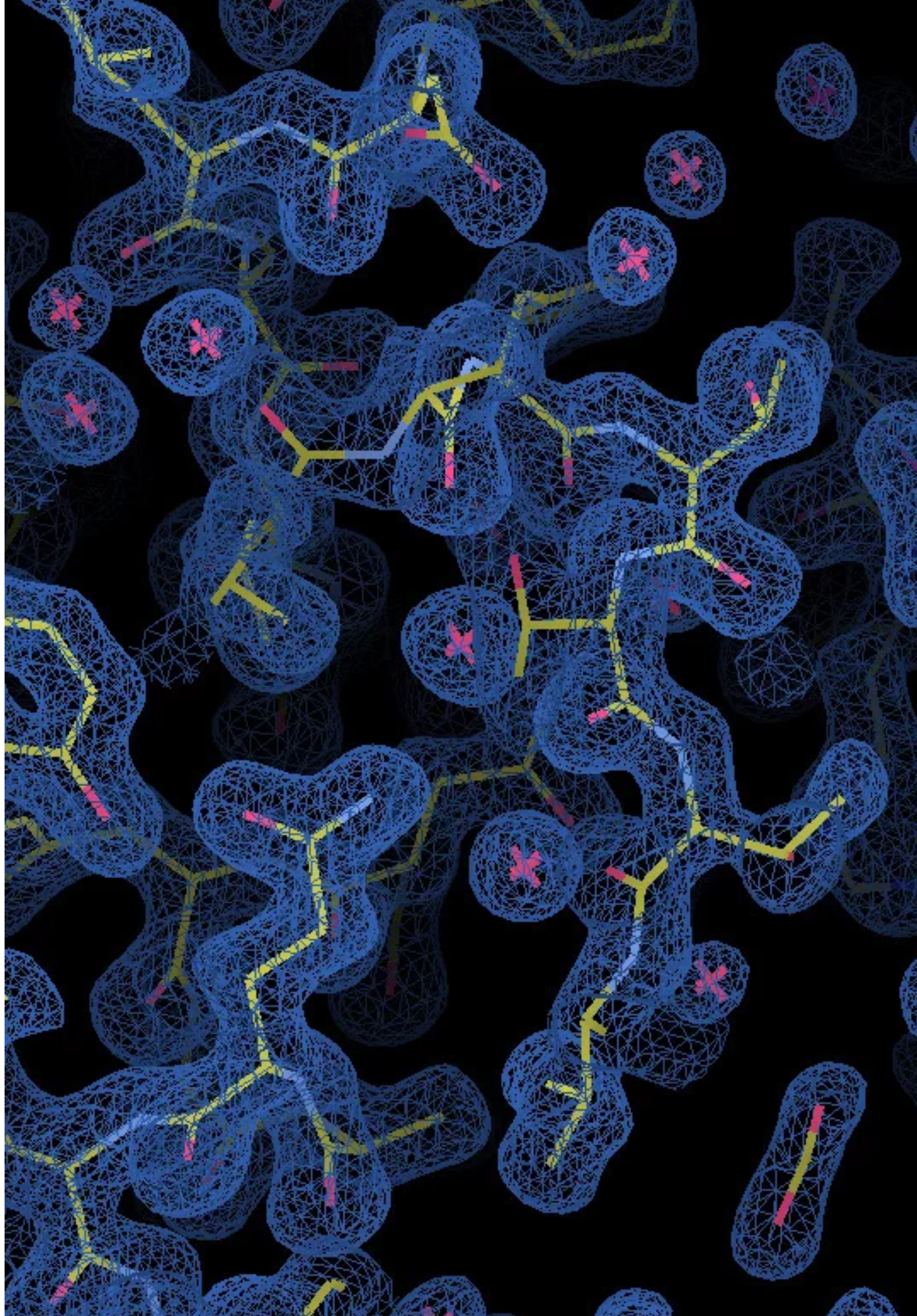
Protein Structure Determination with Cryo-EM

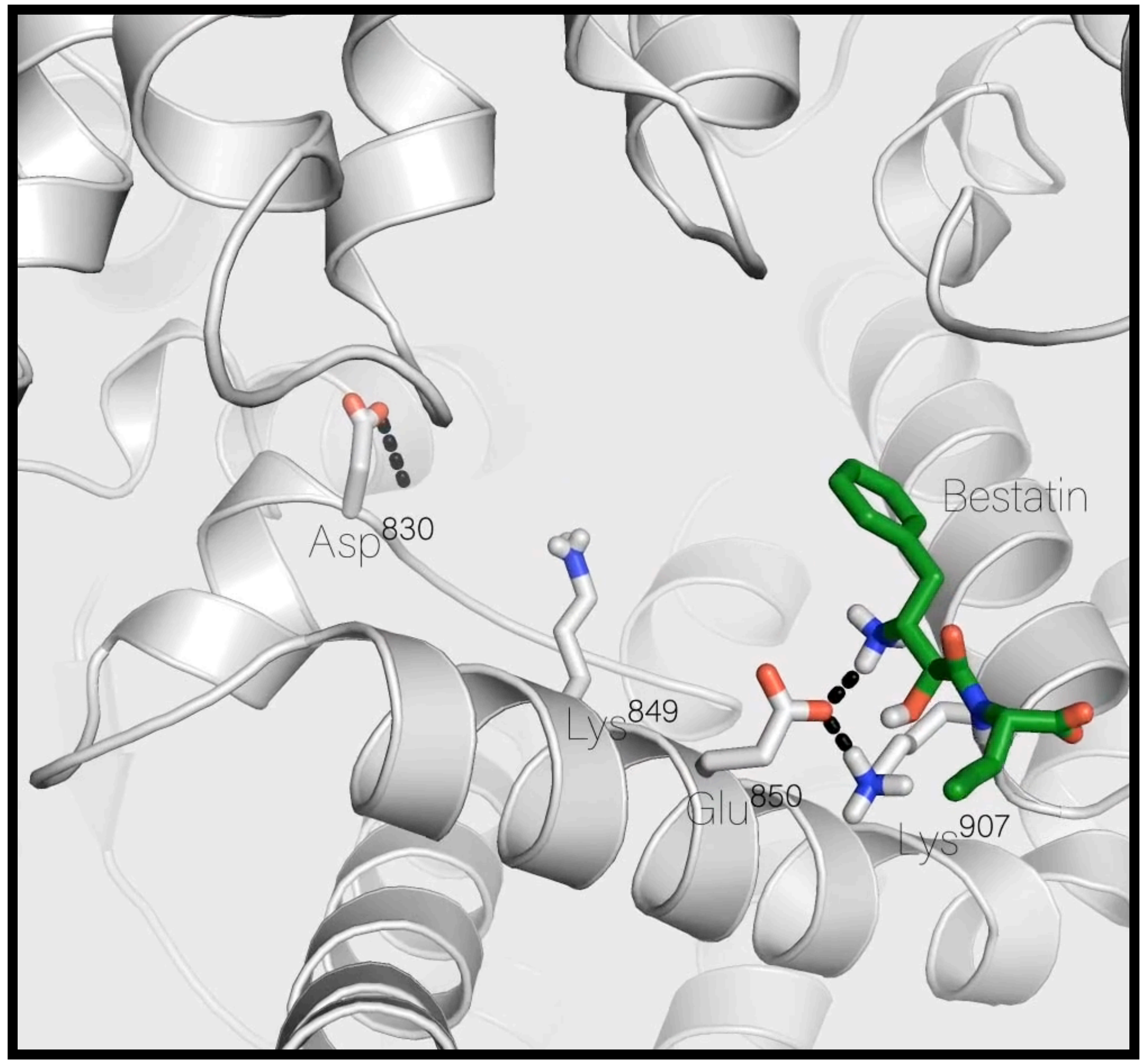
Rishwanth Raghu

COS598L, February 2026



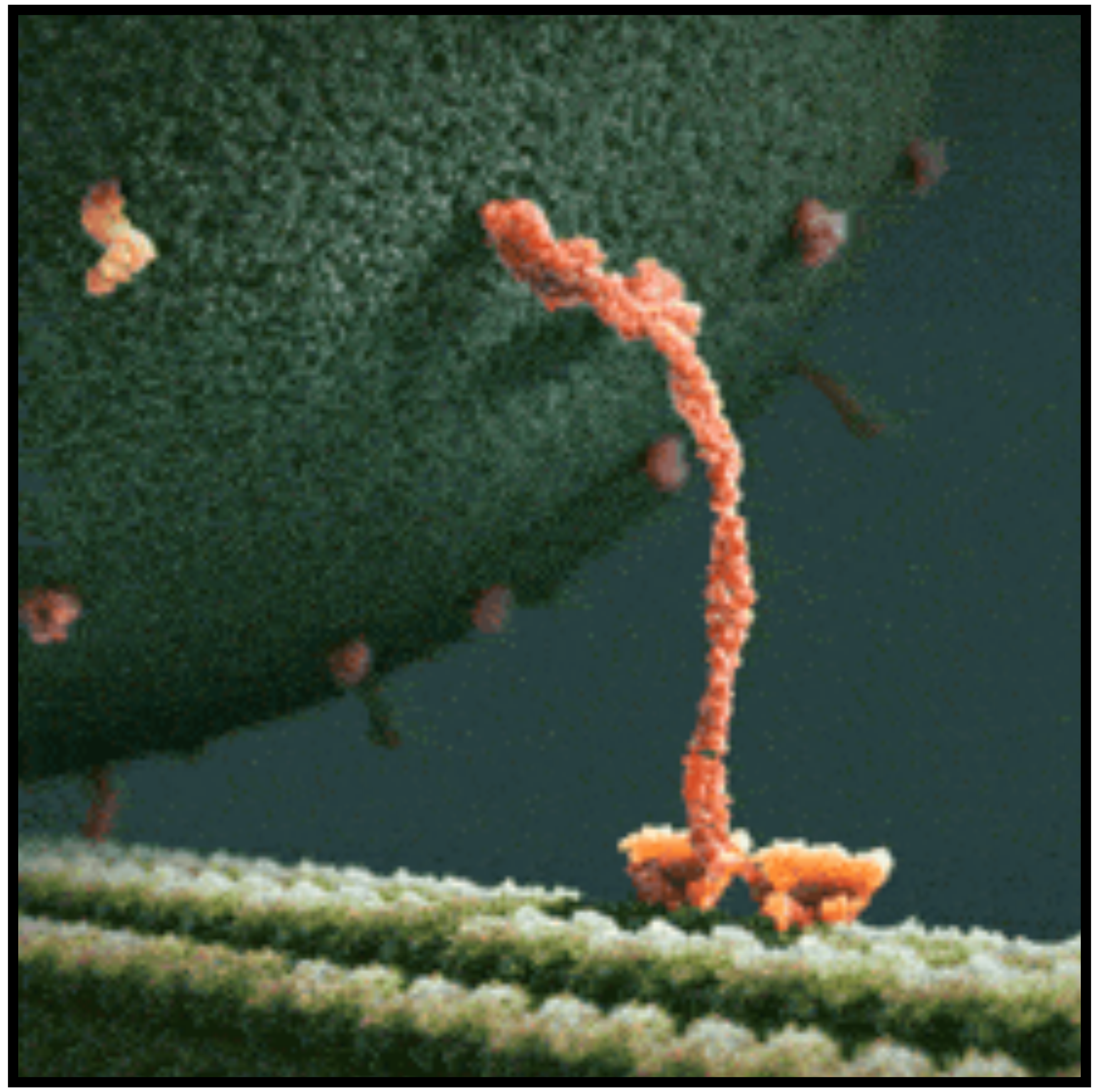
Protein structure and dynamics encode function across scales

