EM IMAGES

Ellen D. Zhong
MIT
zhonge@mit.edu

Tristan Bepler
MIT
tbepler@mit.edu

Joseph H. Davis*
MIT
jhdavis@mit.edu

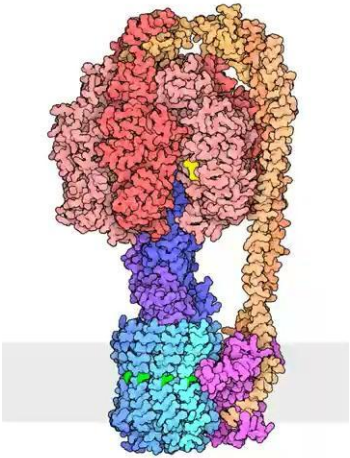
Bonnie Berger*
MIT
bab@mit.edu

CryoDRGN 

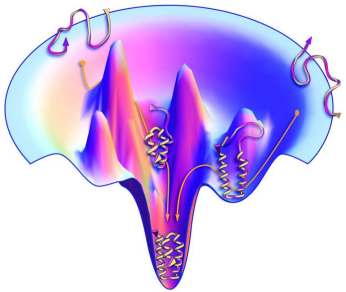
Ariana Vlad and Philipp Janke
09/16/2026

Why Cryo-EM?

Structure encodes function

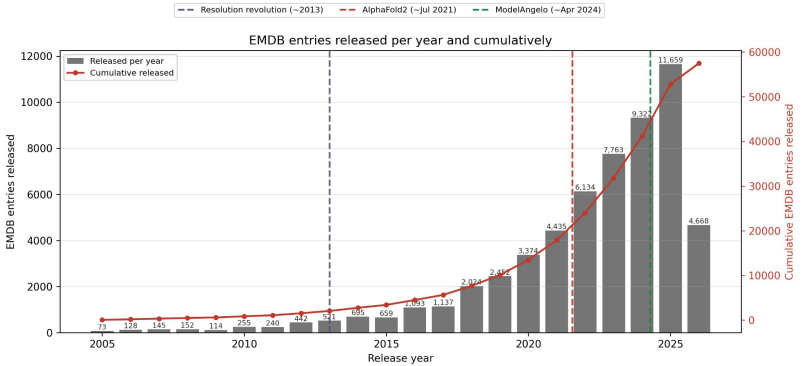


Structure is a collection of conformations sampled from a native energy landscape → this is what we want to measure!

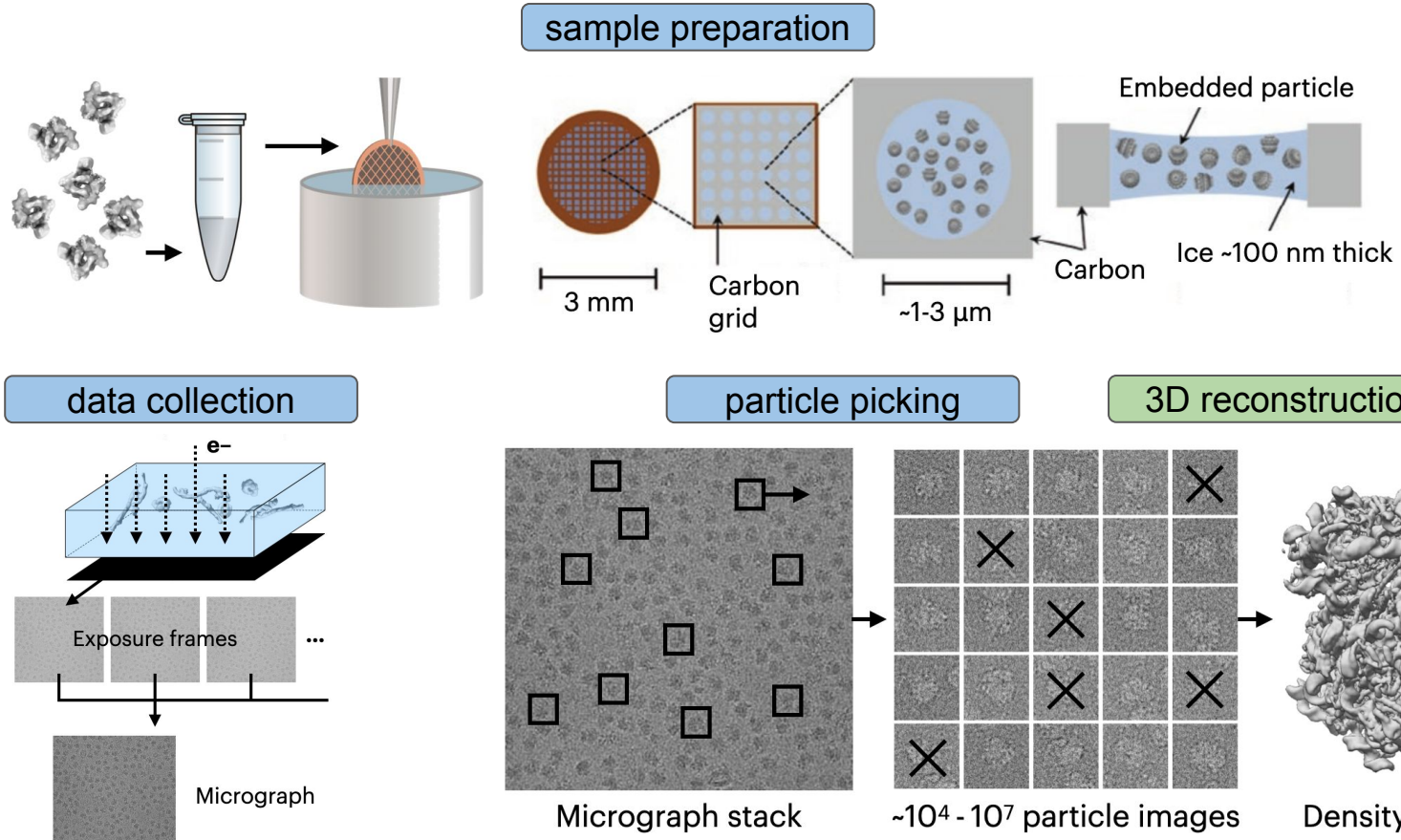


Experiments often break the native landscape through sample processing – cryo-EM flash freezing allows for near-native conformation measurements.

Hardware and software breakthroughs made cryo-EM the fastest growing structure determination method – perfect moment for robust software development.