M1 Aminopeptidase



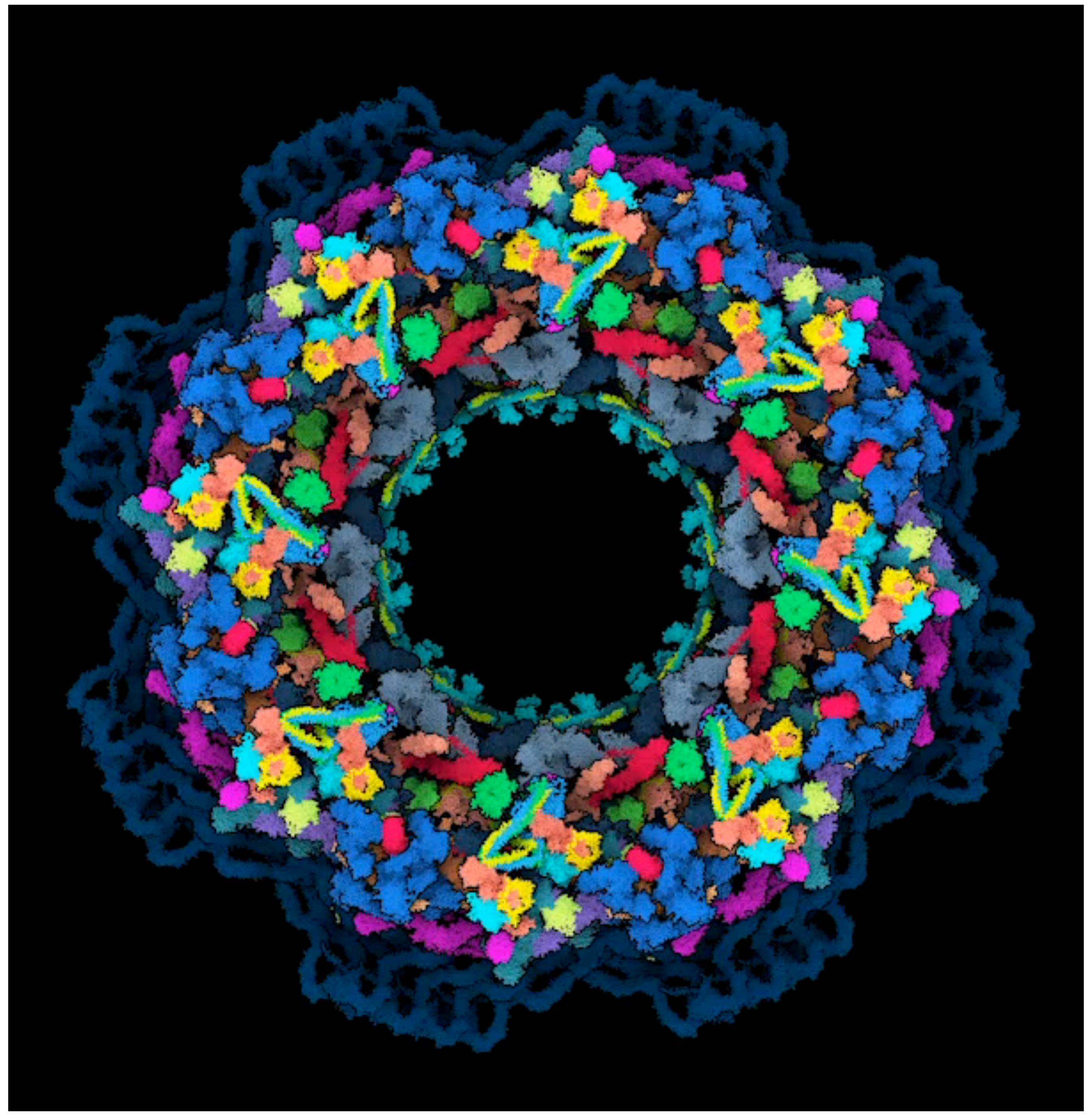
Moore et al. PLOS Comp Bio 2018

Dynein



Art of the Cell

Nuclear Pore Complex

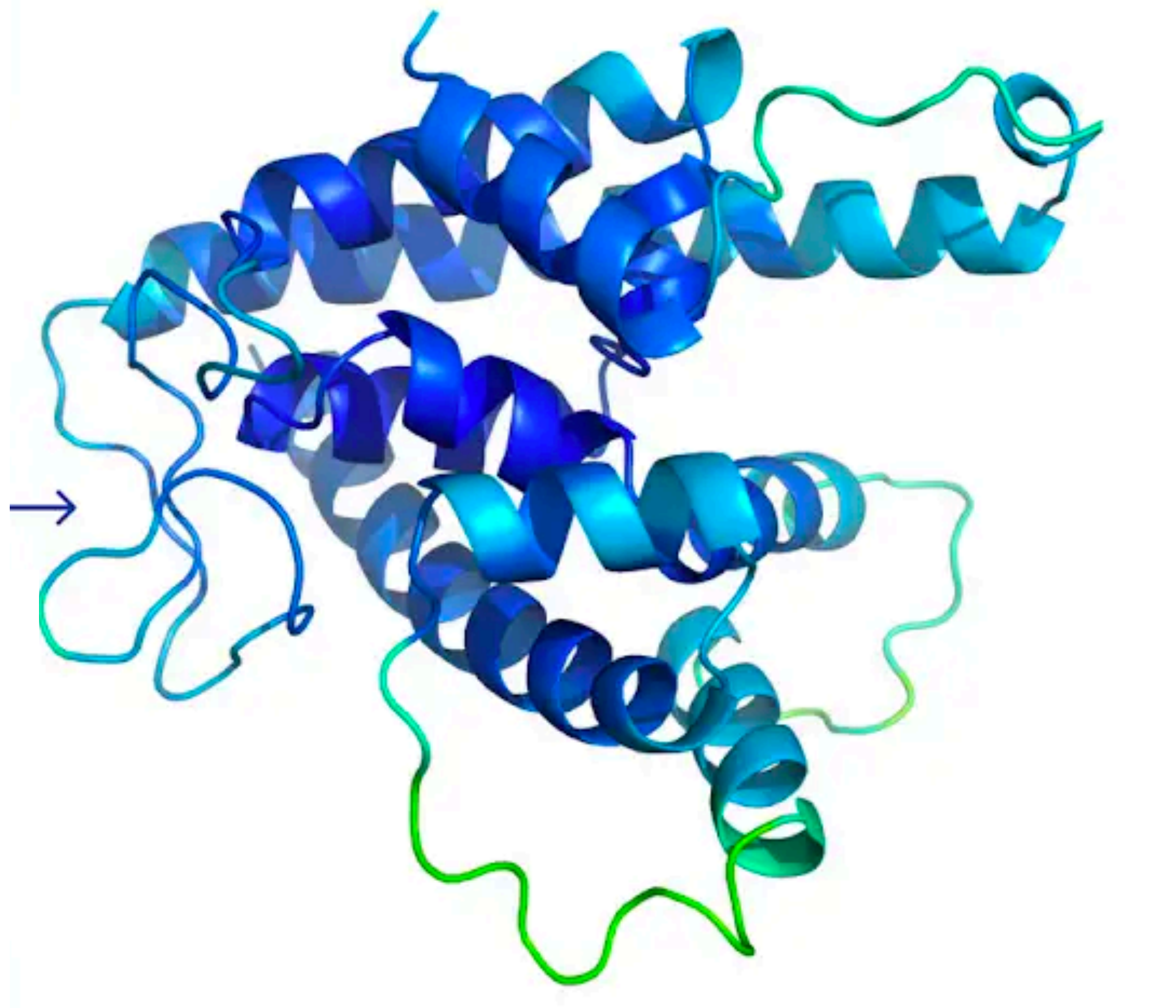


Mosalaganti et al. Science 2022

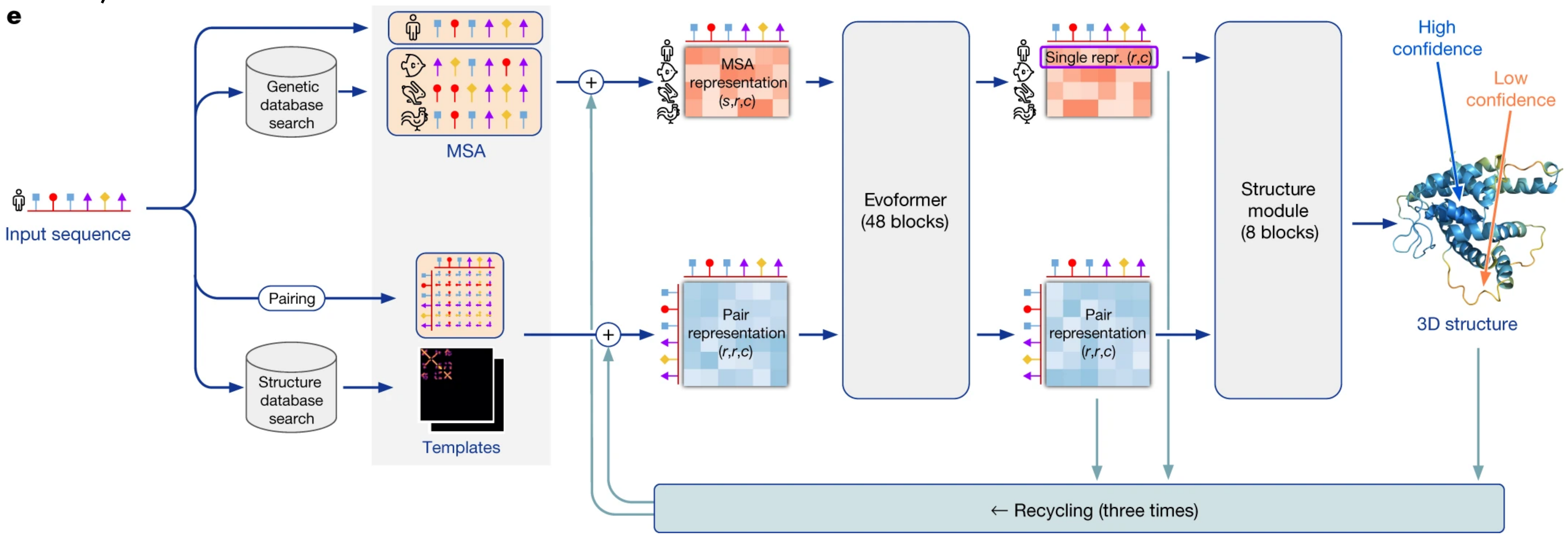
Protein Structure Prediction: AlphaFold

Input sequence
`MRKPRTPFTT…`

→ **Statistical Machine** →

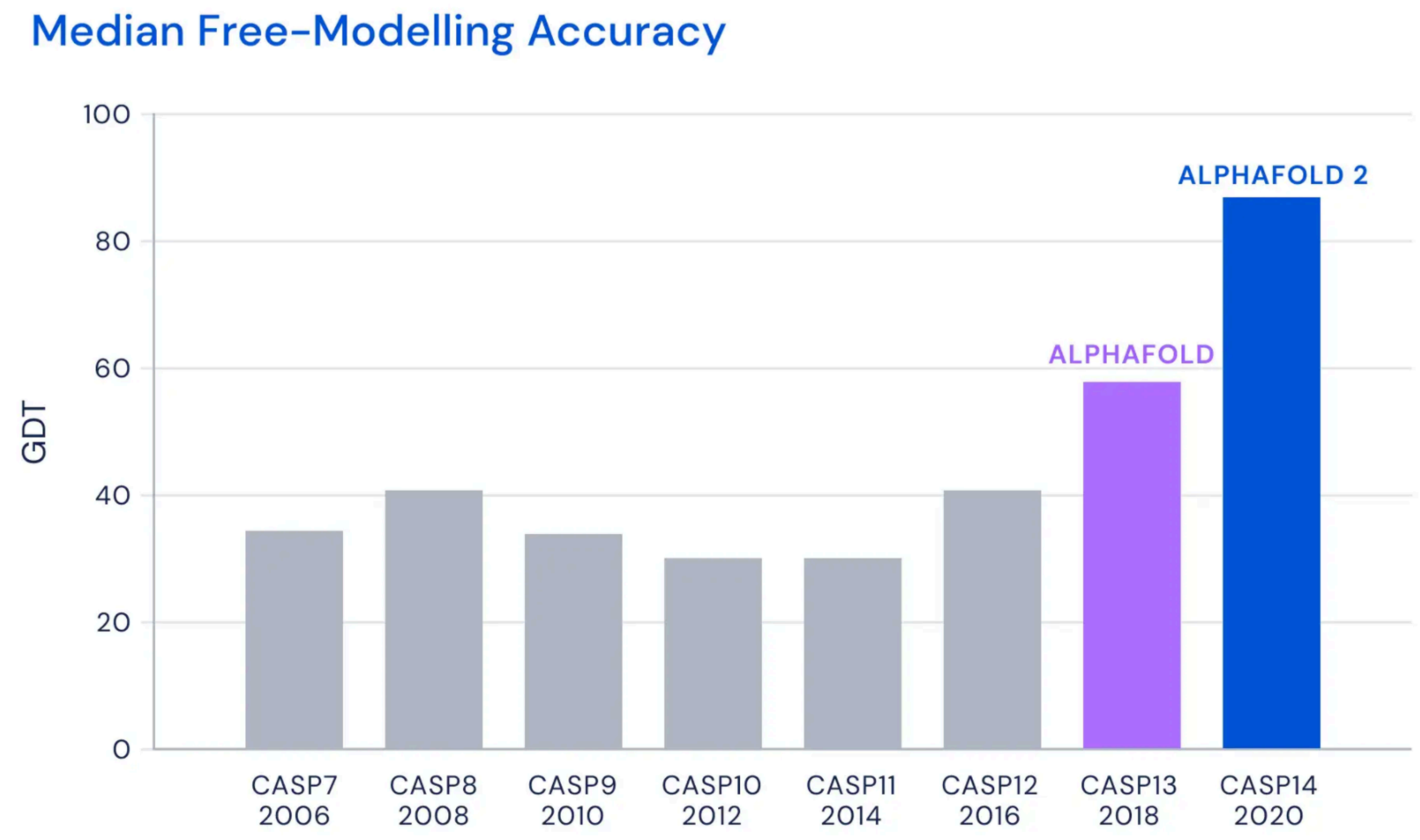


- Sequence-to-structure prediction
- MSA encoding + structure decoding
- Trained on structures from Protein Data Bank



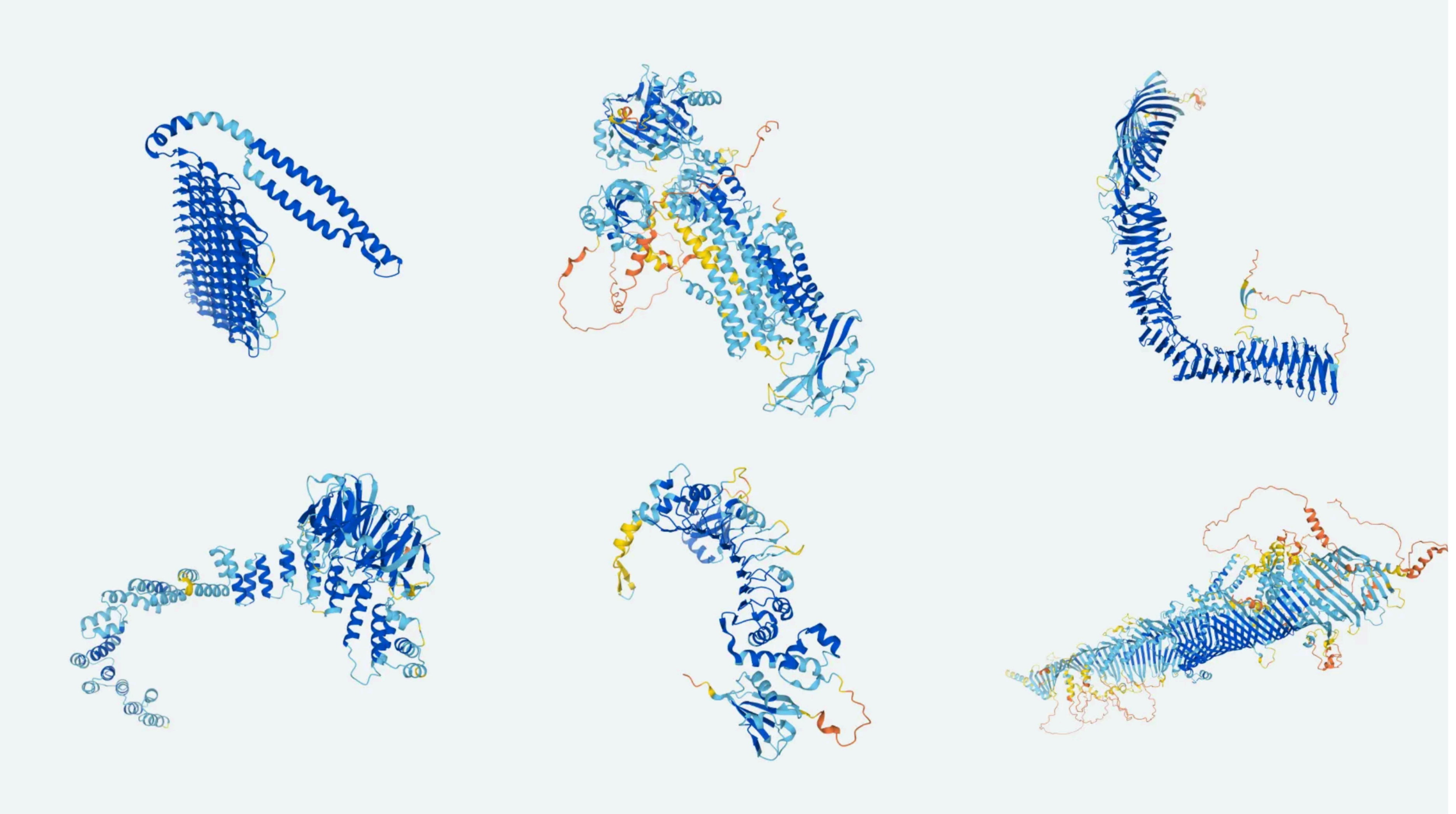
Jumper et al. “Highly accurate protein structure prediction with AlphaFold.” Nature 2021.

Protein Structure Prediction: AlphaFold

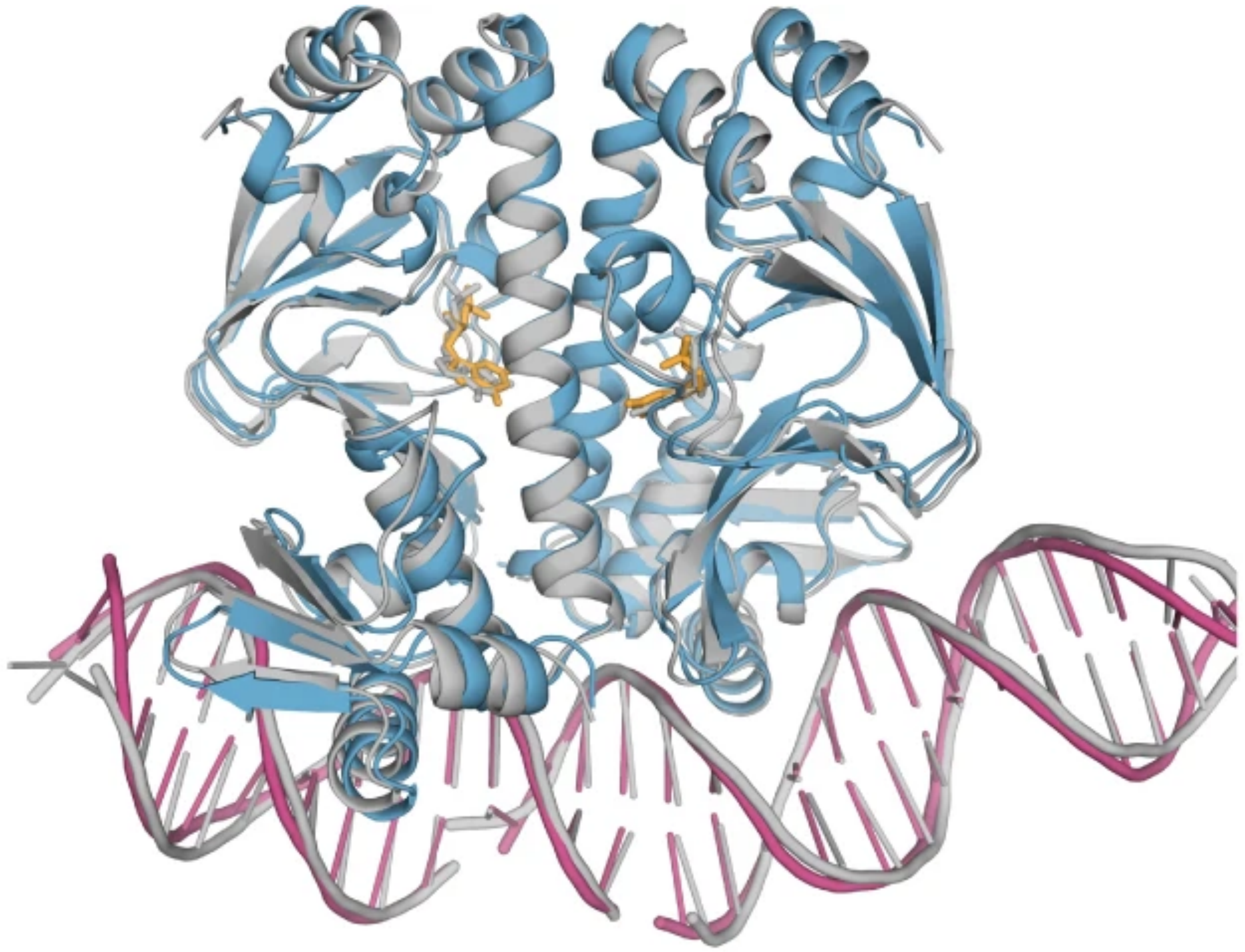


92.4 GDT @ CASP

(Within experimental error)



>214 million predicted structures in AFDB



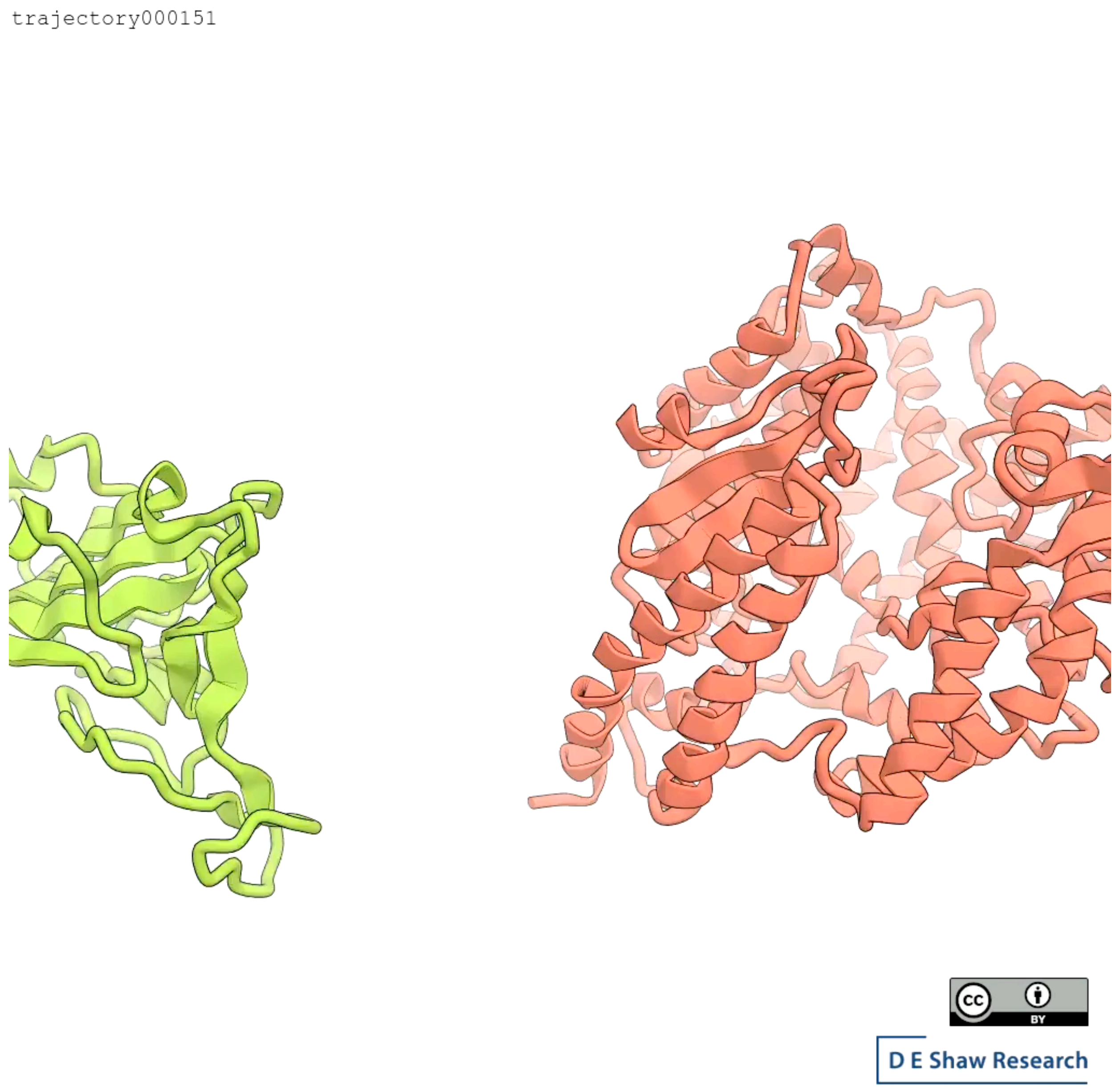
AF3 modeling of proteins, nucleic acids, and ligands

Tunyasuvunakool et al. “Highly accurate protein structure prediction for the human proteome.” Nature 2021.
Abramson et al. “Accurate structure prediction of biomolecular interactions with AlphaFold3.” Nature 2024.