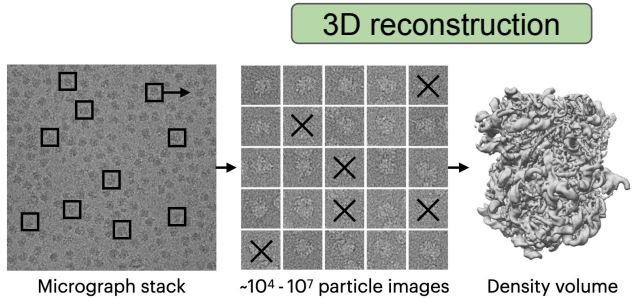


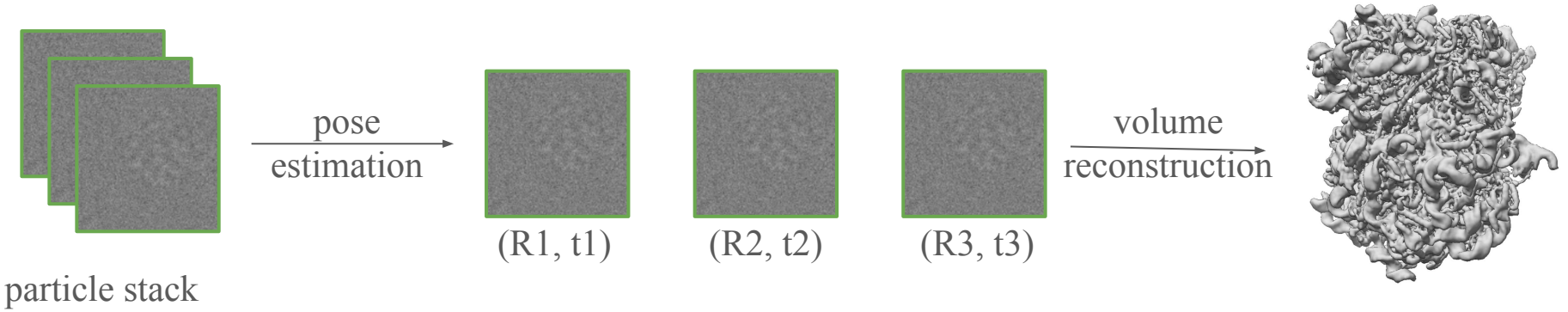
Cryo-EM crash course



credit: Robert Heeter

The cryo-EM reconstruction task





Task: Given a set of N particle images, find the optimal poses and volume.

The image formation model

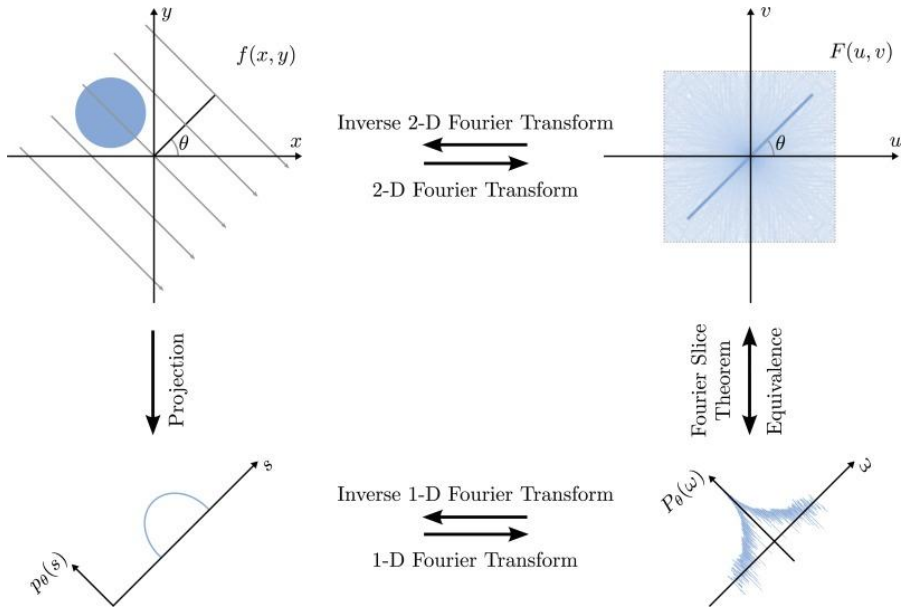
$$\overset{\text{image}}{X(r_x, r_y)} = \overset{\text{PSF}}{g} * \int_{\mathbb{R}} \overset{\text{volume}}{V}(\overset{\text{pose}}{R^T}\mathbf{r} + t)\, dr_z + noise \qquad \mathbf{r} = (r_x, r_y, r_z)^T$$

The image formation model

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

image: $X(r_x, r_y)$ — PSF: g — volume: V — pose: R, t

Fourier Slice Theorem



Taubmann et al. 2018

The image formation model

volume

image PSF pose

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

image CTF volume N(0, I)

$$\hat{X}(k_x, k_y) = \hat{g} S(t) A(R) \hat{V}(k_x, k_y) + \epsilon$$

pose operators

The image formation model

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

image: $X(r_x, r_y)$ — PSF: g — volume: V — pose: R, t

$$\hat{X}(k_x, k_y) = \hat{g} S(t) A(R) \hat{V}(k_x, k_y) + \epsilon$$

image: $\hat{X}$ — CTF: $\hat{g}$ — pose operators: $S(t), A(R)$ — volume: $\hat{V}$ — N(0, I): ϵ

find pose + volume that maximizes probability

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

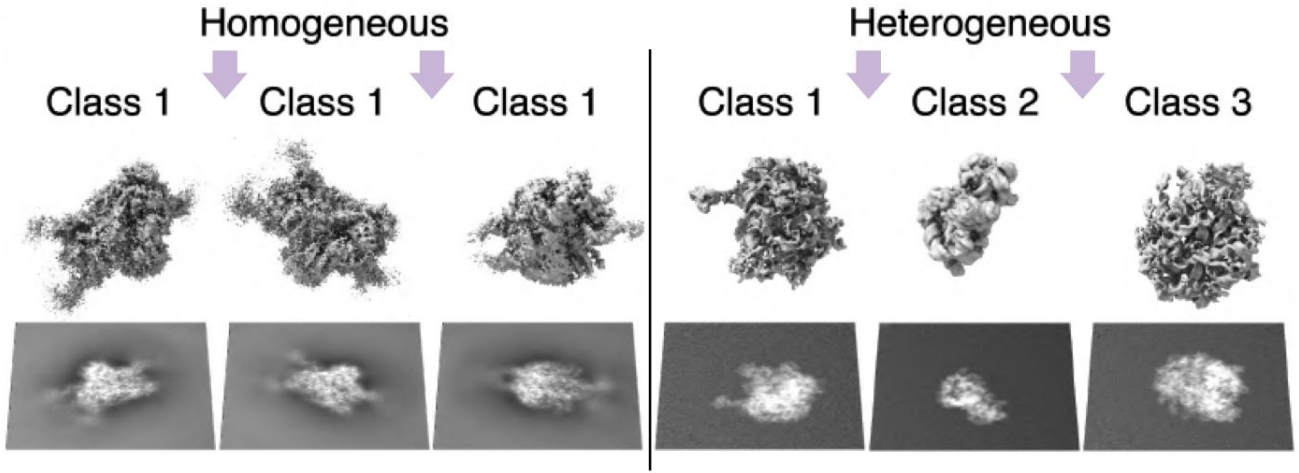
Traditional cryo-EM reconstruction

Homogeneous reconstruction:

- alternating iterative refinement of pose and volume
  - expectation-maximization
  - branch-and-bound pose estimation + gradient descent

Heterogeneous reconstruction:

- discrete volumes: *multiclass refinement*
- the number of clusters K obtained is a hyperparameter that has to be iterated on

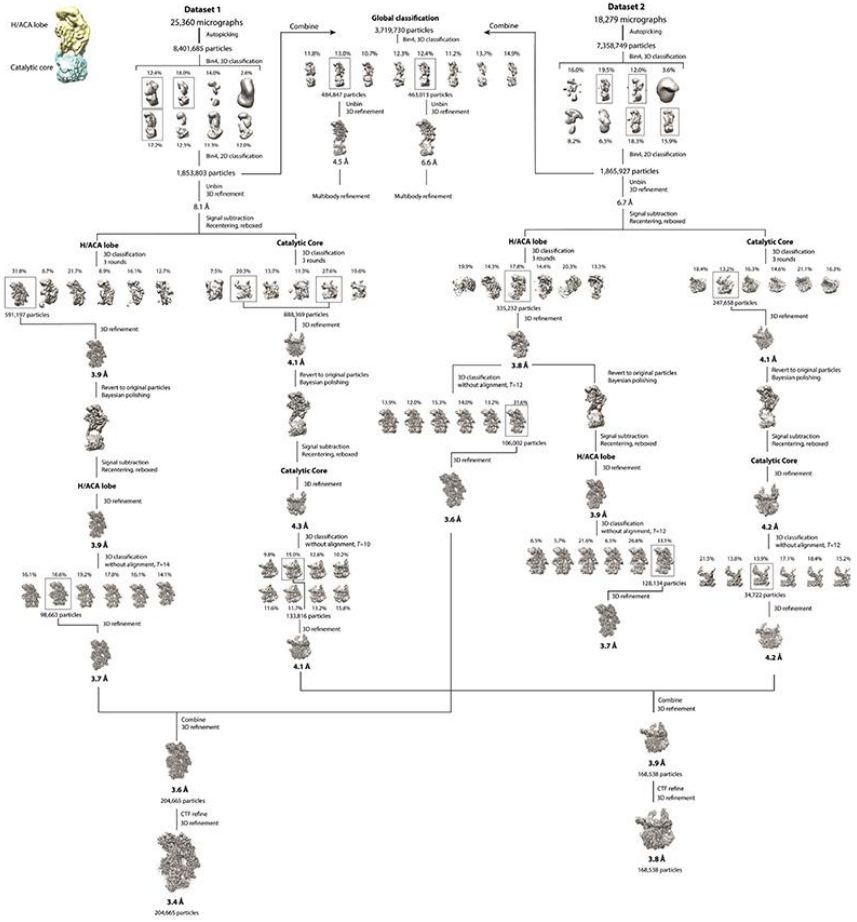


Qu et al., 2025

Traditional cryo-EM reconstruction

Homogeneous recons nstruction:

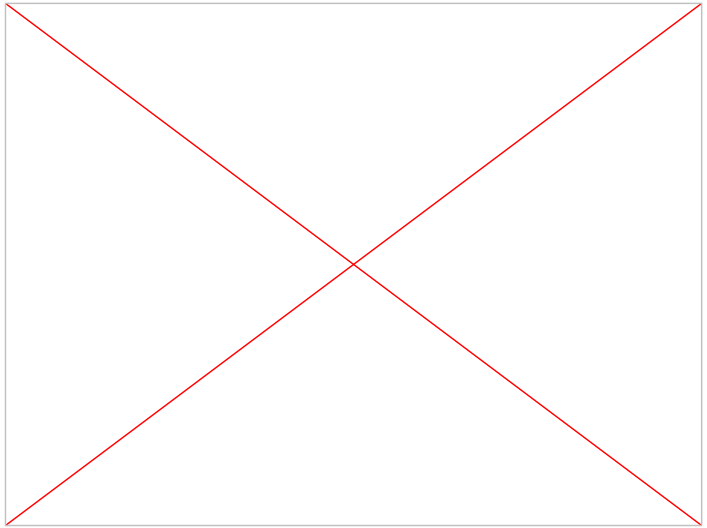
- alternating itera s: *multiclass refinement* pose and volum lusters K obtained is a
  - expectation that has to be iterated on
  - branch-and estimation



Qu et al., 2025

Motivation

- A lot important biology happens through conformational changes
- Number and nature of the conformational states is unknown during reconstruction
    - assumptions change quality of results and several iterations are needed before finding the “best” hyperparameter fit for the data
- Some heterogeneity is continuous and can’t be modeled by K classes
    - hinge motions, opening/closing of channels



credit: Rishwanth Raghu

The cryoDRGN model

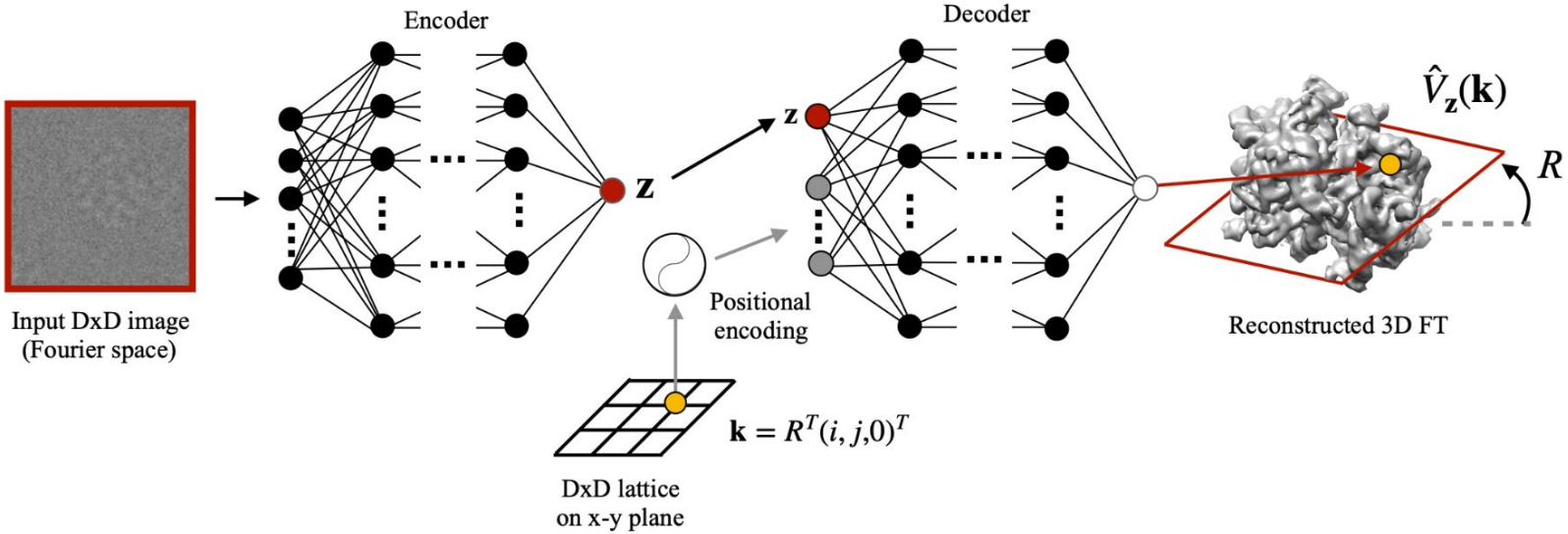


Figure 2: CryoDRGN model architecture. We use a VAE to perform approximate inference for latent variable z denoting image heterogeneity. The decoder reconstructs an image pixel by pixel given z and $pe(k)$, the positional encoding of 3D Cartesian coordinates. The 3D coordinates corresponding to each image pixel are obtained by rotating a DxD 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).