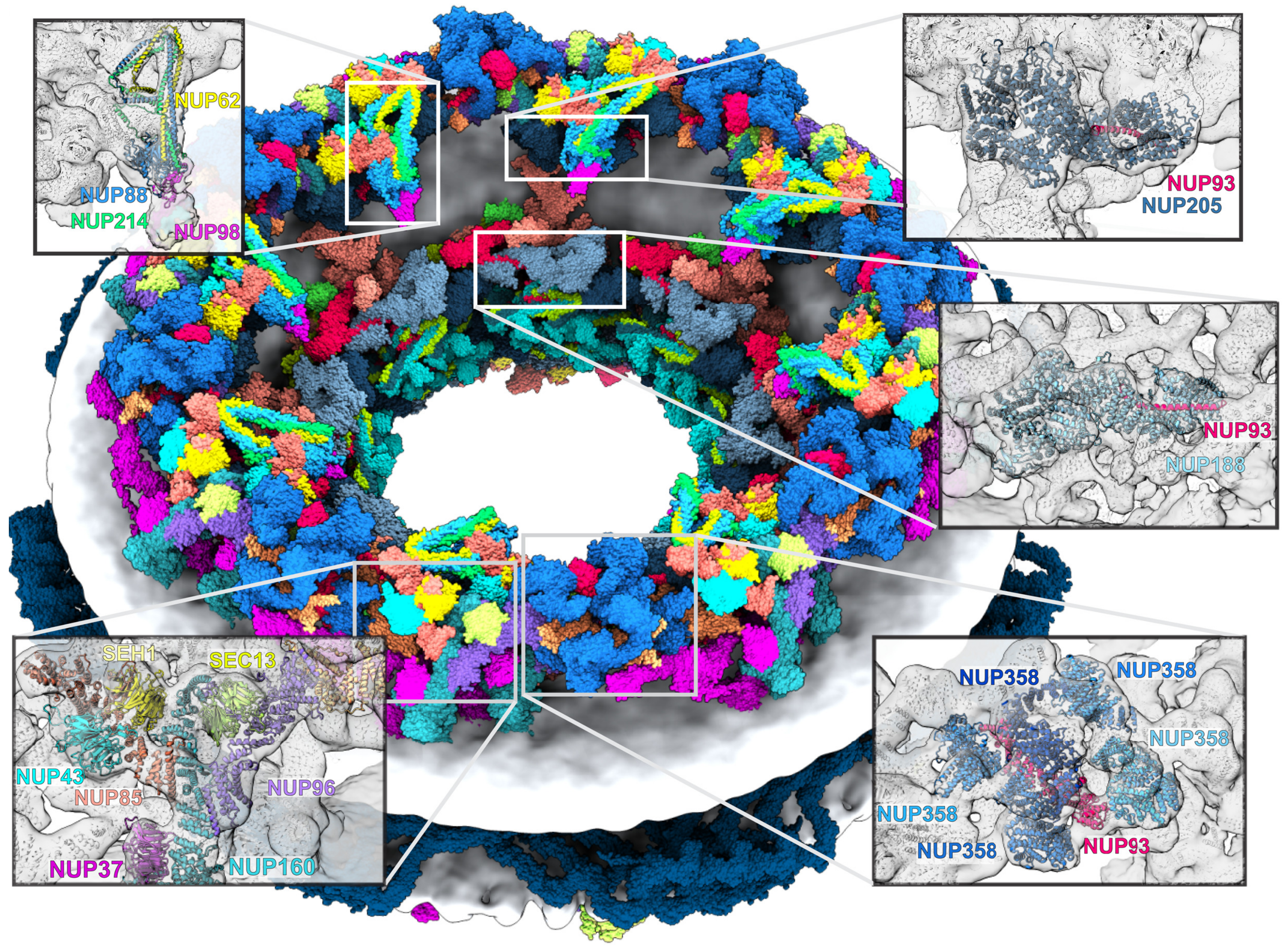
Did AlphaFold “solve” the protein structure prediction problem?

What is the role of structure *determination*?

- AlphaFold provides valuable hypotheses, but experiment remains the best form of structure validation in biological discovery
- AlphaFold has limitations in predicting structures that are dynamic, large, or have few sequence homologs
- Experiment is the source of "ground truth" for training ML models



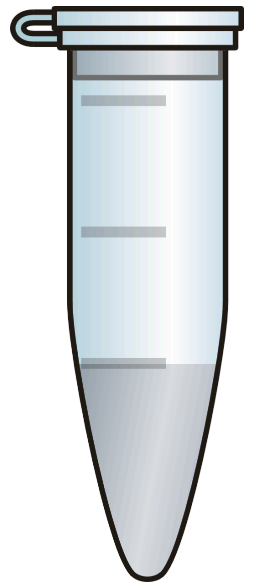
Molecular dynamics simulation. DE Shaw Research.



Nuclear pore supercomplex. Mosalaganti et al.

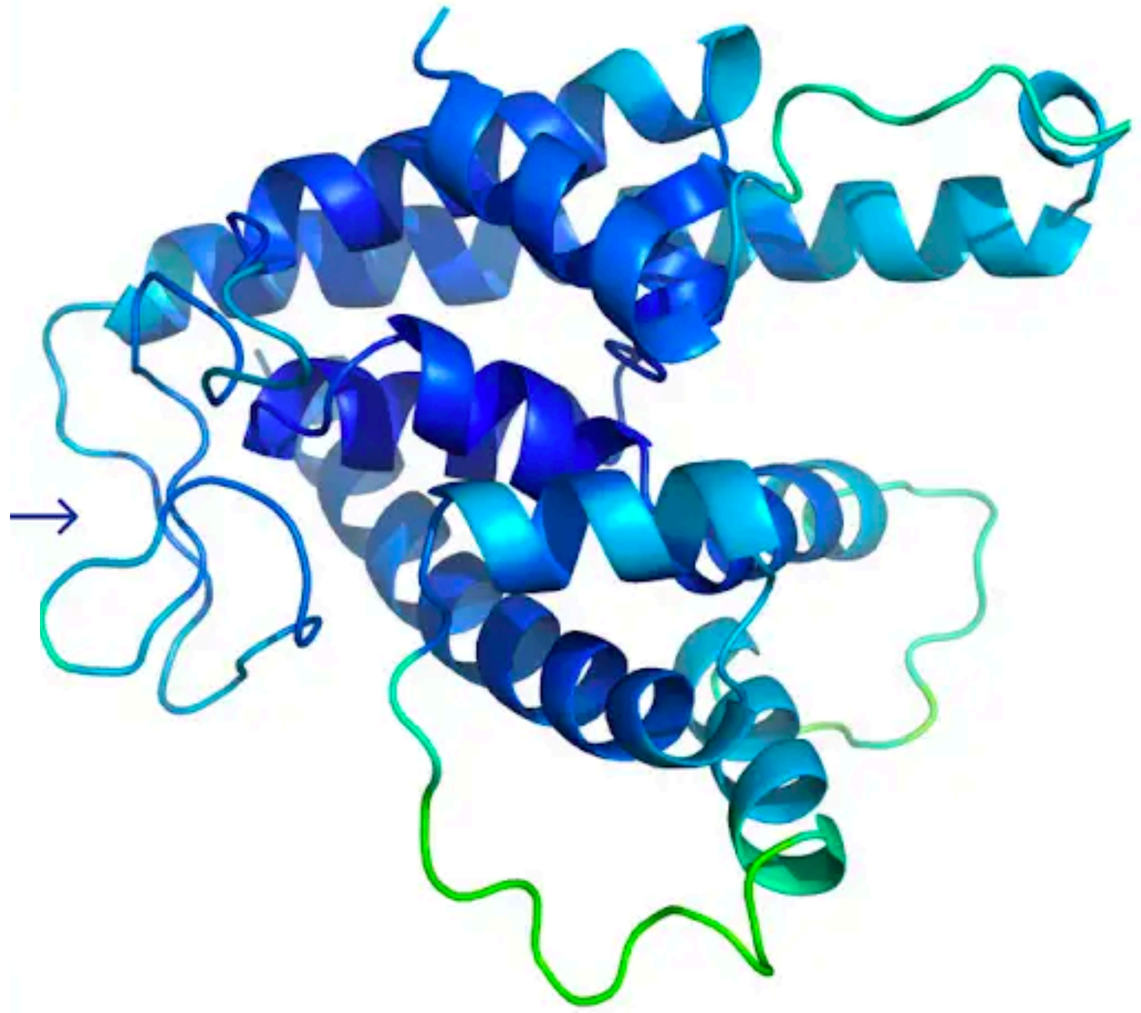
Illustration of the E. coli bacterium. David S Goodsell.

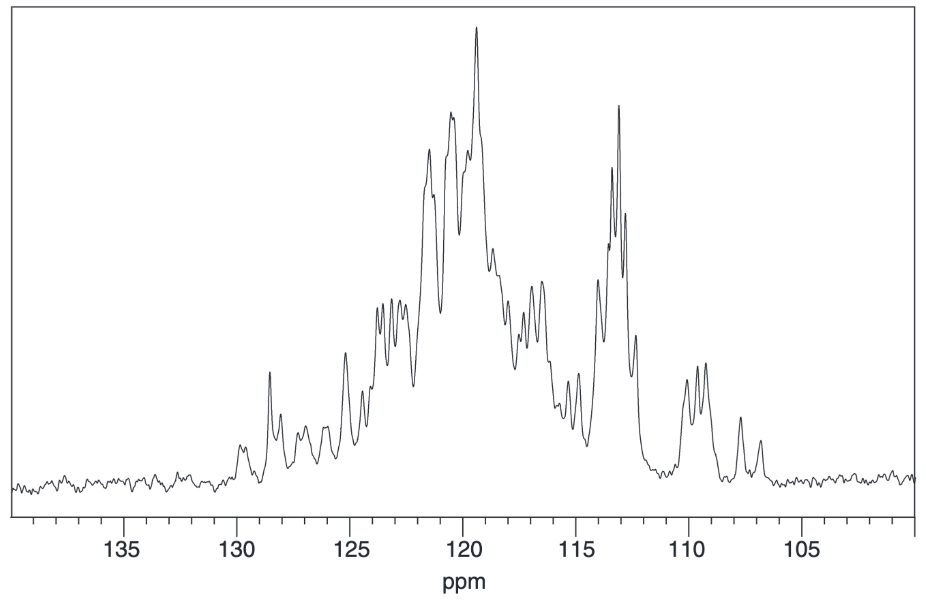
Protein Structure Determination



Protein sample

→ Experiment →

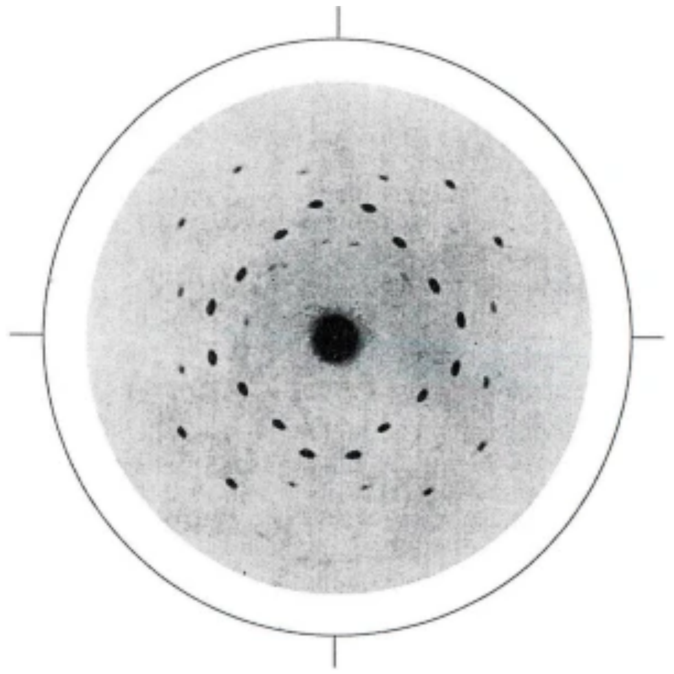




NMR

Solution state

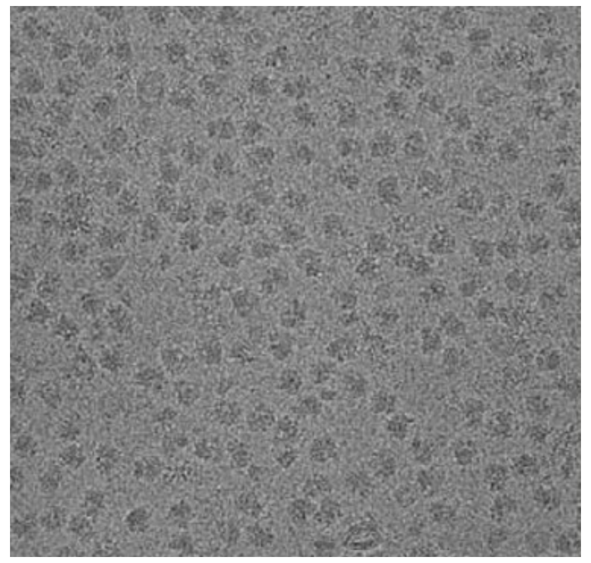
Limited to smaller proteins



Crystallography

High resolution

Requires crystallization



Cryo-EM

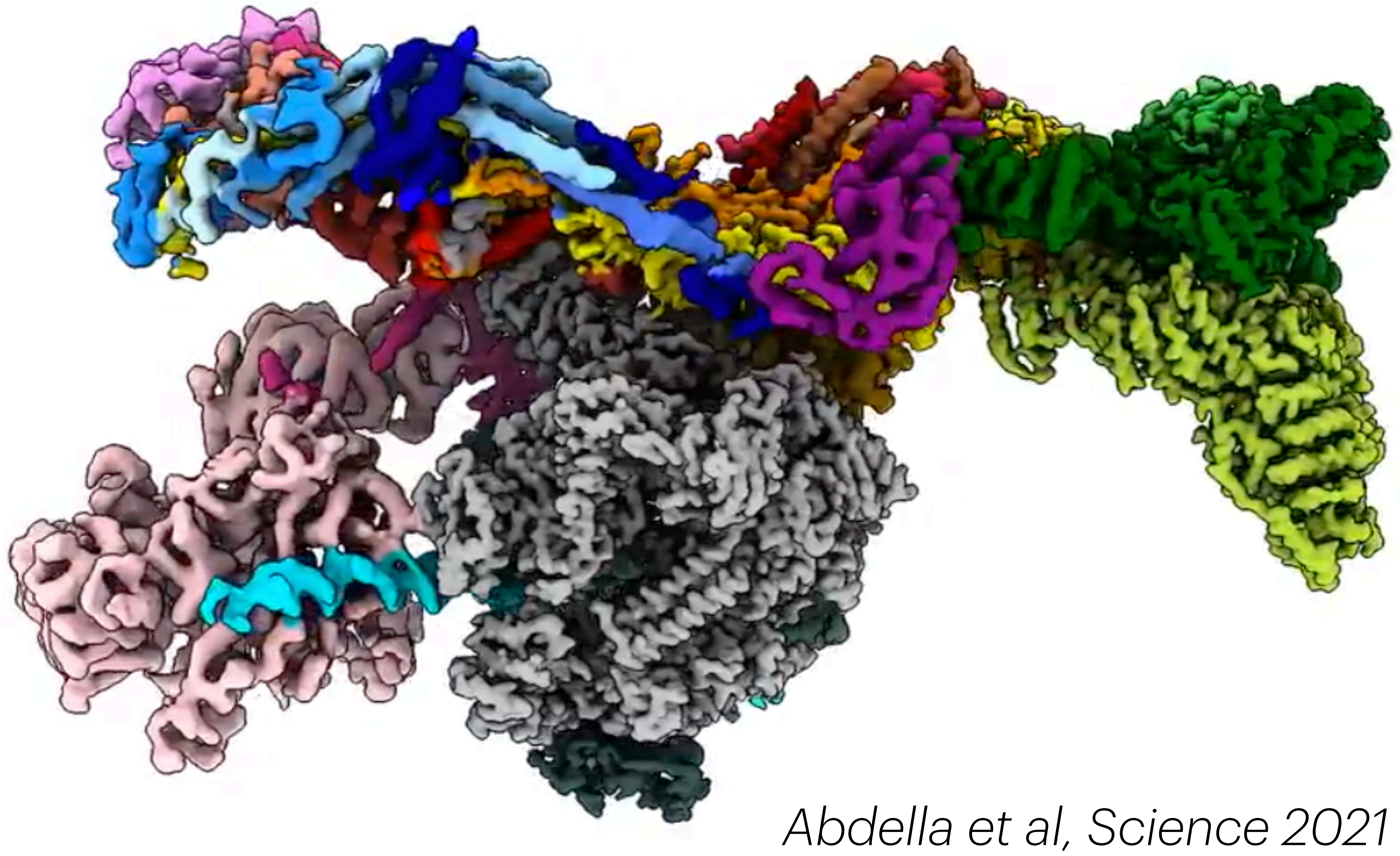
Captures dynamics

Limited to larger proteins

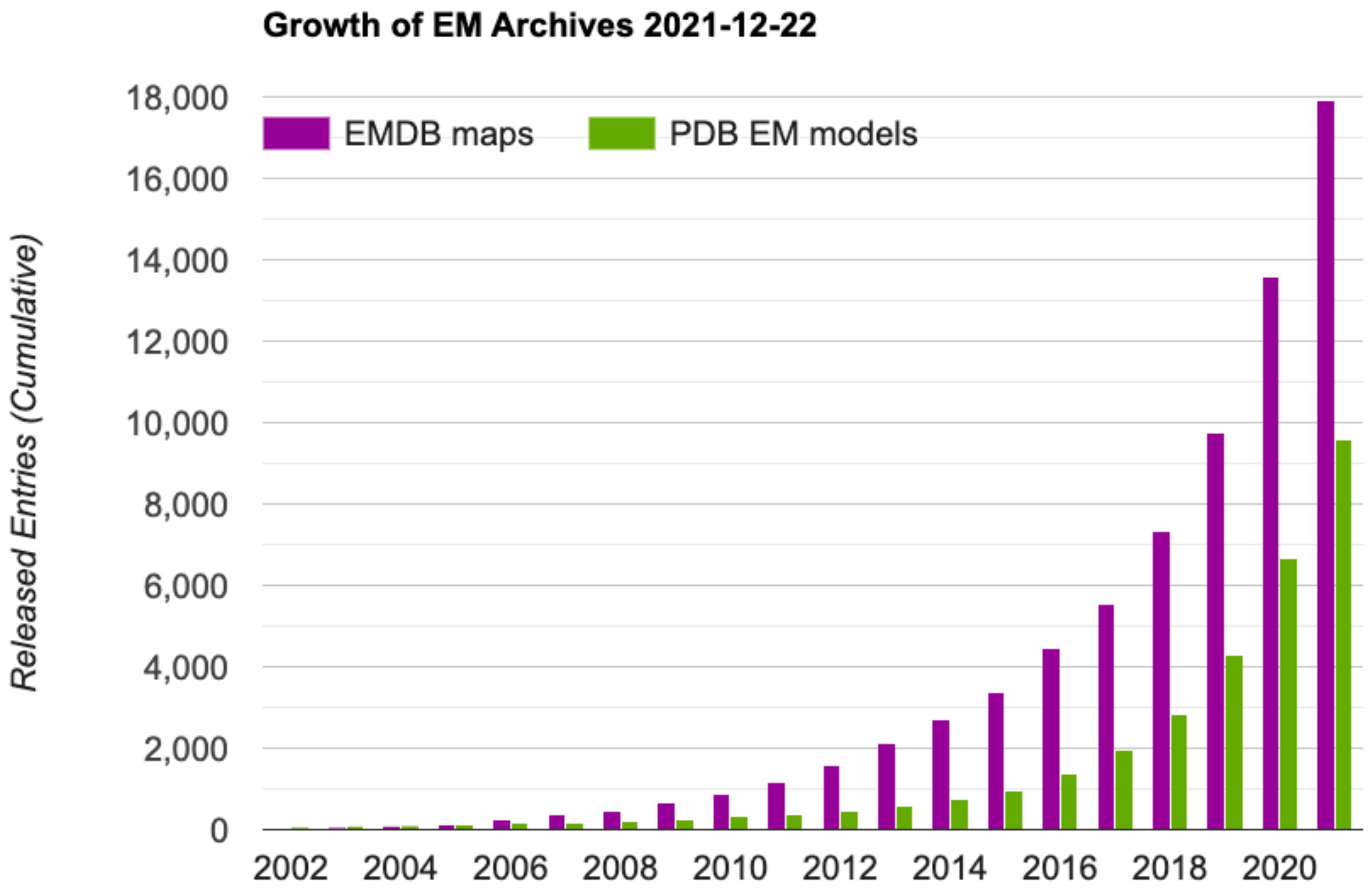
Wong et al. Elife 2014
Thomas. Nature 2012
Edwards and Reid. Protein Science 2000

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



Roadmap

- Overview of the cryo-EM pipeline
- Classical cryo-EM reconstruction
    - Homogeneous reconstruction
    - Multiclass reconstruction
- Modern heterogeneous reconstruction
    - CryoDRGN

Roadmap

- **Overview of the cryo-EM pipeline**
- Classical cryo-EM reconstruction
  - Homogeneous reconstruction
  - Multiclass reconstruction
- Modern heterogeneous reconstruction
  - CryoDRGN

Sample Preparation

Sample Preparation

Data Collection

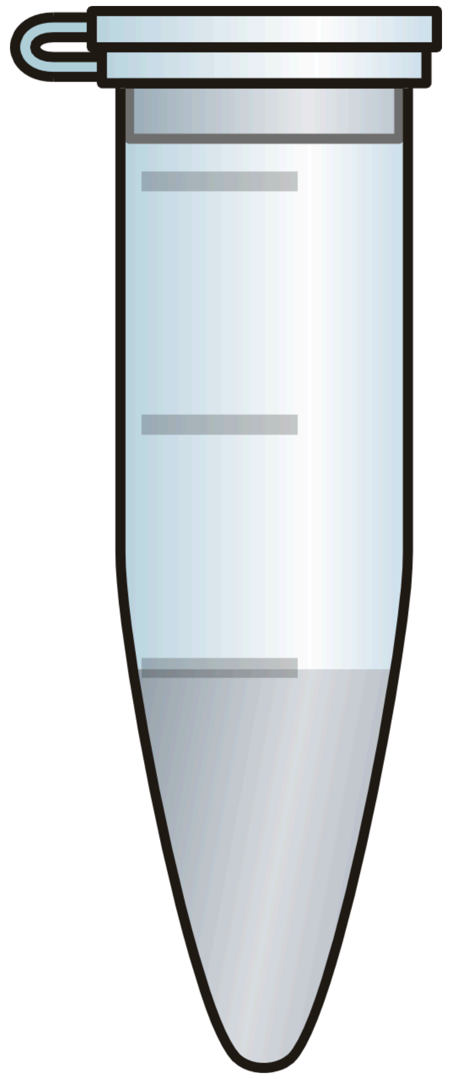
Particle Picking

3D Reconstruction

Model Building

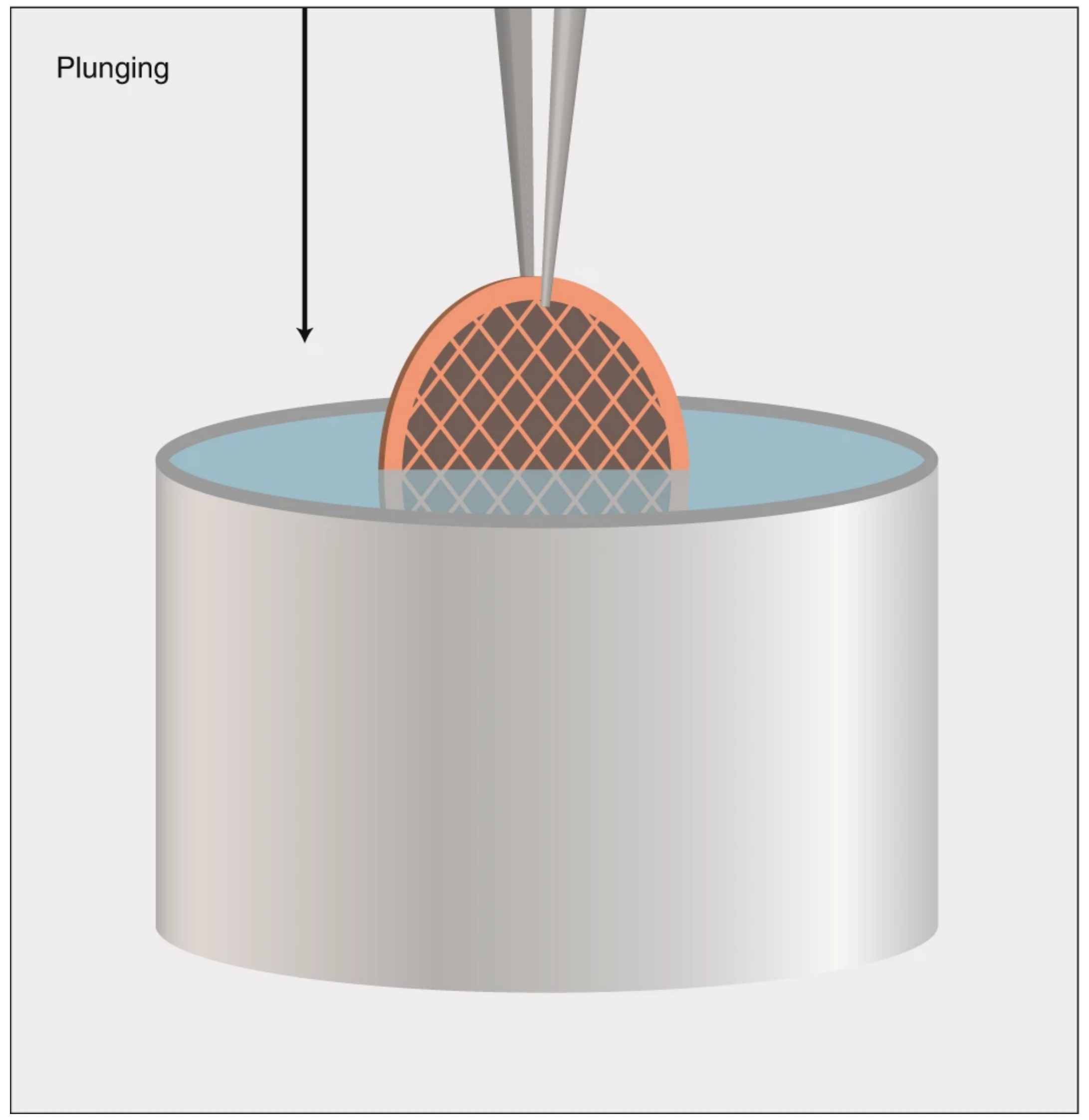
The protein of interest is purified and flash frozen into a thin specimen

- Freezing traps molecules in place, and increases radiation tolerance
- More proteins are amenable to freezing than crystallization
- Molecules are sampled from an <u>“equilibrium” distribution of conformations</u>



Purified Protein

Vitrification →



Cryo-EM Sample

Data Collection

Sample Preparation

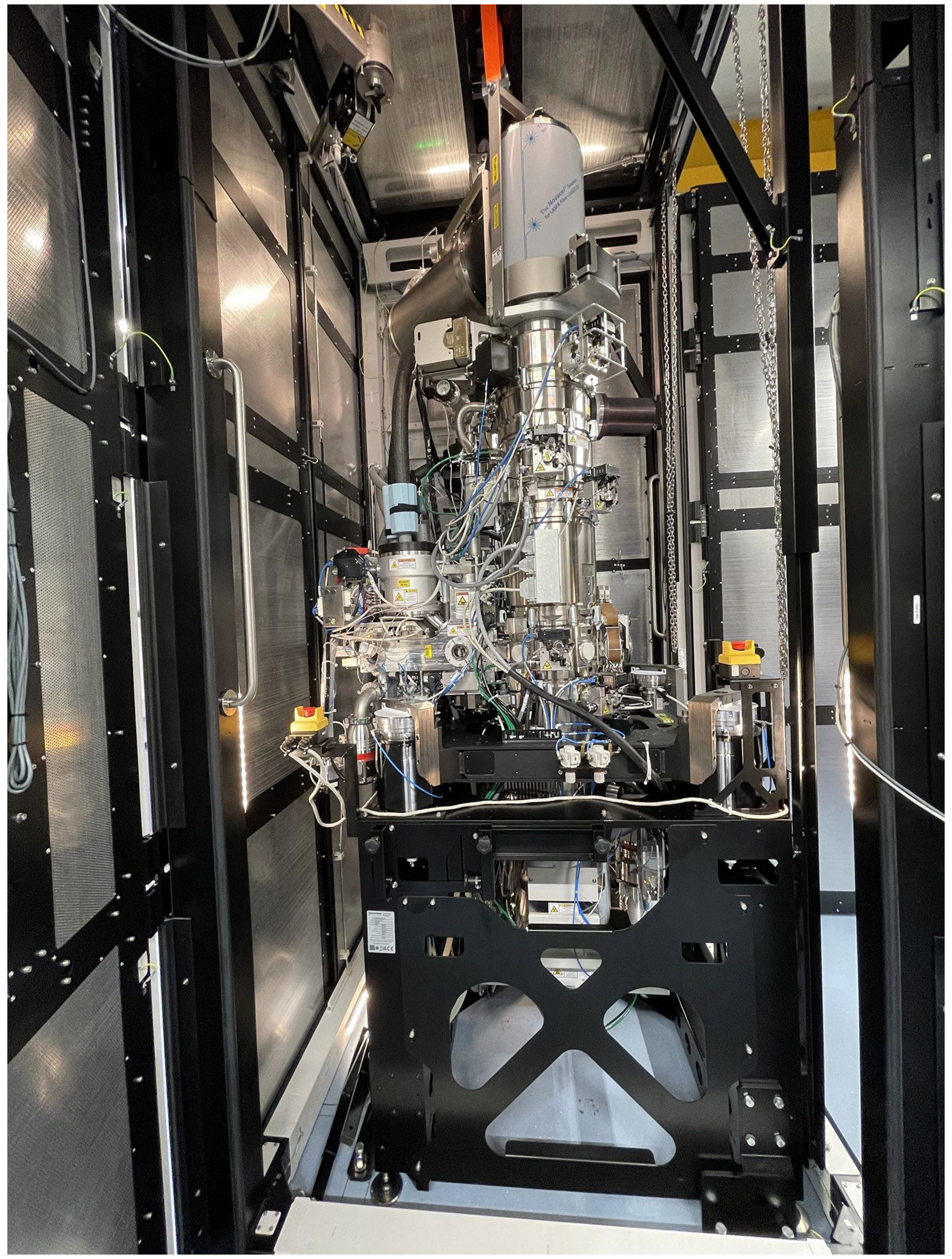
Data Collection

Particle Picking

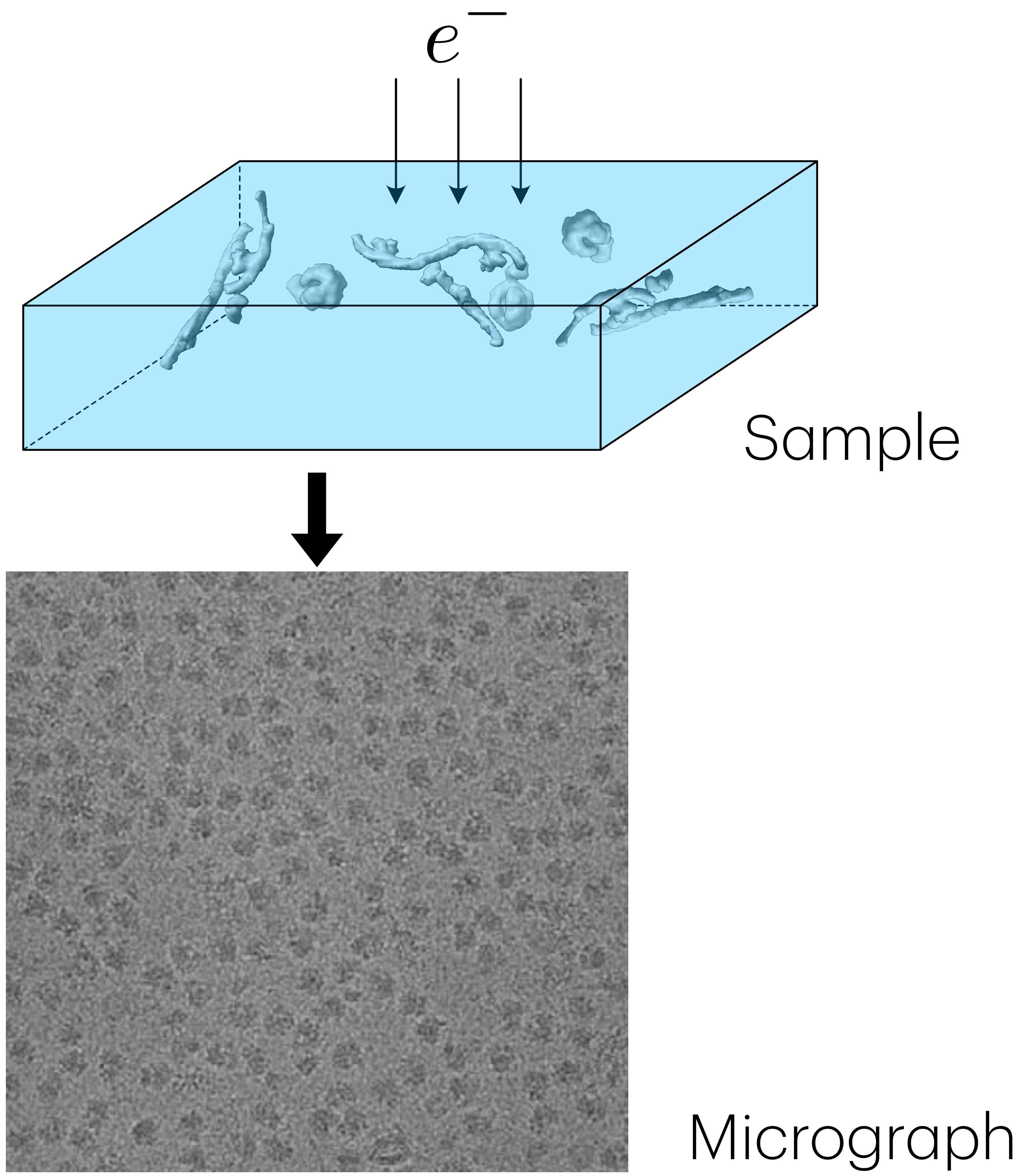
3D Reconstruction

Model Building

The sample is irradiated with electrons to produce 2D projection data of electron scattering potential