The cryoDRGN model – encoder

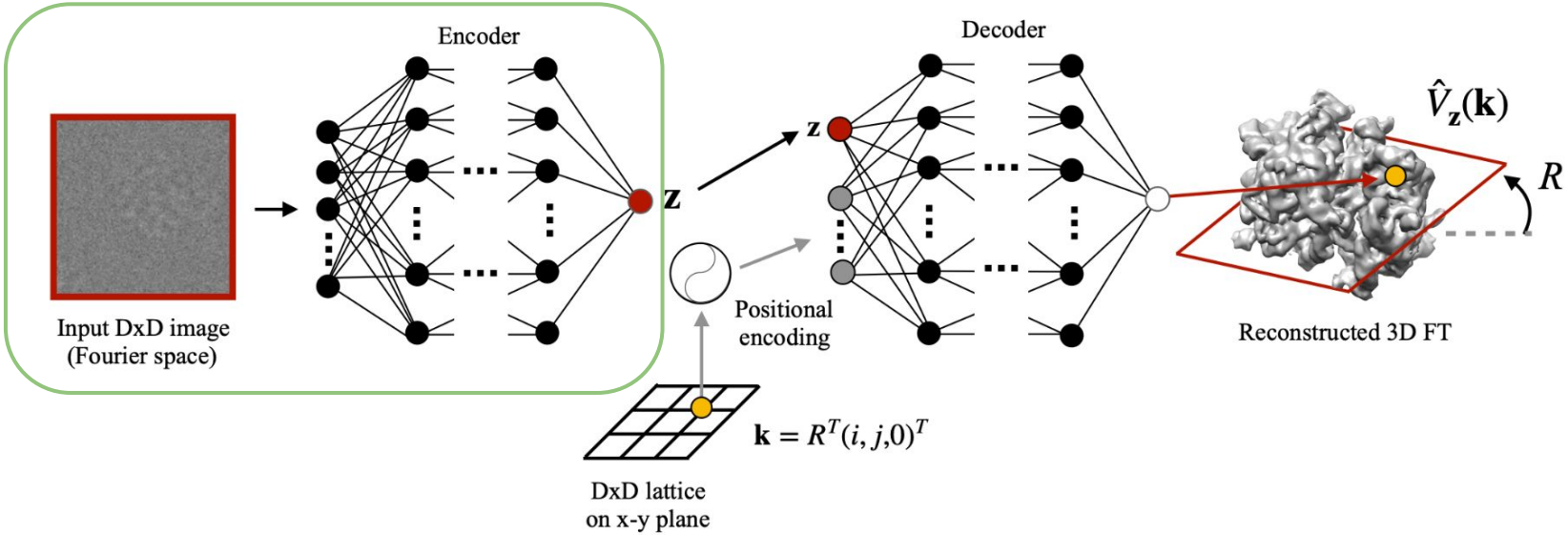
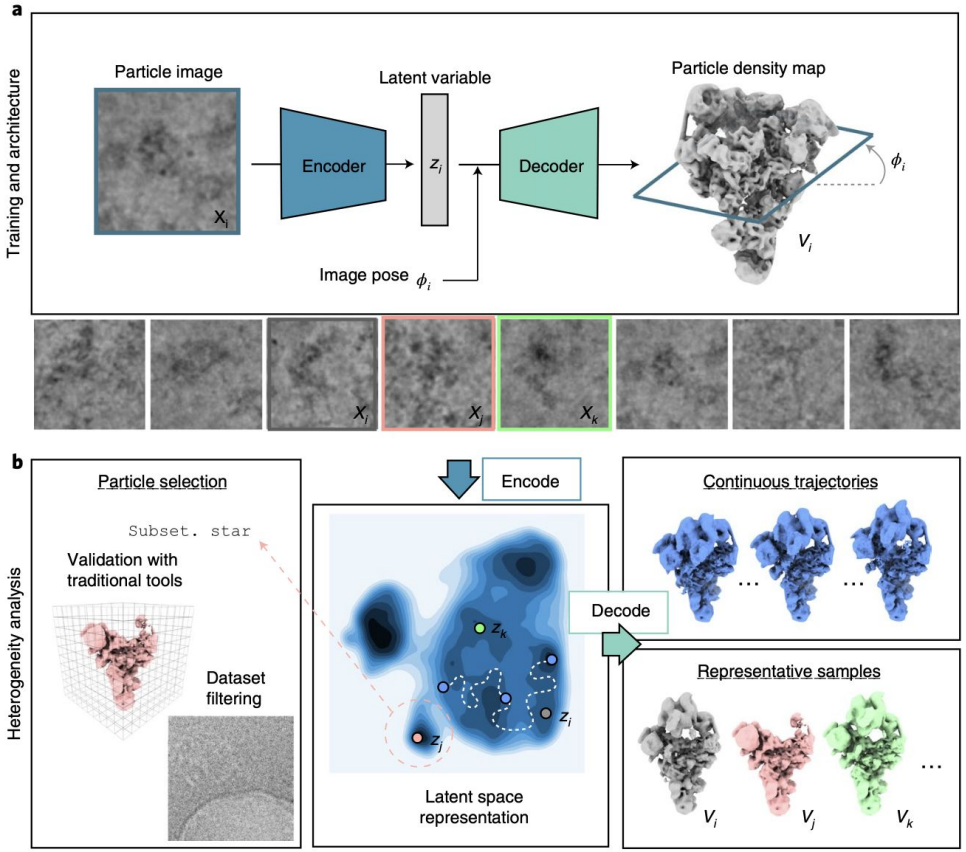


Figure 2: CryoDRGN model architecture. We use a VAE to perform approximate inference for latent variable z denoting image heterogeneity. The decoder reconstructs an image pixel by pixel given z and $pe(k)$, the positional encoding of 3D Cartesian coordinates. The 3D coordinates corresponding to each image pixel are obtained by rotating a DxD 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).