Laboratory of Biological Electron Microscopy



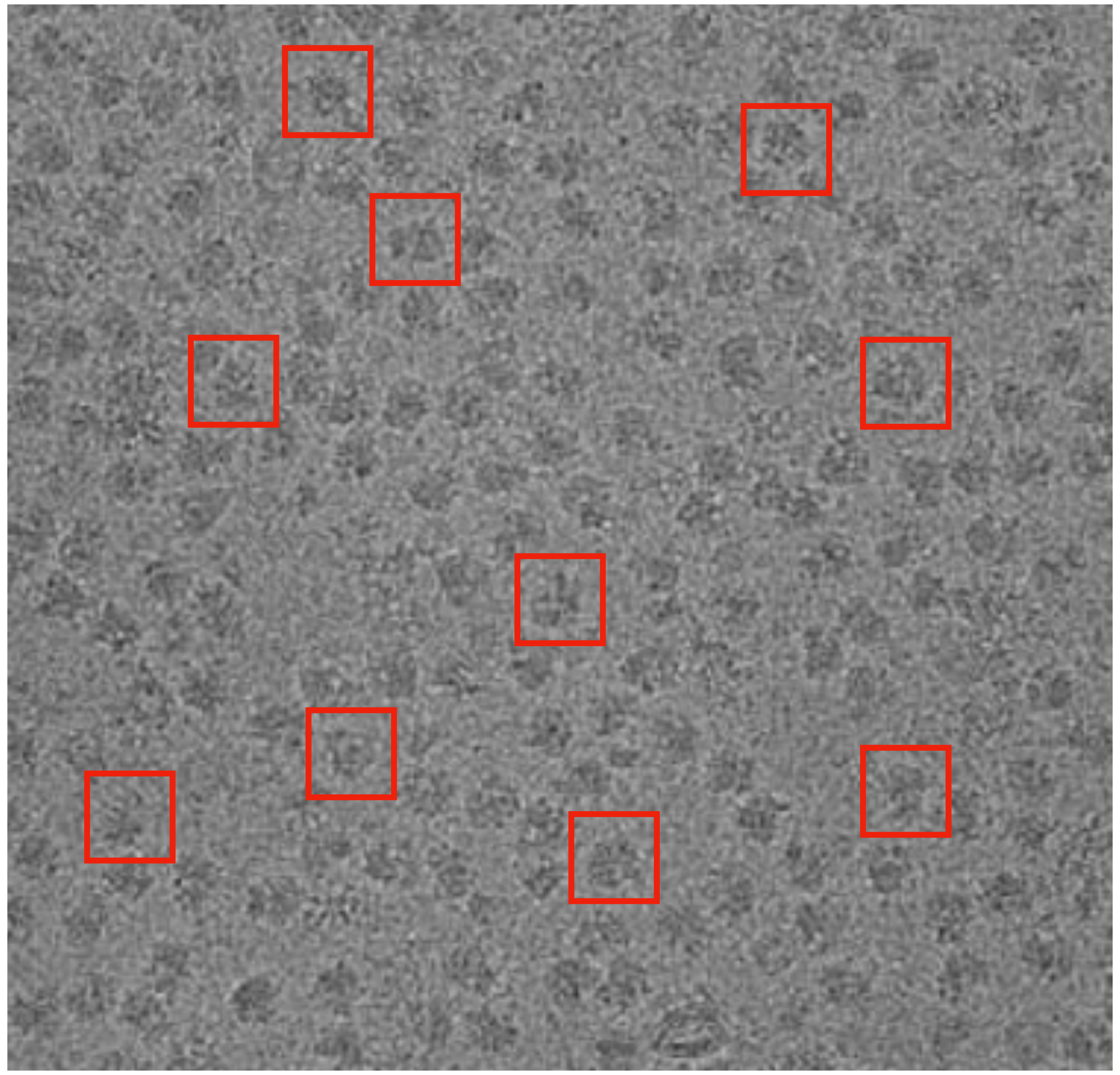
[EMPIAR-10028]

Particle Picking

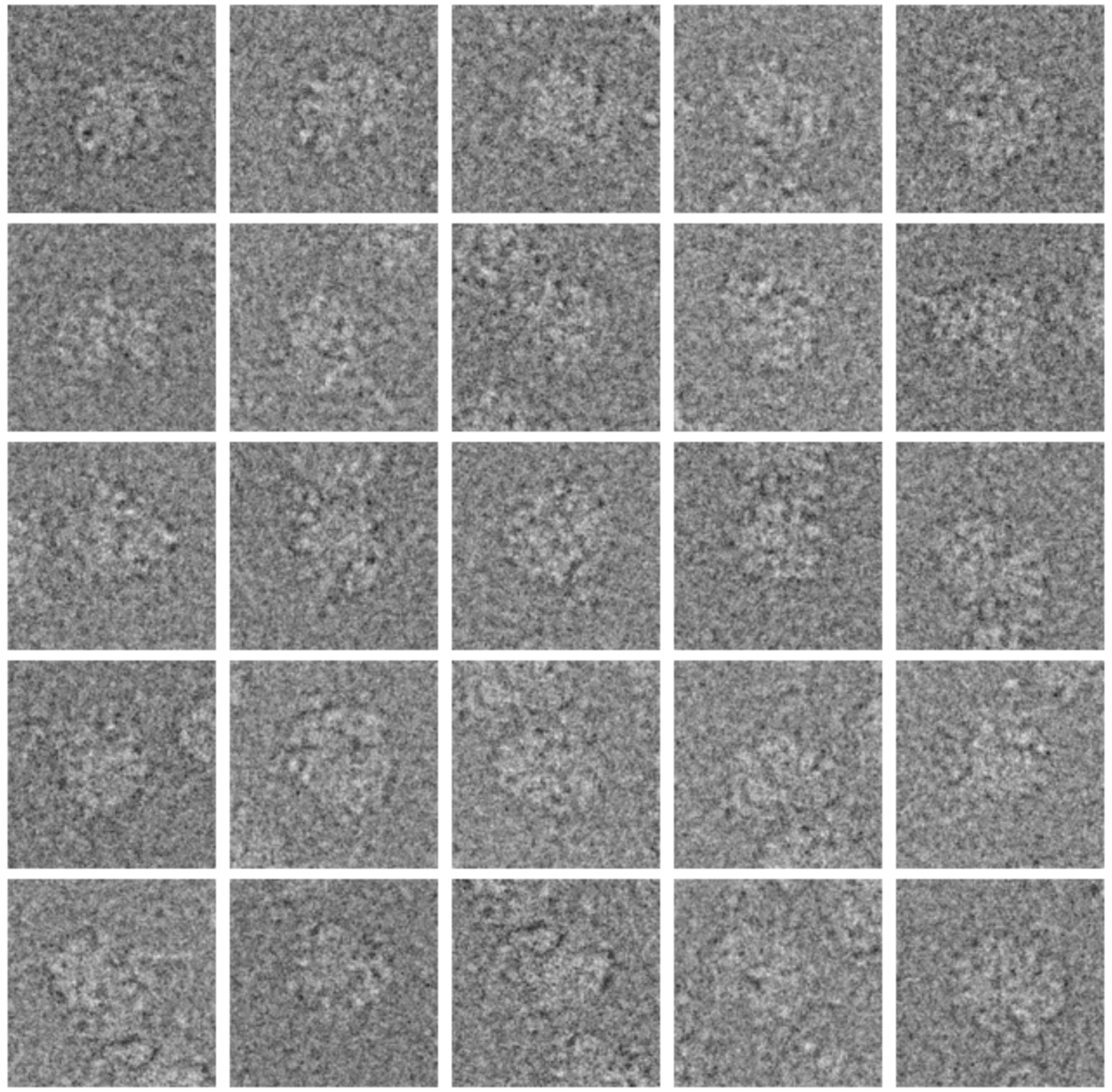
Sample Preparation
Data Collection
Particle Picking
3D Reconstruction
Model Building

Individual particles are extracted from micrographs using manual selection, machine learning methods, or prior structural templates

[EMPIAR-10028]



Micrograph



10^4-10^7 Particle Images

2D Classification

Sample Preparation

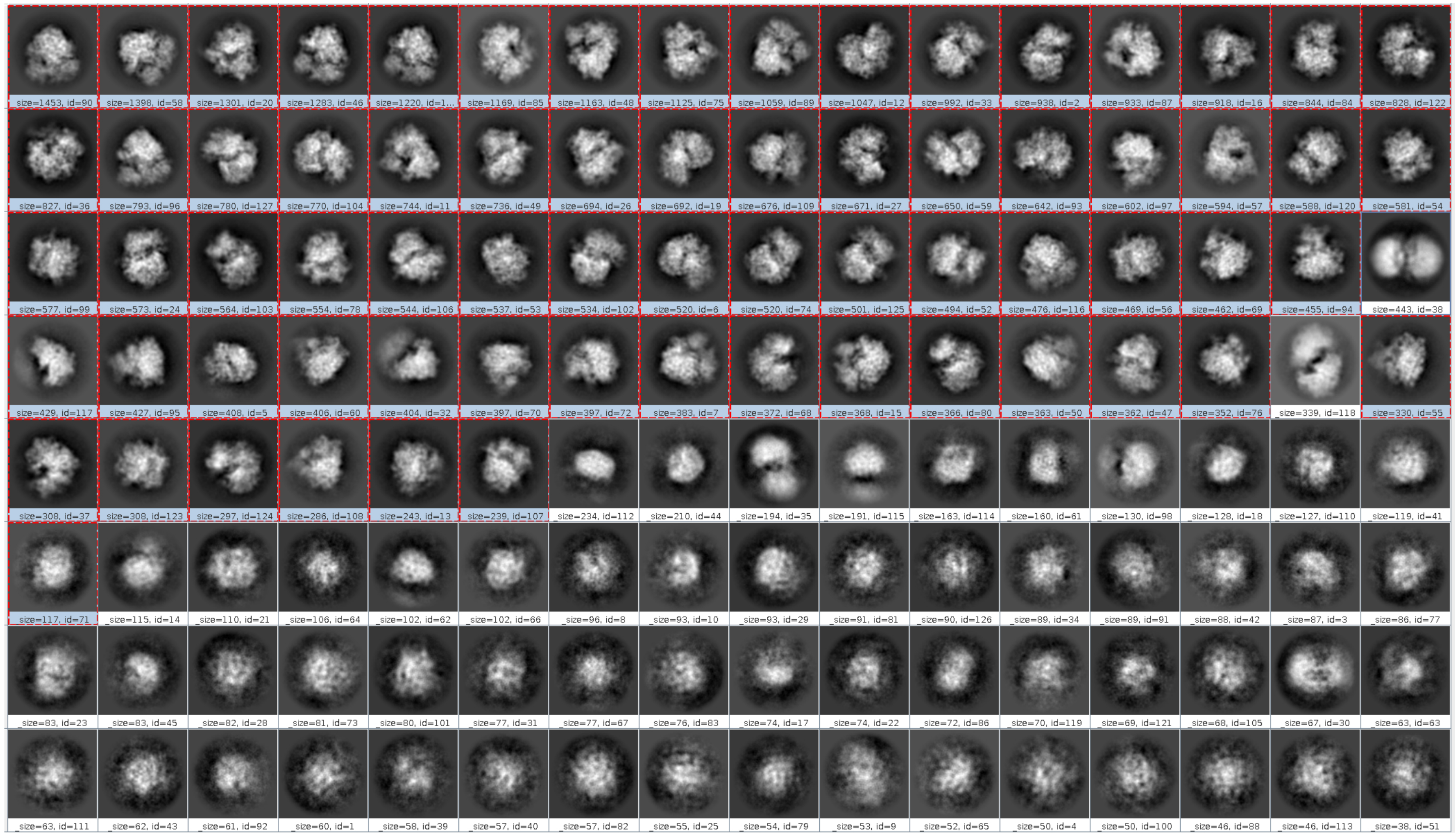
Data Collection

Particle Picking

3D Reconstruction

Model Building

Cryo-EM datasets often contain junk or outlier particles, which may be filtered out by selecting good classes of averaged particles



Jiménez-Moreno et al. JoVE Biochemistry, 2021

3D Reconstruction

Sample Preparation

Data Collection

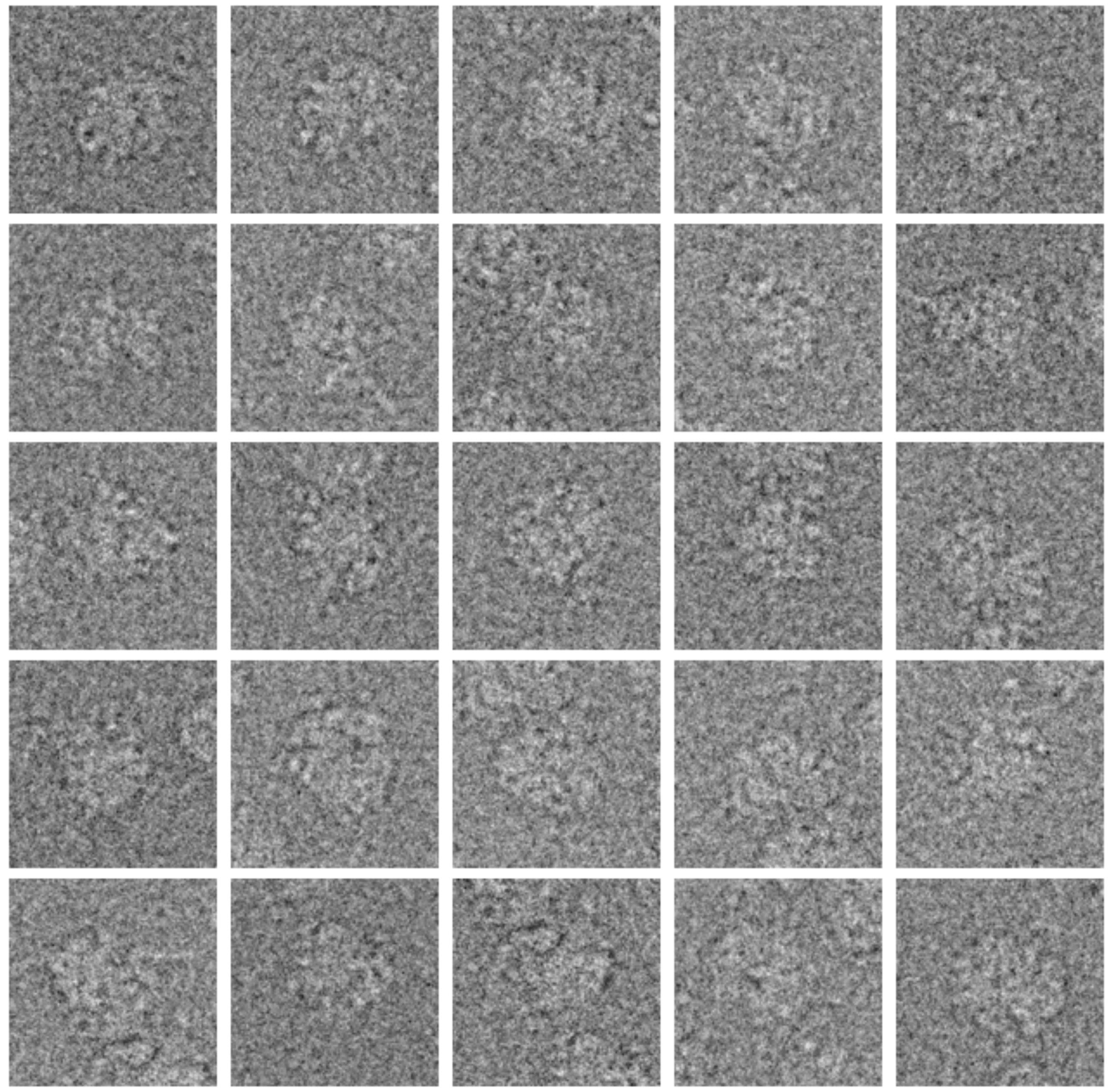
Particle Picking

3D Reconstruction

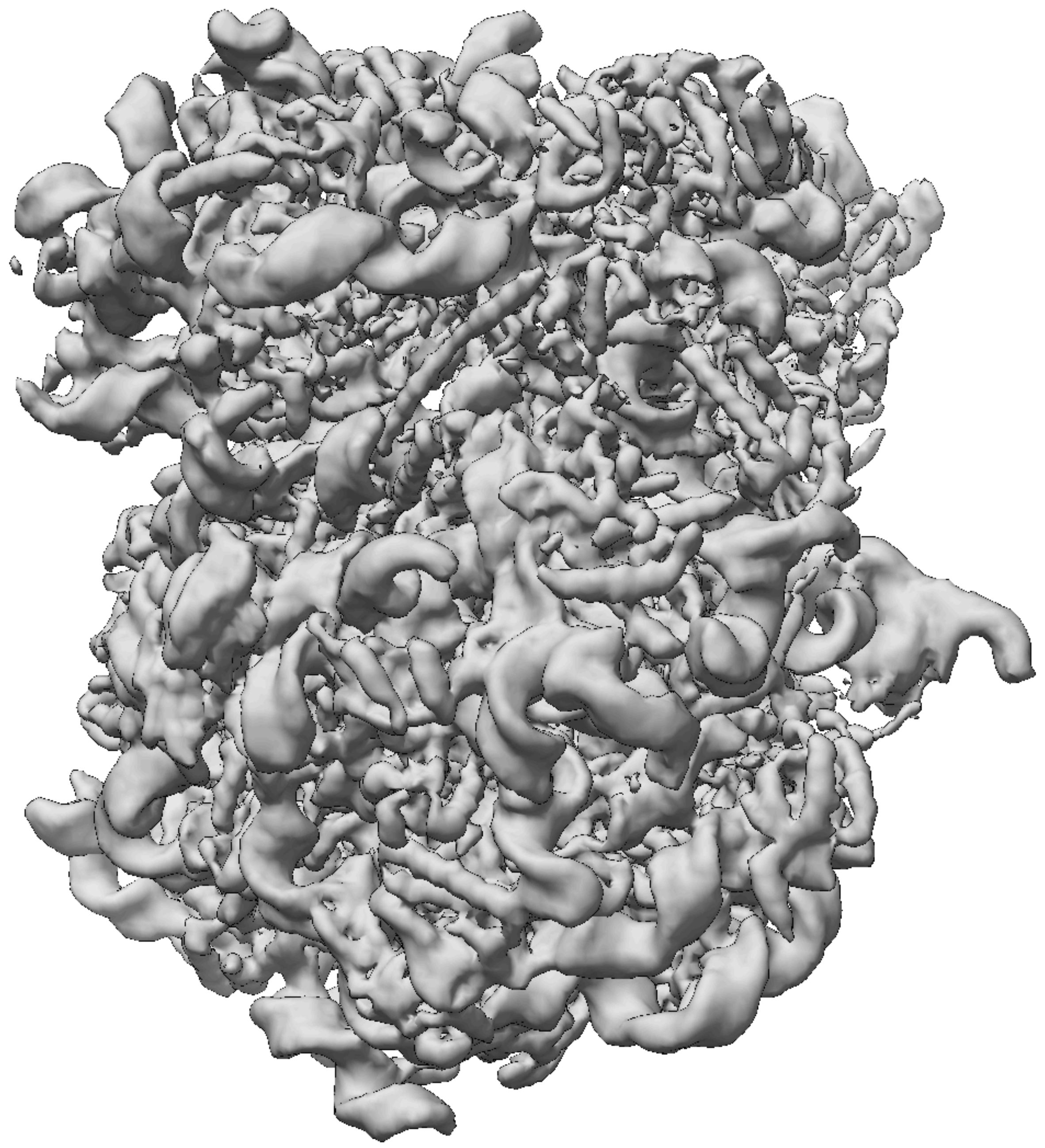
Model Building

The noisy 2D projection images are reconstructed into a 3D volume representing the protein’s structure