The cryoDRGN model – encoder



- cryoDRGN works exclusively in Fourier space
- Images are represented through Hartley space representation
  - needed since MLPs need real numbers
- Instead of classifying the image, the encoder maps it to a continuous latent variable
  - note that the pose is not encoded in the latent!

The cryoDRGN model – decoder

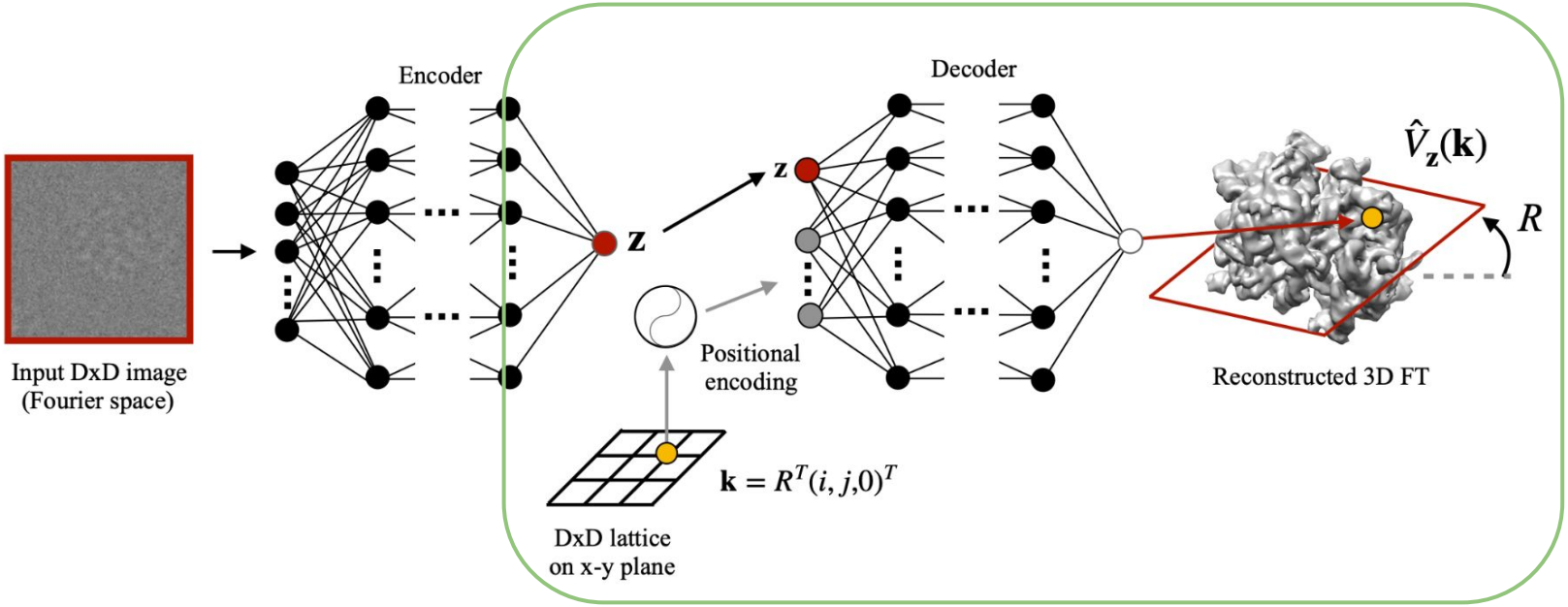
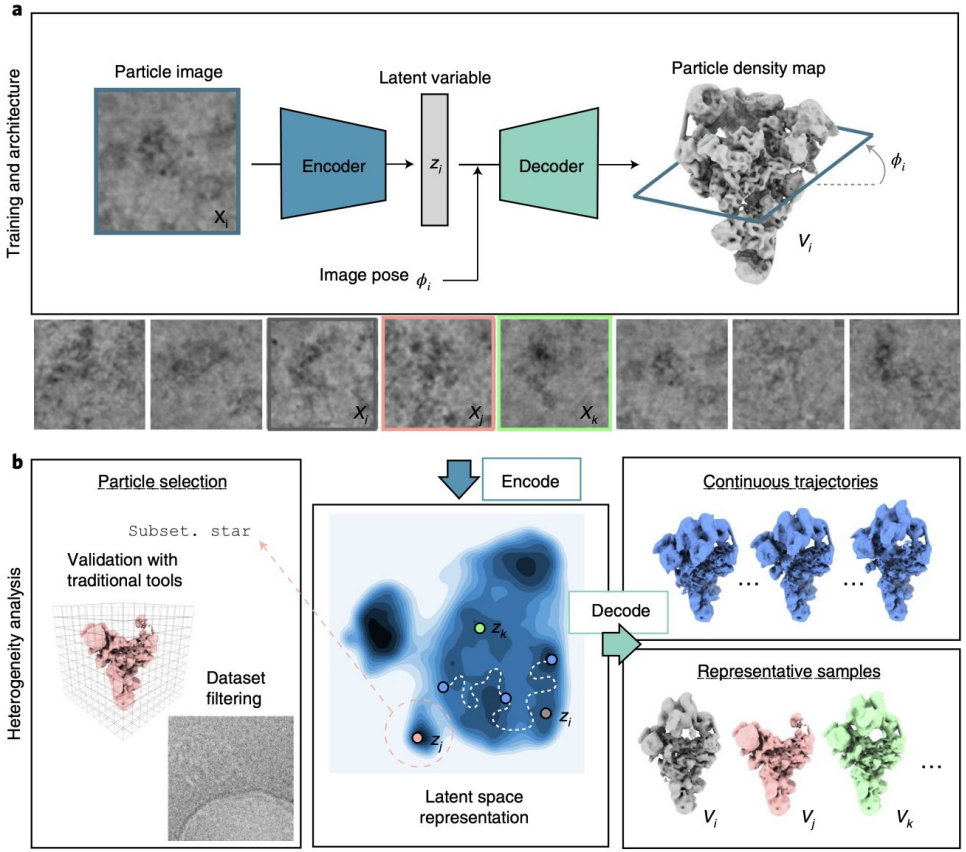


Figure 2: CryoDRGN model architecture. We use a VAE to perform approximate inference for latent variable z denoting image heterogeneity. The decoder reconstructs an image pixel by pixel given z and $pe(k)$, the positional encoding of 3D Cartesian coordinates. The 3D coordinates corresponding to each image pixel are obtained by rotating a DxD 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).