[EMPIAR-10028]



2D Particle Images



3D Density Map

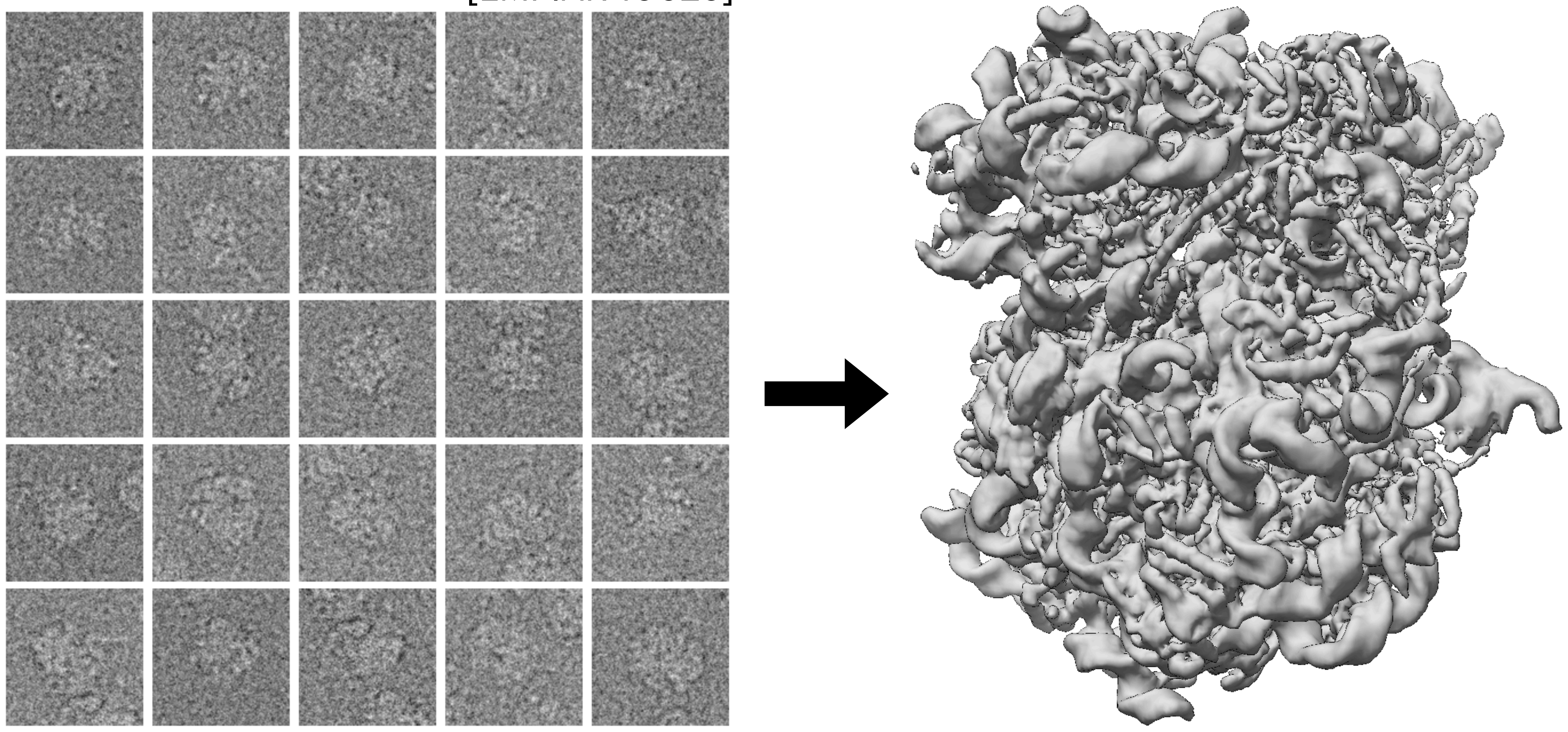
What makes 3D reconstruction challenging in cryo-EM?

Key principle: Biological samples are highly radiation-sensitive, but high-resolution structure can be recovered by combining information across thousands or millions of particles

3D Reconstruction

Sample Preparation
↓
Data Collection
↓
Particle Picking
↓
3D Reconstruction
↓
Model Building

The noisy 2D projection images are reconstructed into a 3D volume representing the protein's structure



[EMPIAR-10028]

2D Particle Images

3D Density Map

Challenges

- Unknown particle poses
- Low signal to noise ratio
- Microscope aberrations

Key principle: Biological samples are highly radiation-sensitive, but high-resolution structure can be recovered by combining information across thousands or millions of particles

3D Reconstruction

Sample Preparation

Data Collection

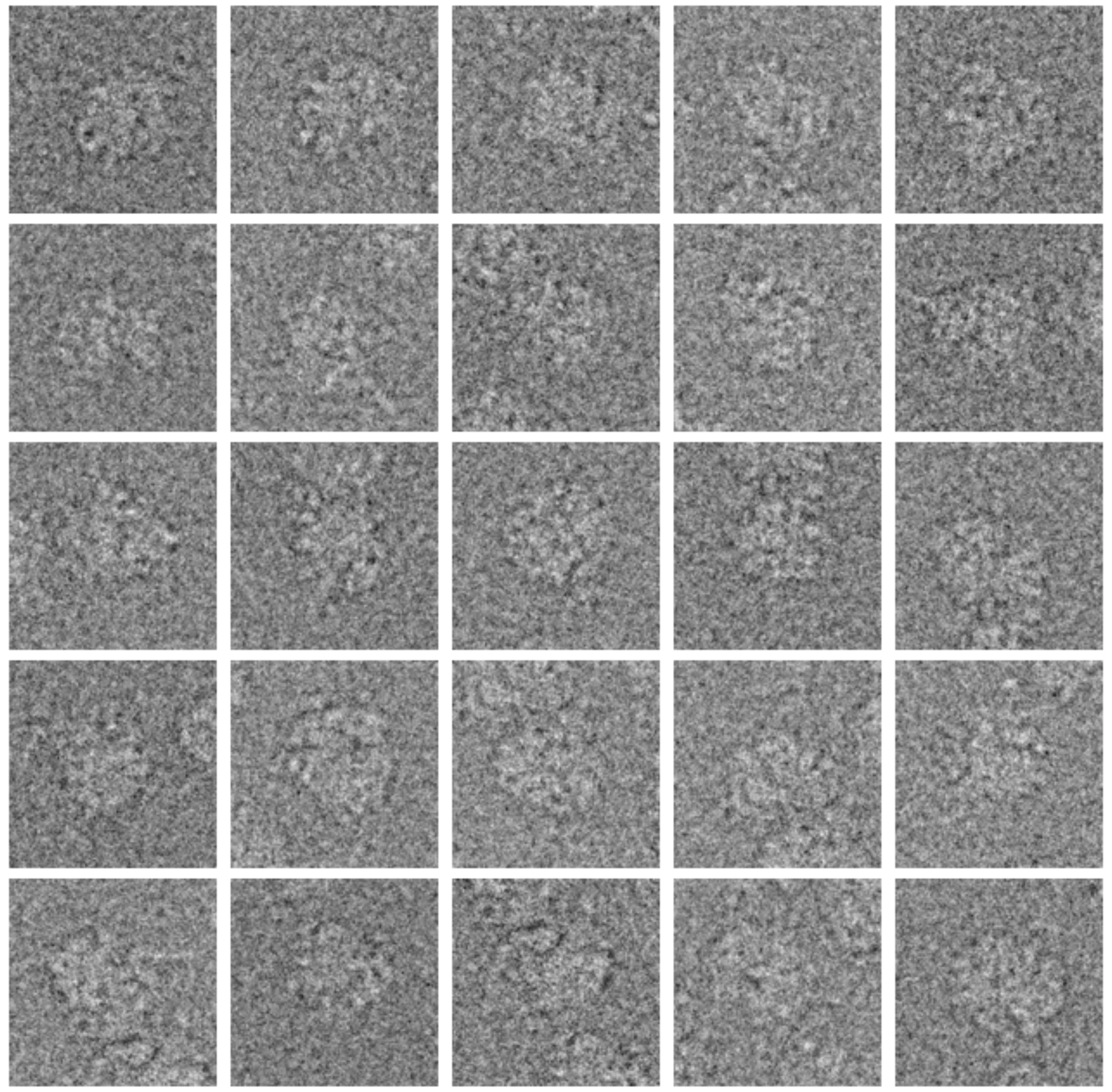
Particle Picking

3D Reconstruction

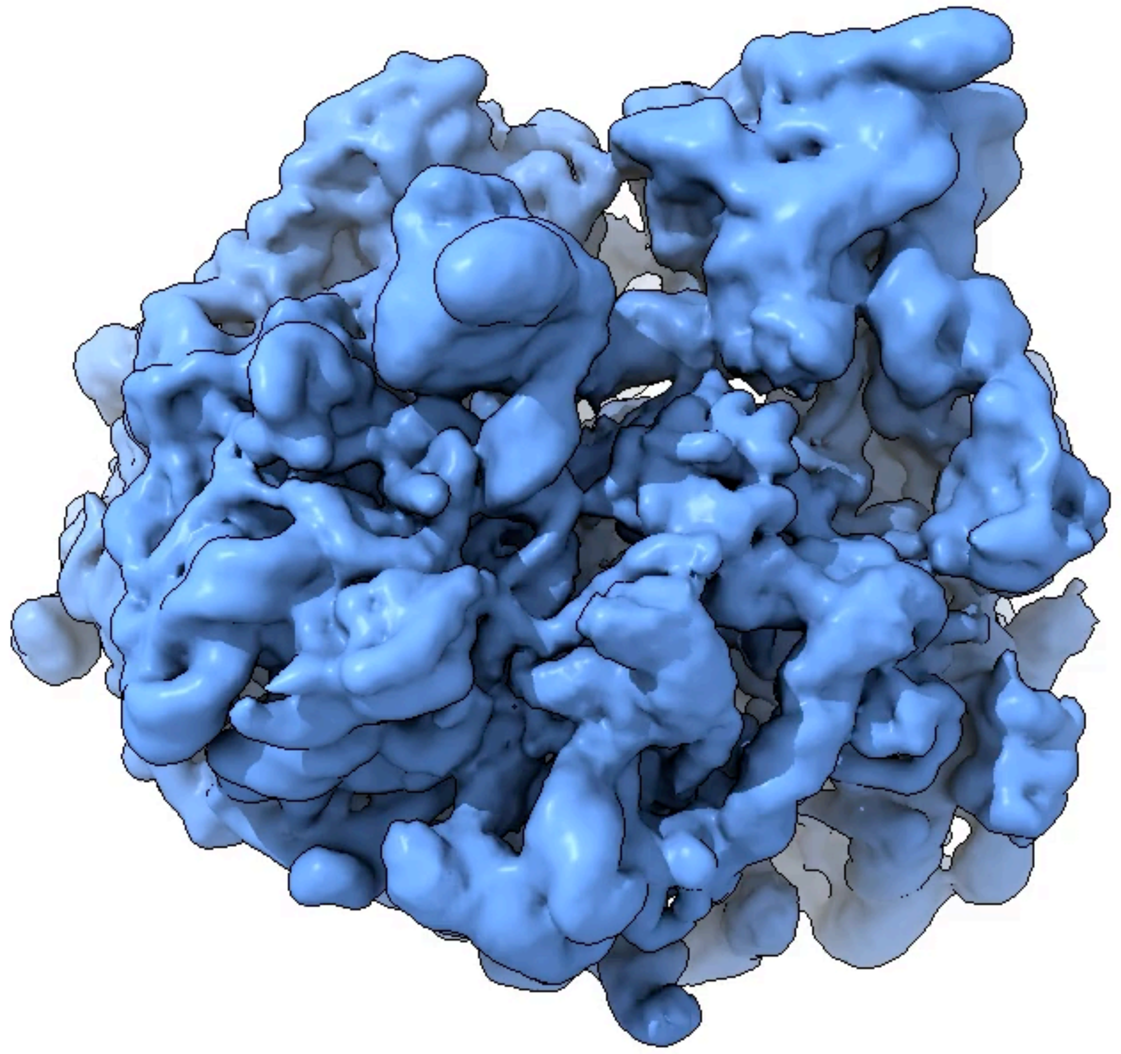
Model Building

As each image arises from a unique particle, several 3D volumes may be reconstructed as well, representing the protein's dynamic structure

[EMPIAR-10028]



2D Particle Images



3D Density Maps

Challenges

- Unknown particle poses
- Low signal to noise ratio
- Microscope aberrations
- ***Structural heterogeneity***

Model Building

Sample Preparation

↓

Data Collection

↓

Particle Picking

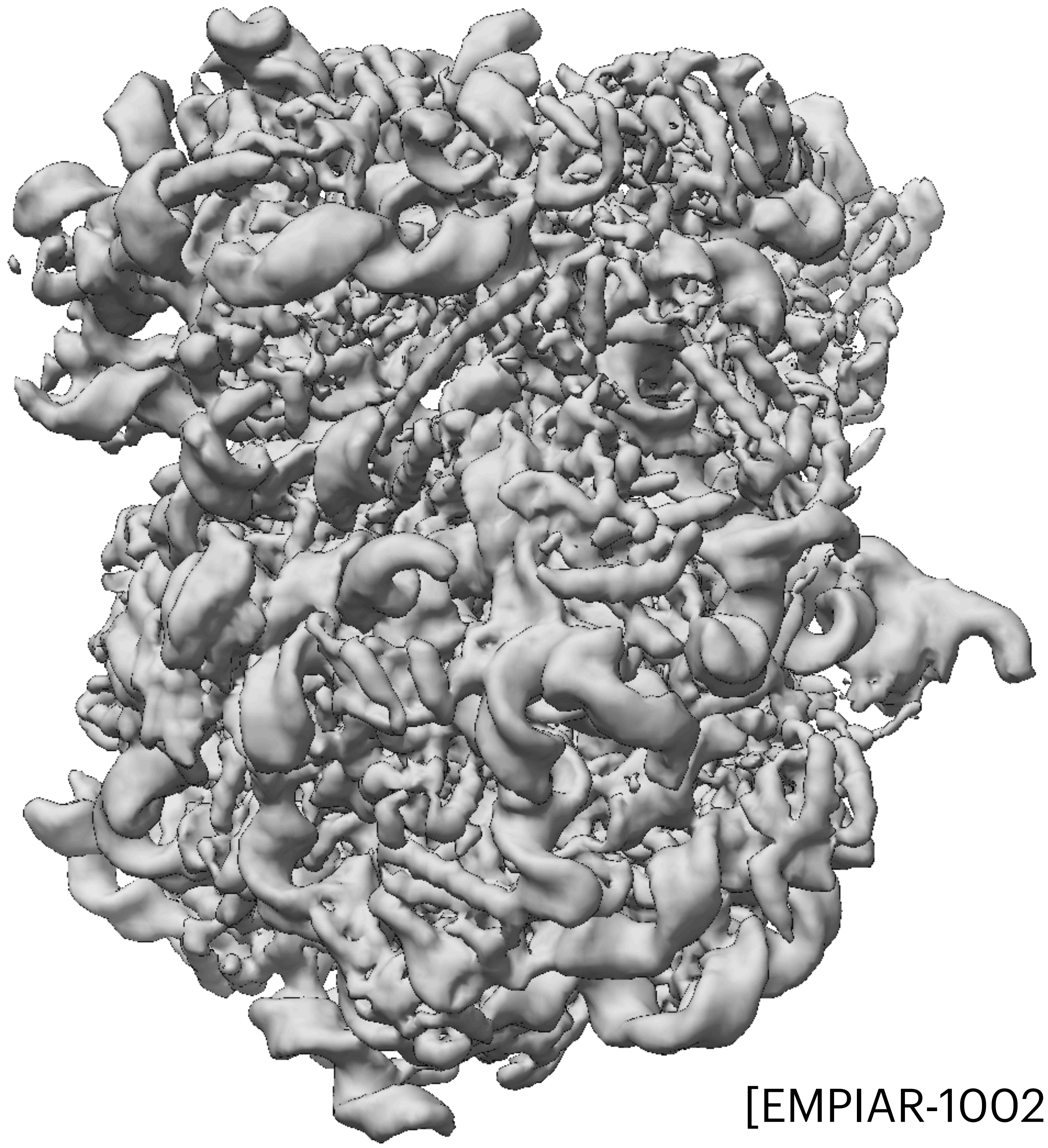
↓

3D Reconstruction

↓

Model Building

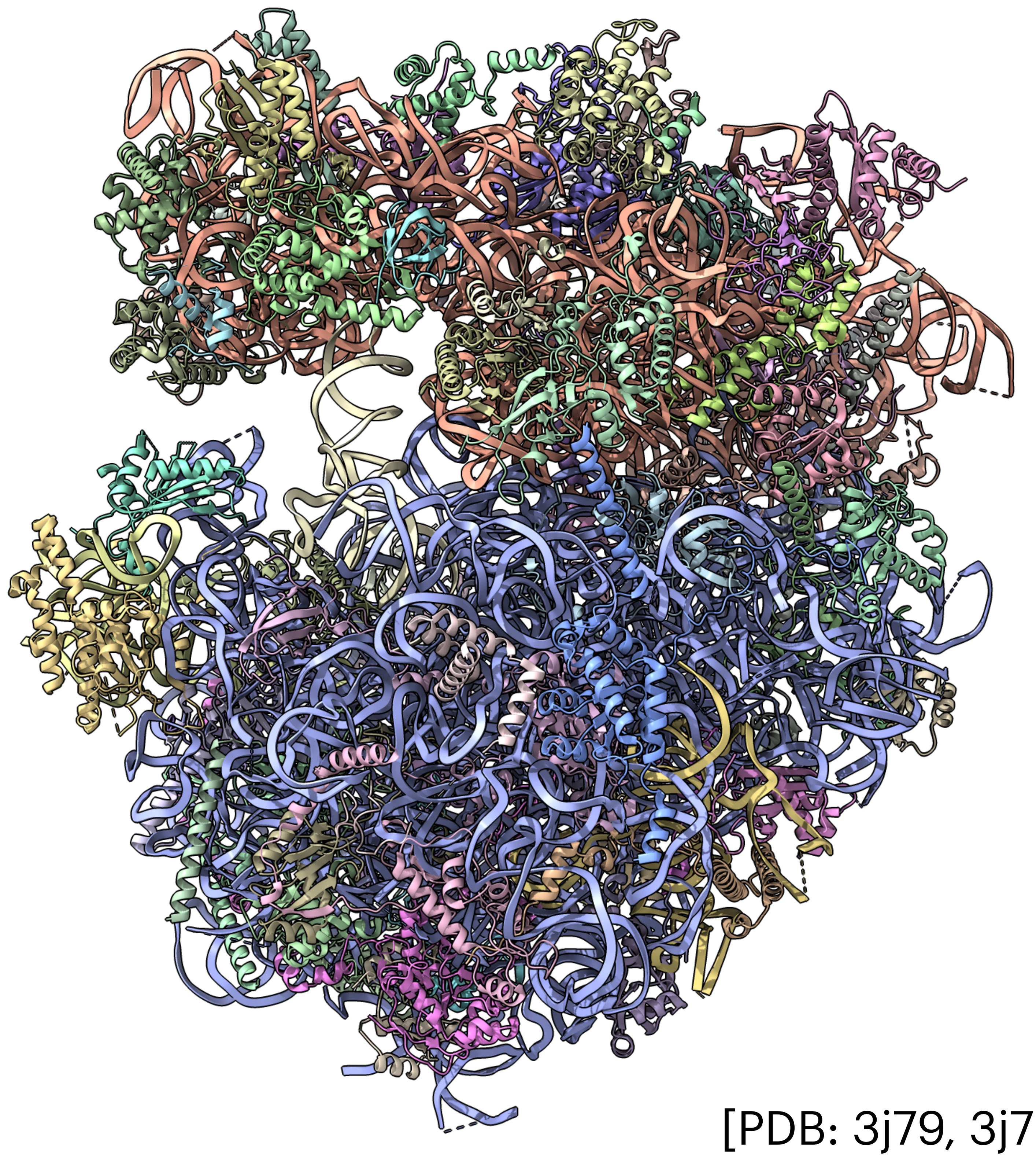
Amino acid identities and positions can be fit into high-resolution cryo-EM density maps to build atomic models