The cryoDRGN model – decoder



- Moving away from the classical grid representation → continuous implicit function
    - instead of sampling position (x,y,z), the decoder samples (kx, ky, kz)
- Decoder reconstructs a slice in Fourier space, then compares to the input particle image
- Sampling is possible from any point of the latent space: representative structures vs. continuous trajectories

The cryoDRGN model – pose search

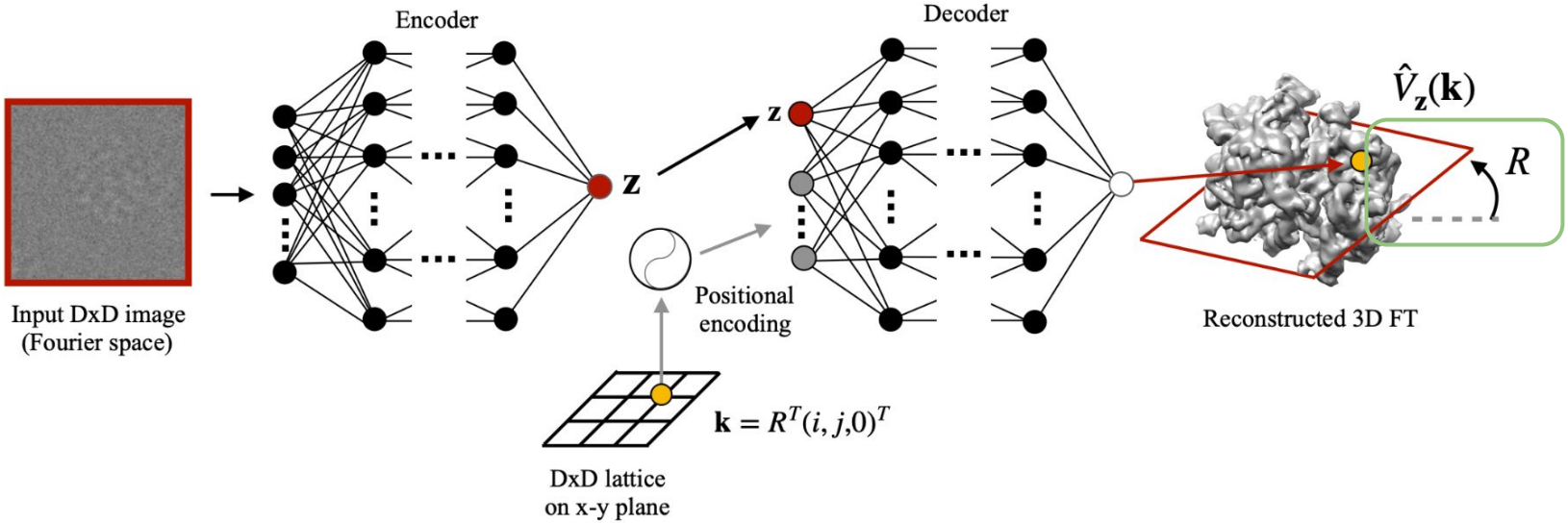


Figure 2: CryoDRGN model architecture. We use a VAE to perform approximate inference for latent variable z denoting image heterogeneity. The decoder reconstructs an image pixel by pixel given z and $pe(k)$, the positional encoding of 3D Cartesian coordinates. The 3D coordinates corresponding to each image pixel are obtained by rotating a DxD 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).

The cryoDRGN model – pose search

- Pose space is huge – cryoDRGN discretizes it uniformly and iteratively
- Prune the poses by comparing a cheap lower bound evaluated at each position to an upper bound evaluated at one point
- At each iteration, increase k_max
  - computation gets more expensive
  - low-frequency dominates pose estimation
- Final resolution ~1°

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 $iter = 1 \ \ldots \ N_{iter}$ **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(ϕ_i)
    $\Phi \leftarrow \Phi_{new}$
    $k \leftarrow k + (k_{max} - k_{min})/(N_{iter} - 1)$ ▷ Increase frequency band limit
  return ϕ^*

How the ICLR paper builds up the result

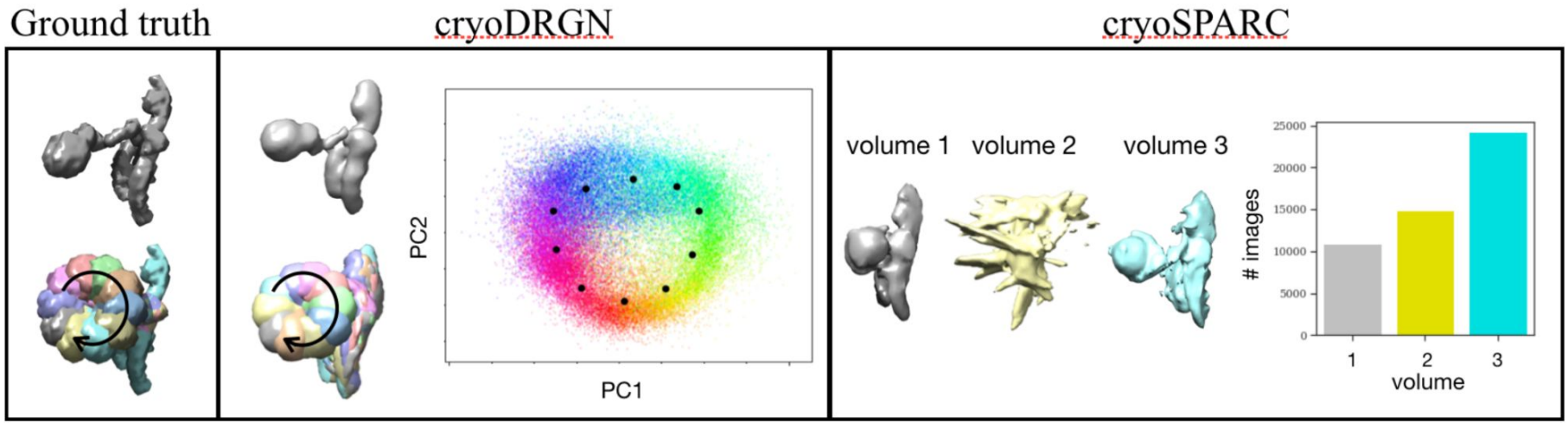
ICLR 2020 · Original proof of concept

Homogeneous, unknown pose	Heterogeneous, known pose	Heterogeneous, unknown pose
•One structure •Pose inferred •Decoder only	•Multiple conformations •Pose supplied •Encoder + Decoder	•Multiple conformations •Pose inferred •Encoder + decoder + pose inference
Validates reconstruction	Recovers known heterogeneity	Fully unsupervised heterogeneous reconstruction

The paper adds complexity step-by-step.

Can cryoDRGN recover continuous heterogeneity?

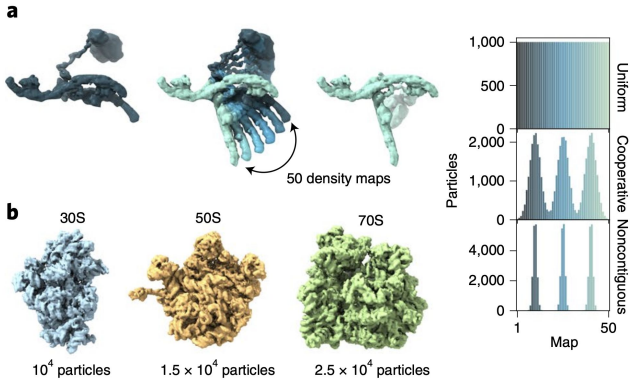
ICLR 2020 · Original proof of concept · Figure 4



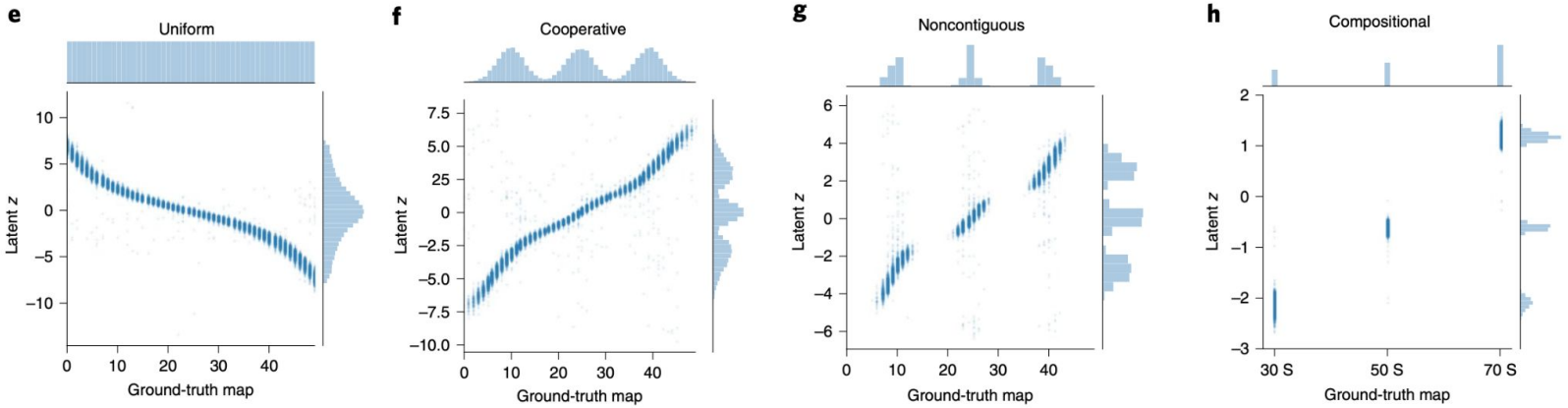
- CryoDRGN's learned manifold preserves the circular motion and its topology
- CryoSPARC's three discrete classes blur together different conformations (only 2 out of 3 structures resemble the ground truth)

Does the learned latent space reflect the true structural landscape?

Nature Methods 2021 · Follow-up work · Figure 3a,b,e–h



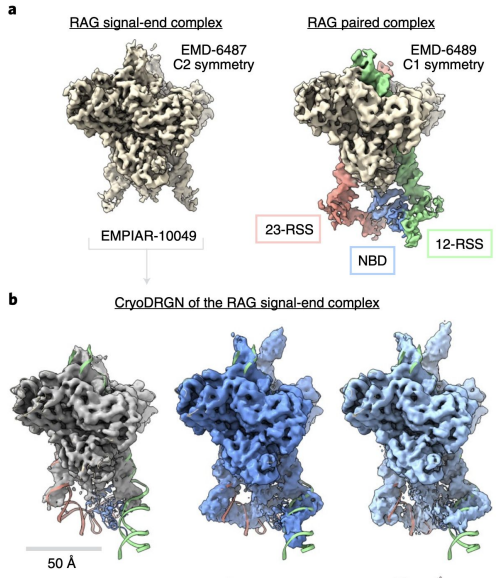
- Tested uniform, cooperative, noncontiguous and compositional variation.
- Latent z tracks the simulated ground-truth reaction coordinate (degree of freedom).



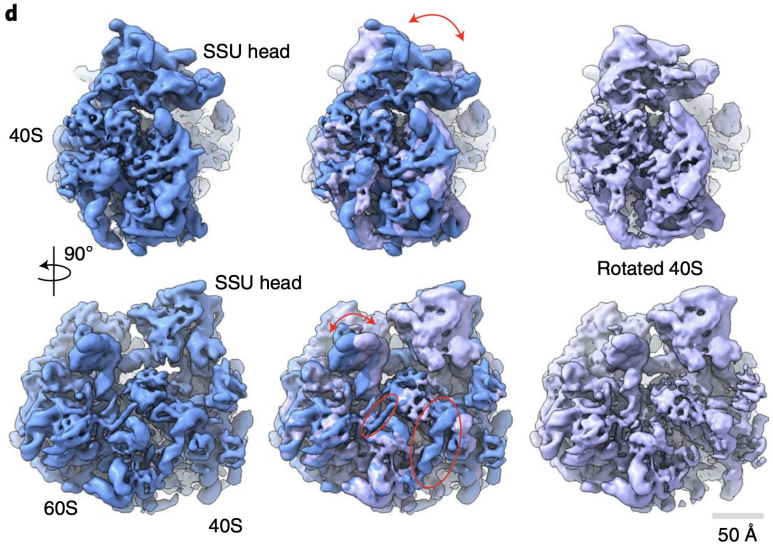
Heterogeneity hidden by a homogeneous reconstruction