[EMPIAR-10028]

Cryo-EM Density Map

→



[PDB: 3j79, 3j7a]

Atomic Model

Roadmap

- Overview of the cryo-EM pipeline
- Classical cryo-EM reconstruction
  - **Homogeneous reconstruction**
  - Multiclass reconstruction
- Modern heterogeneous reconstruction
  - CryoDRGN

Cryo-EM Image Formation Model

Image Formation Model

$$X_i = C_i * \mathcal{P}_{\phi_i} V + \eta_i$$

2D Image

Contrast transfer function

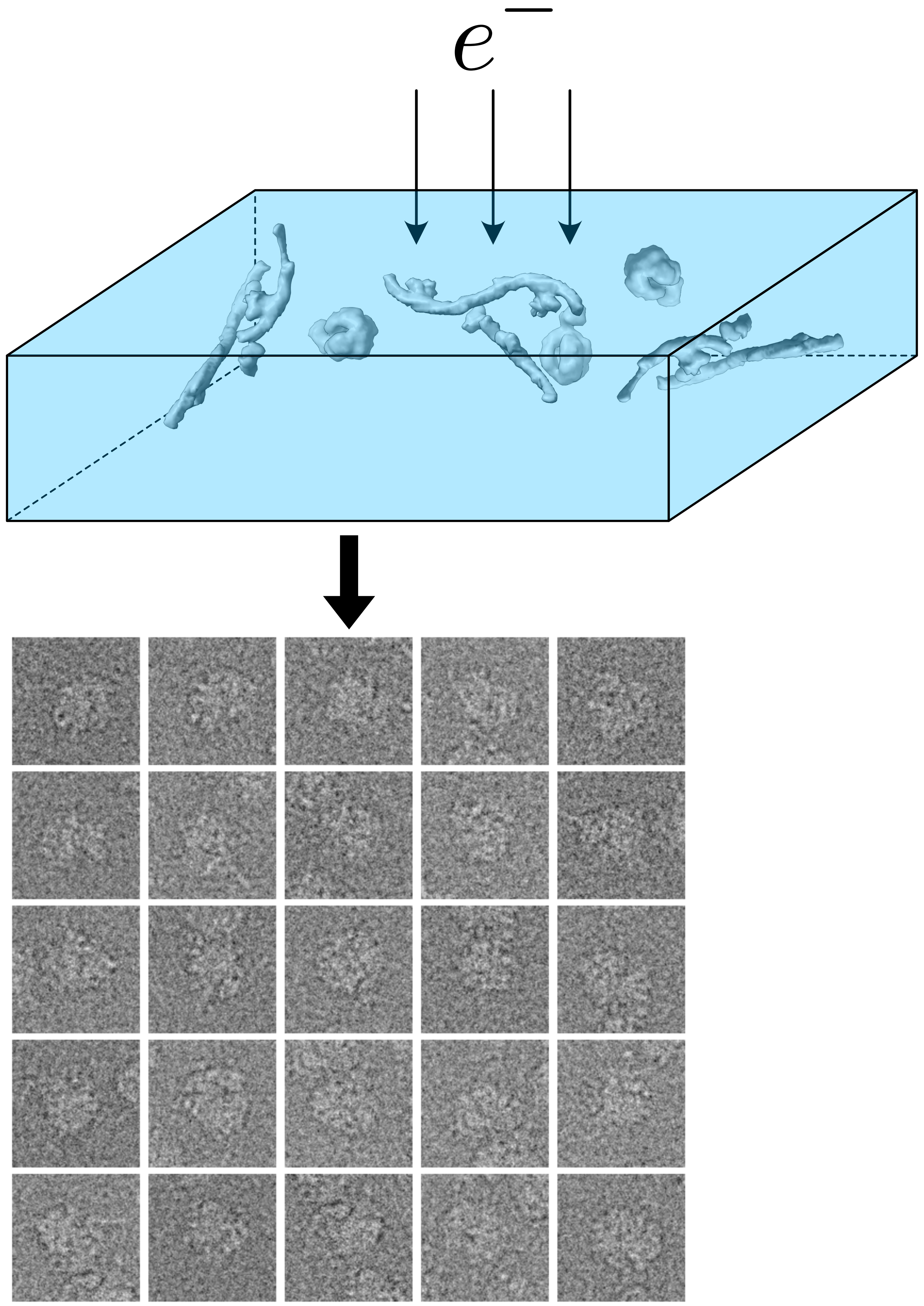
Projection about 5D pose

3D Cryo-EM density

Gaussian noise $\mathcal{N}(0, \sigma^2)$

5D pose:

- 3D rotation, due to random particle orientation within sample
- 2D translation, due to imperfect centering of particle within image



[EMPIAR-10028]

Cryo-EM Reconstruction Problem

Image Formation Model

$$X_i = C_i * \mathcal{P}_{\phi_i} V + \eta_i$$

2D Image

Contrast transfer function

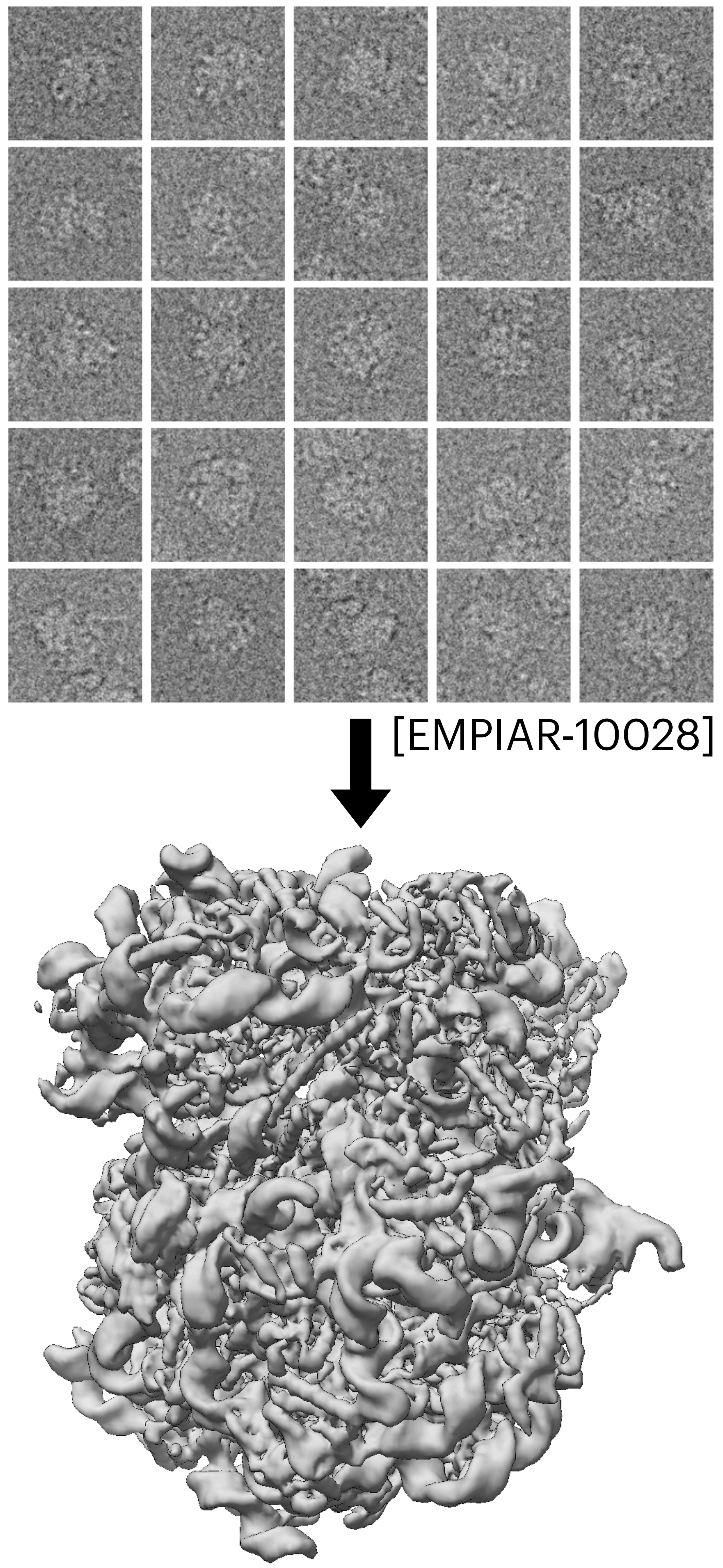
Projection about 5D pose

3D Cryo-EM density

Gaussian noise $\mathcal{N}(0, \sigma^2)$

Cryo-EM reconstruction problem:
Find optimal volume *V* given the data

Assume CTF values C_i are known but poses ϕ_i are unknown



Maximum Likelihood Objective

Recall the image formation model:

$$X_i = C_i * \mathcal{P}_{\phi_i} V + \eta_i \implies X_i \sim \mathcal{N}(C_i * \mathcal{P}_{\phi_i} V, \sigma^2)$$

The likelihood of a single image (given volume and pose) is:

$$P(X_i | V, \phi) = \frac{1}{(2\pi\sigma^2)^{D^2/2}} \exp\{-\frac{||X_i - C_i * \mathcal{P}_\phi V||^2}{2\sigma^2}\}$$

D = Image side length

We seek the volume *V* that maximizes the log-likelihood over all images, marginalizing over poses:

$$\hat{V}^{\mathrm{MLE}} = \underbrace{\arg\max_{V}}_{\text{Maximize over volumes}} \underbrace{\sum_{i=1}^{n}}_{\text{Sum over images}} \log \underbrace{\int_{\mathrm{SO}(3)\times\mathbb{R}^2}}_{\text{Integrate over pose space}} \underbrace{\exp\{-\frac{||X_i - C_i * \mathcal{P}_\phi V||^2}{2\sigma^2}\}}_{\text{Image likelihood given a pose}} d\phi$$

Expensive, non-convex

Singer and Sigworth 2020

Expectation-Maximization for MLE

Instead of maximizing the marginalized log-likelihood, maximize the expected log-likelihood instead:

$$\mathbb{E}_{\phi_1,\dots,\phi_n}[\sum_{i=1}^{n} \log P(X_i|V,\phi_i)]$$

(E-Step) Update the posterior distribution of poses, given the current volume:

$$\forall i \in [n], s \in S \qquad \boldsymbol{w_i(s) = P(\phi_i = s|V^{(t)}, X_i) = \frac{P(X_i|V^{(t)}, \phi_i = s)}{\sum_{s' \in S} P(X_i|V^{(t)}, \phi_i = s')}}$$

Discretized candidate poses

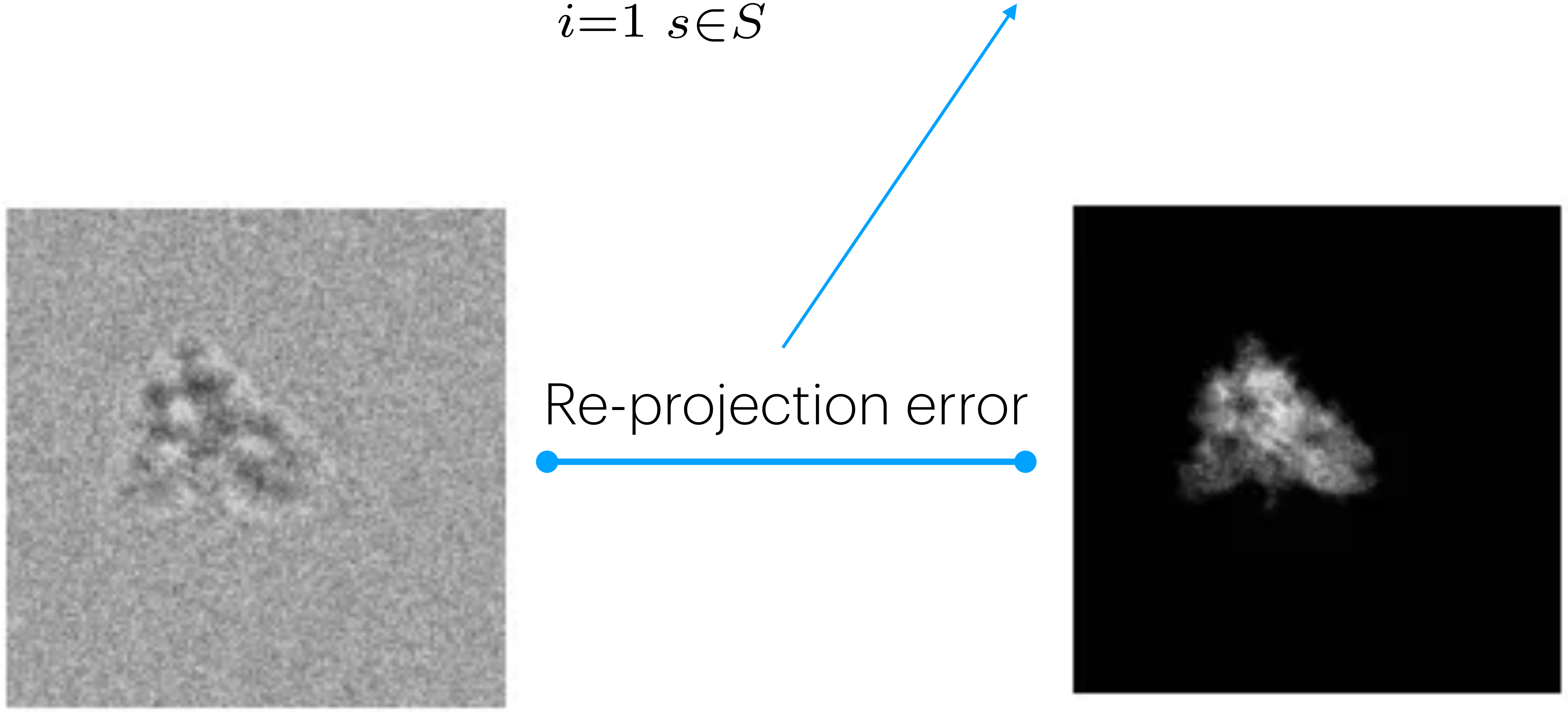
(M-Step) Update the volume, given the current pose posterior

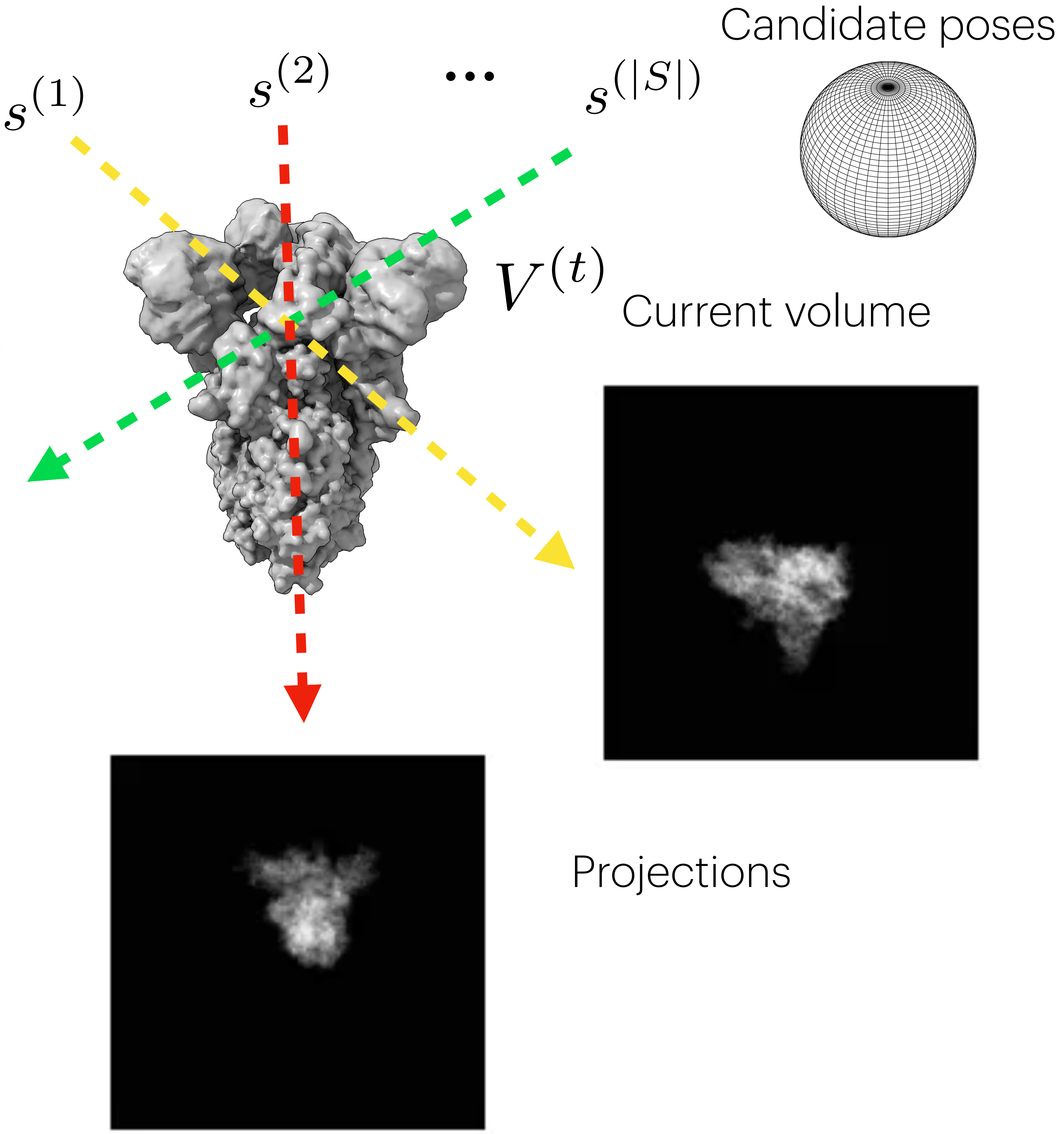
$$V^{(t+1)} = \arg\min_{V} \sum_{i=1}^{n} \sum_{s \in S} w_i(s) ||X_i - C_i * \mathcal{P}_s V||^2$$

$$V^{(t+1)} = A^{-1}(\sum_{i=1}^{n} \sum_{s \in S} w_i(s) \mathcal{P}_s^T C_i * X_i) \qquad A = \sum_{i=1}^{n} \sum_{s \in S} w_i(s) (\mathcal{P}s)^T C_i * C_i * \mathcal{P}s$$