Nature Methods 2021 · Follow-up work · Figure 4a,b,d

RAG complex



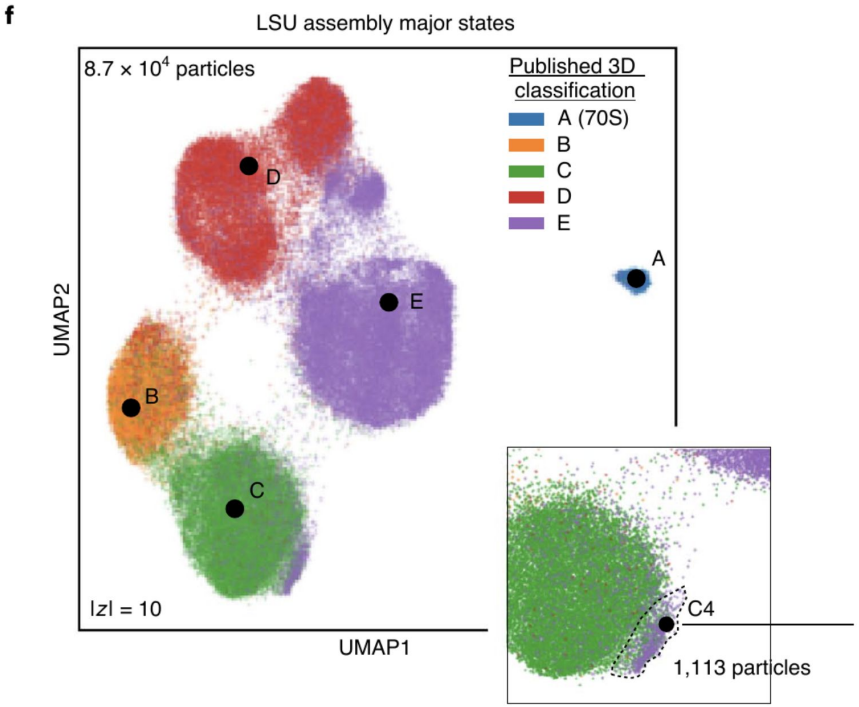
Pf80S ribosome

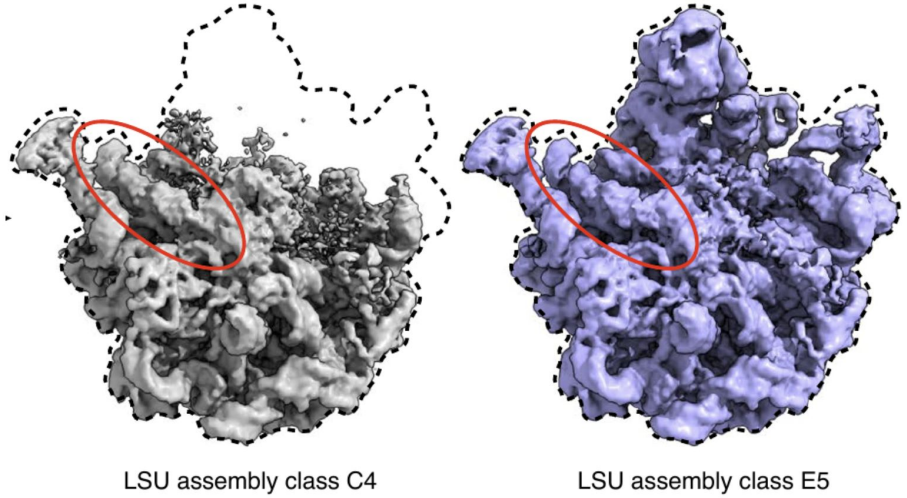


- RAG: cryoDRGN recovers variable RSS DNA and NBD conformations that are blurred out in the consensus reconstruction.
- Pf80S maps reveal 40S rotation and coordinated local rearrangements.

Unsupervised discovery of a rare ribosome assembly state

Nature Methods 2021 · Follow-up work · Figure 5f,i

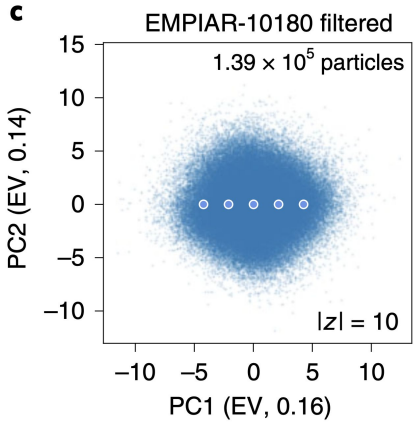




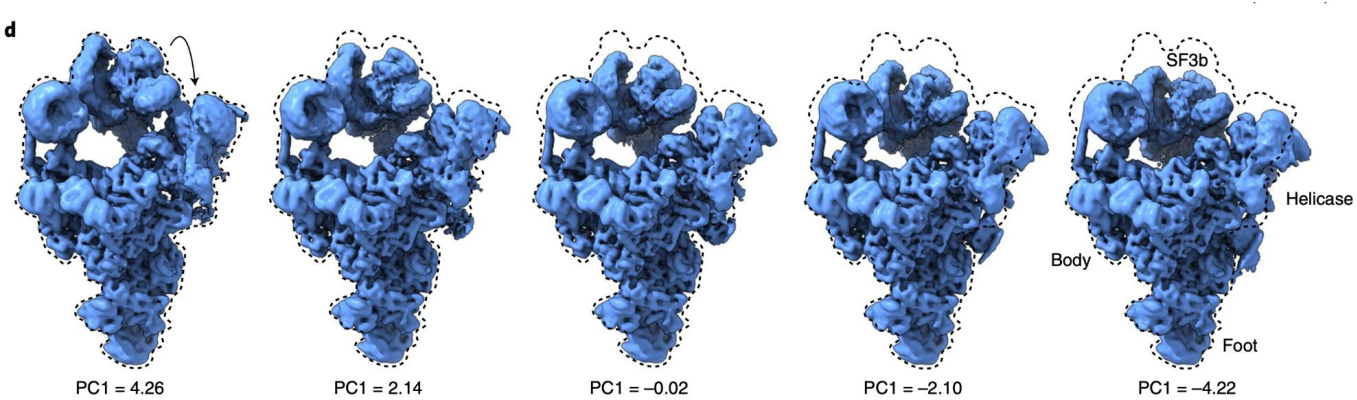
- C4: rare ~1% population (1,113 particles) missed by the original classification.
- Standard refinement of those same particles reproduces the structure.

From discrete states to continuous molecular motion

Nature Methods 2021 · Follow-up work · Figure 6c,d

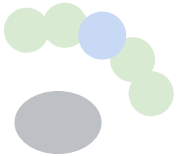


- Traversing latent PC1 reveals coordinated spliceosome motion.
- Latent trajectories show structural variation, not necessarily physical dynamics or preferred pathways.



What did cryoDRGN teach us?

One average

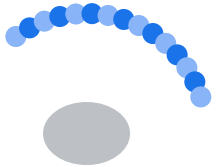


Flexible regions can blur

A few discrete classes

Within-class variation remains

A continuous family



Decode maps throughout latent space

- A continuous representation can reveal variation that averaging hides.
- Validate discoveries with particle subsets and conventional reconstruction.
- Pose errors and sparse latent regions still limit interpretation.

Strength, Limitations, and Open Questions

Strengths

- Flexible representation
- Discovery tool

Limitations

- Little information about kinetics
- Reliance on externally estimated particle poses

Open Questions

- Can pose and conformation be inferred jointly?
- Can we infer kinetics?

What came next?

Selected milestones through 2025

2020 / 2021

cryoDRGN

Continuous neural structure model

ICLR / Nature Methods

2021

cryoDRGN2

Jointly infer poses and structures on real cryo-EM data

ICCV

2024

cryoDRGN-ET

Learn heterogeneity from subtomogram tilt-series images

Nature Methods

2025

cryoDRGN-AI

Combine global pose search with local optimization

Nature Methods

- Later work removes pose assumptions and extends heterogeneous reconstruction to ab initio and in-cell cryo-ET settings.

Discussion

Q1: Describe two ways cryoDRGN differs from a standard VAE or from previous cryo-EM reconstruction approaches.

Q2: How does cryoDRGN represent a 3D volume with a neural network? What roles do the Fourier slice theorem and positional encoding play in the decoder?

Q3: How can we best interpret cryoDRGN's latent space? What does the trajectory from A to B represent?

Q4: If cryoDRGN finds a rare structural state that previous analysis missed (C4), what would convince you that the state is biologically meaningful rather than just an artifact of the analysis?

Q5: What new biological possibilities become possible if we can reconstruct continuous heterogeneity instead of a few discrete states?