E-Step: Weighted Projection Matching

$$V^{(t+1)} = \arg\min_{V} \sum_{i=1}^{n} \sum_{s \in S} w_i(s) \underline{||X_i - C_i * \mathcal{P}_s V||^2}$$



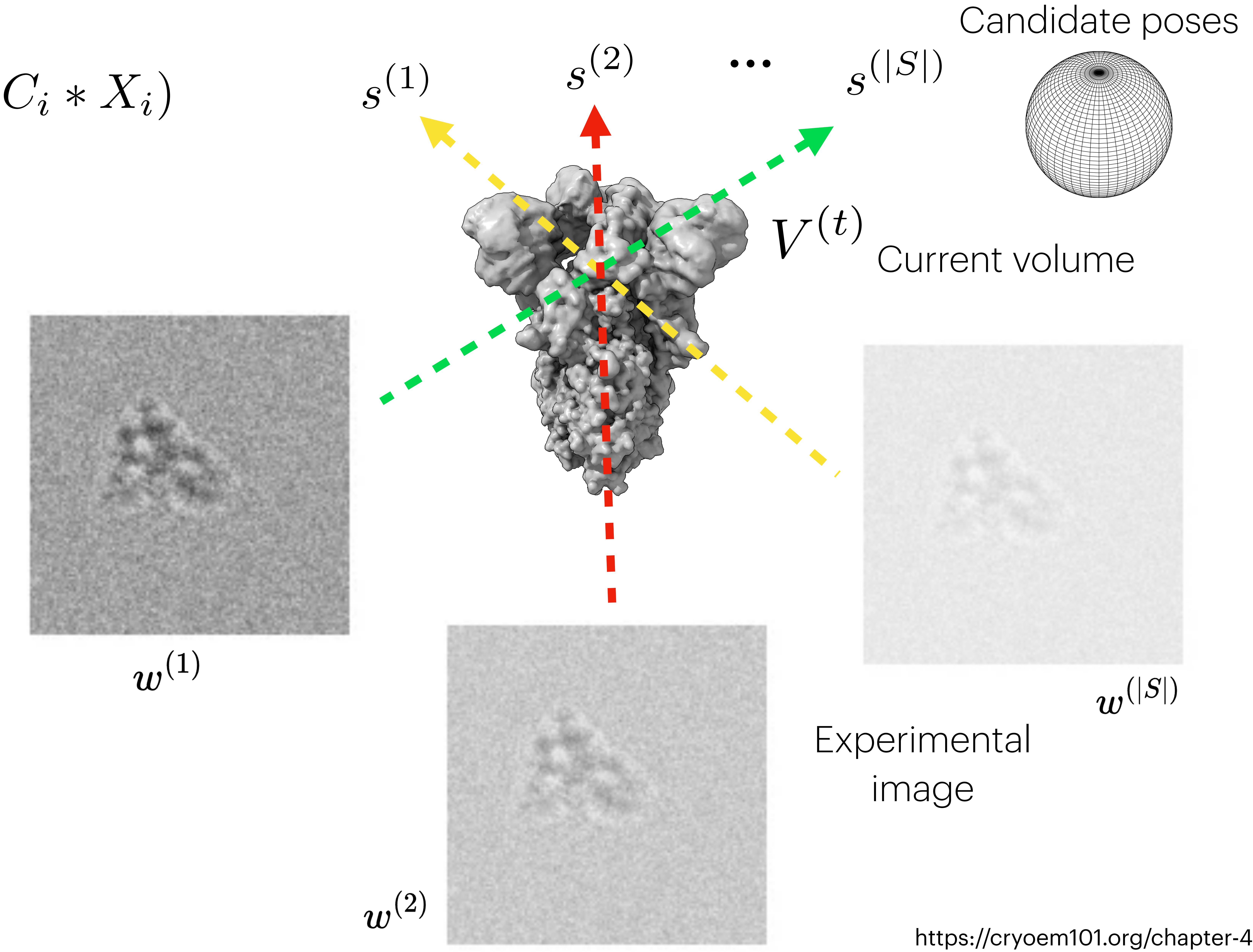


- If image is well-aligned, weights will be peaked
- If image pose is uncertain, weights will be more uniform, and contribute little to the reconstruction

M-Step: Weighted Backprojection

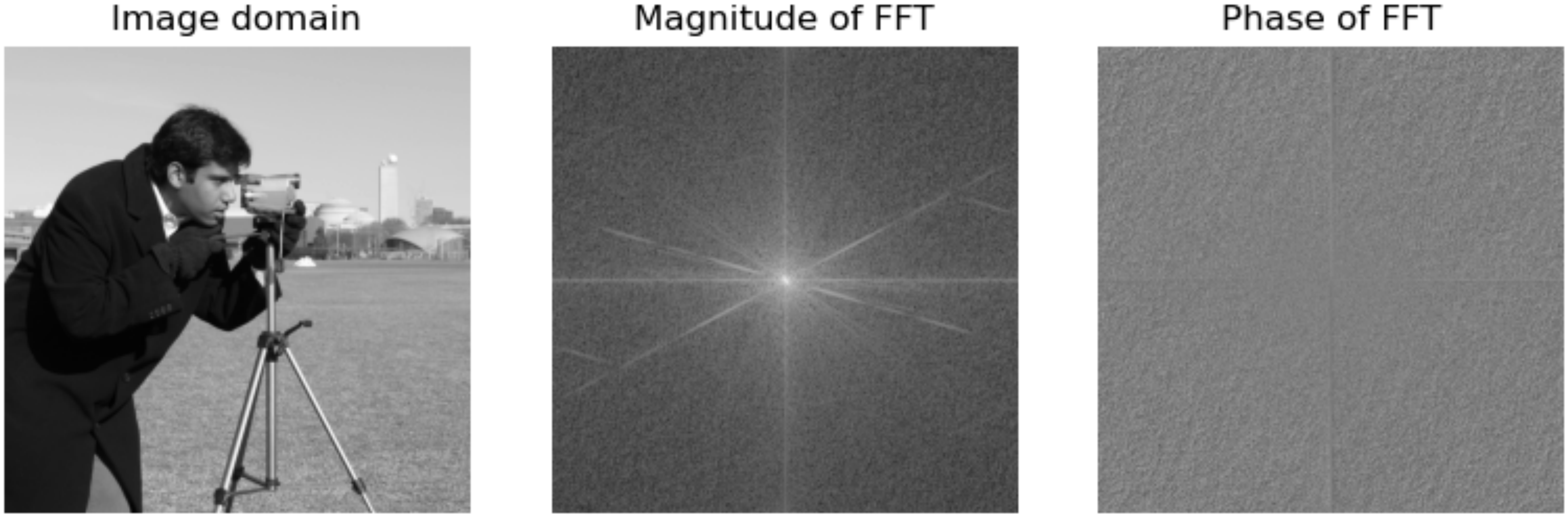
$$V^{(t+1)} = A^{-1}\Big(\sum_{i=1}^{n}\sum_{s\in S} w_i(s)\mathcal{P}_s^T C_i * X_i\Big)$$

- "Inverse" projection of the original image at all candidate poses, using pose weights computed in E-step



https://cryoem101.org/chapter-4/

Real vs Fourier Space



Real space

Intensity as a function of
spatial coordinate

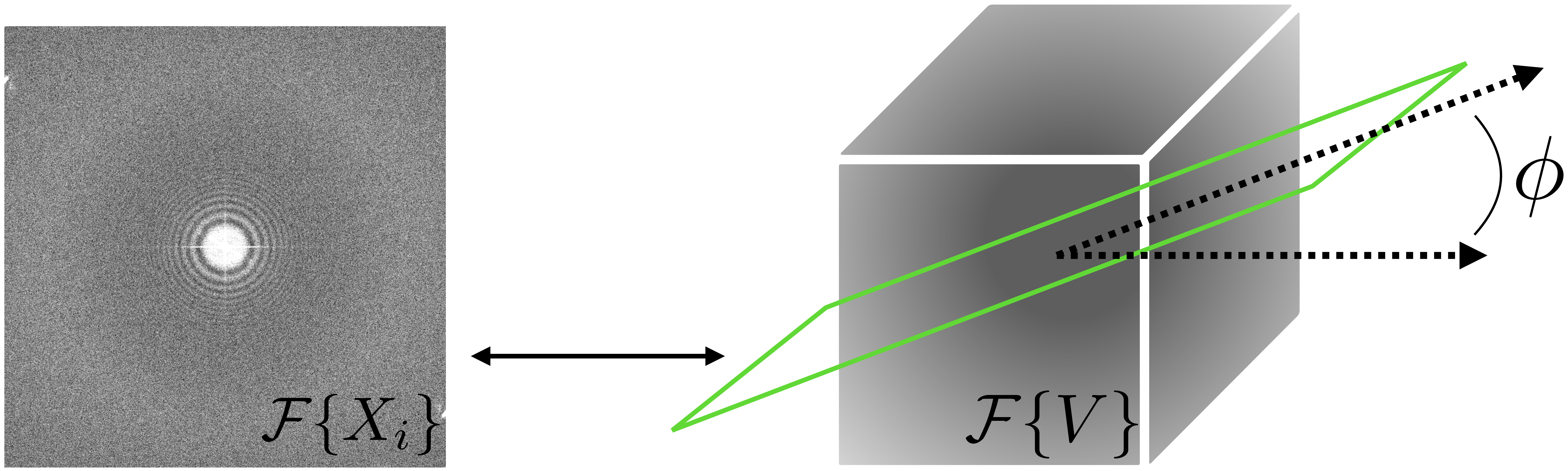
Fourier Space

Amplitude and phase as a
function of frequency

https://peterbbryan.medium.com/understanding-2d-fourier-space-59808b644a13

Fourier Slice Theorem

- E-step: Extract Fourier slices out of the current Fourier volume and compute pose weights
- M-step: Insert Fourier transform of input image as slices of Fourier volume, according to pose weight



<u>Fourier Slice Theorem:</u> 2D FT is centrally oriented slice of 3D FT

Real space: $X_i = C_i * \mathcal{P}_{\phi_i} V + \eta_i$

Fourier space: $\mathcal{F}\{X_i\} = \mathcal{F}\{C_i\} \Theta_{\phi_i} \mathcal{F}\{V\}$

What are the tradeoffs in real vs Fourier space reconstruction?

https://cryoem101.org/chapter-4/

Key Challenges in Optimization

1. Convergence to local optima due to poor volume initialization

2. Due to low SNR imaging, easily overfits to noise

Initialization of E-M Algorithm

- The E-M algorithm is guaranteed to improve the likelihood every iteration
- This suggests that given a random initialization of *V*, the algorithm will converge to a local optima
- To help find the global optimum, CryoSPARC (Punjani et al. 2017) introduced gradient descent on the objective to achieve a good starting volume

$$\hat{V}^{\mathrm{MLE}} = \arg\max_{V} \ell(V) = \arg\max_{V} \sum_{i=1}^{n} \log \int_{\mathrm{SO}(3)\times\mathbb{R}^2} \exp\{-\frac{||X_i - C_i * \mathcal{P}_\phi V||^2}{2\sigma^2}\} d\phi$$

$$V^{(t+1)} = V^{(t)} + \alpha_t \sum_{i \in B} \nabla_V \ell(V; X_i)$$

α_t Learning rate

B Batch of images

Singer and Sigworth 2020

Preventing Overfitting to Noise

RELION (Scheres 2012) introduced a Bayesian framework for reconstruction:

$$\hat{V}^{\mathrm{MAP}} = \arg\max_{V} [\underbrace{\log P(X_1, ..., X_n | V)}_{\text{MLE Objective}} + \underbrace{\log P(V)}_{\substack{\text{Volume Prior} \\ \text{(Regularization)}}}]$$

A common prior is to assume each Fourier component is from a zero-mean Gaussian:

$$P(V) = \prod_{d=1}^{D^3} \frac{1}{2\pi\tau_d^2} \exp\{-\frac{|V_d|^2}{2\tau_d^2}\}$$

The frequency-dependent variance of the prior is sometimes optimized as well:

$$(\tau_d^2)^{(t)} = \frac{1}{2} |V_d^{(t)}|^2$$

Singer and Sigworth 2020

Roadmap

- Overview of the cryo-EM pipeline
- Classical cryo-EM reconstruction
    - Homogeneous reconstruction
    - **Multiclass reconstruction**
- Modern heterogeneous reconstruction
    - CryoDRGN

Multiclass Image Formation Model

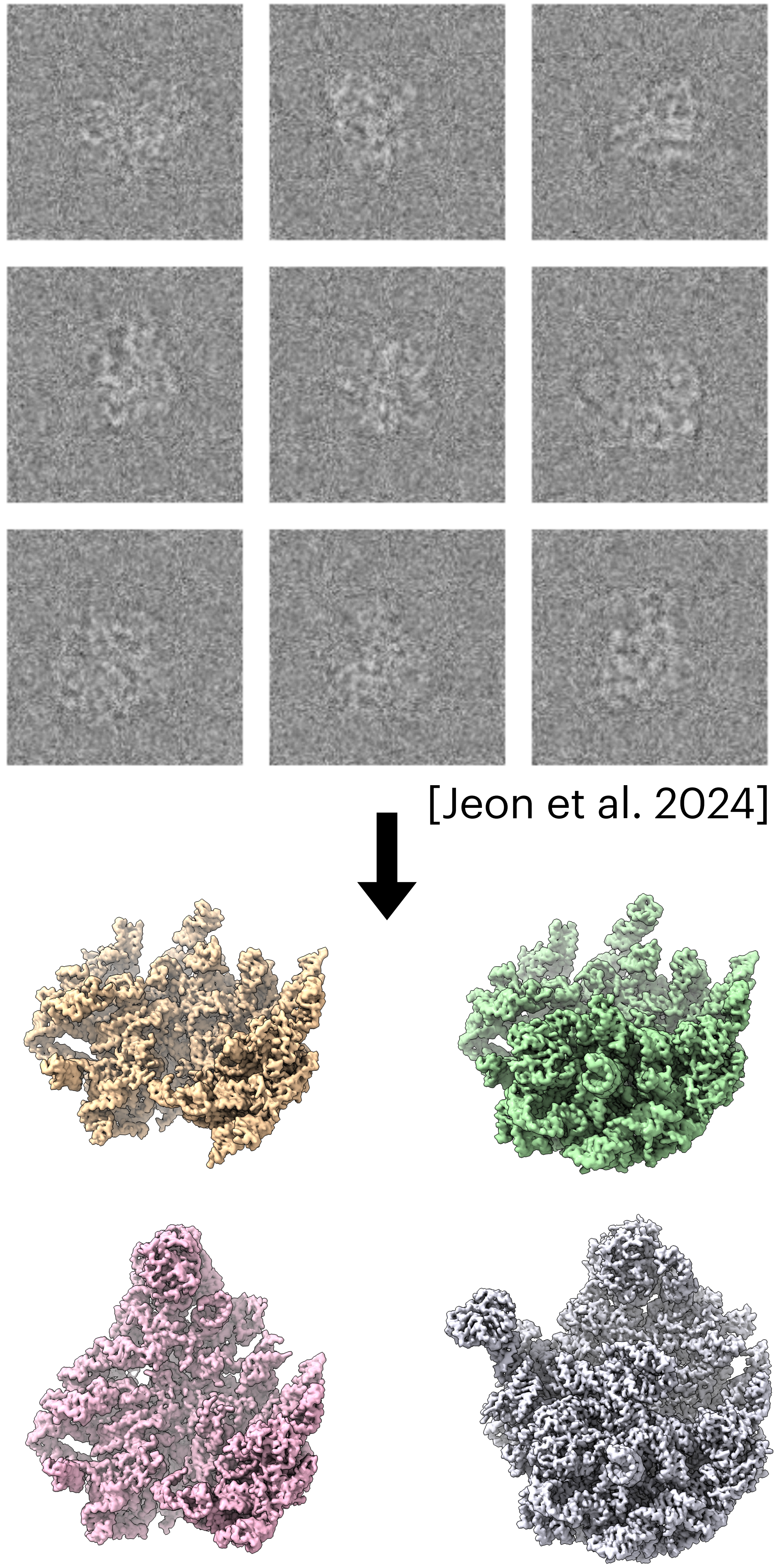
Image Formation Model

$$X_i = C_i * \mathcal{P}_{\phi_i} V_{a_i} + \eta_i$$

$$a_i \in \{1, ..., K\}$$

Particle Class

Multiclass reconstruction problem:
Find optimal volumes $\{V_1, ..., V_k\}$ given the data


[Jeon et al. 2024]

Multiclass Reconstruction

The MLE objective marginalizes over class assignments:

$$\hat{V}_1^{\mathrm{MLE}}, ..., \hat{V}_K^{\mathrm{MLE}} = \arg \max_{V_1, ..., V_K} \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(X_i | \phi_i, V_k)$$

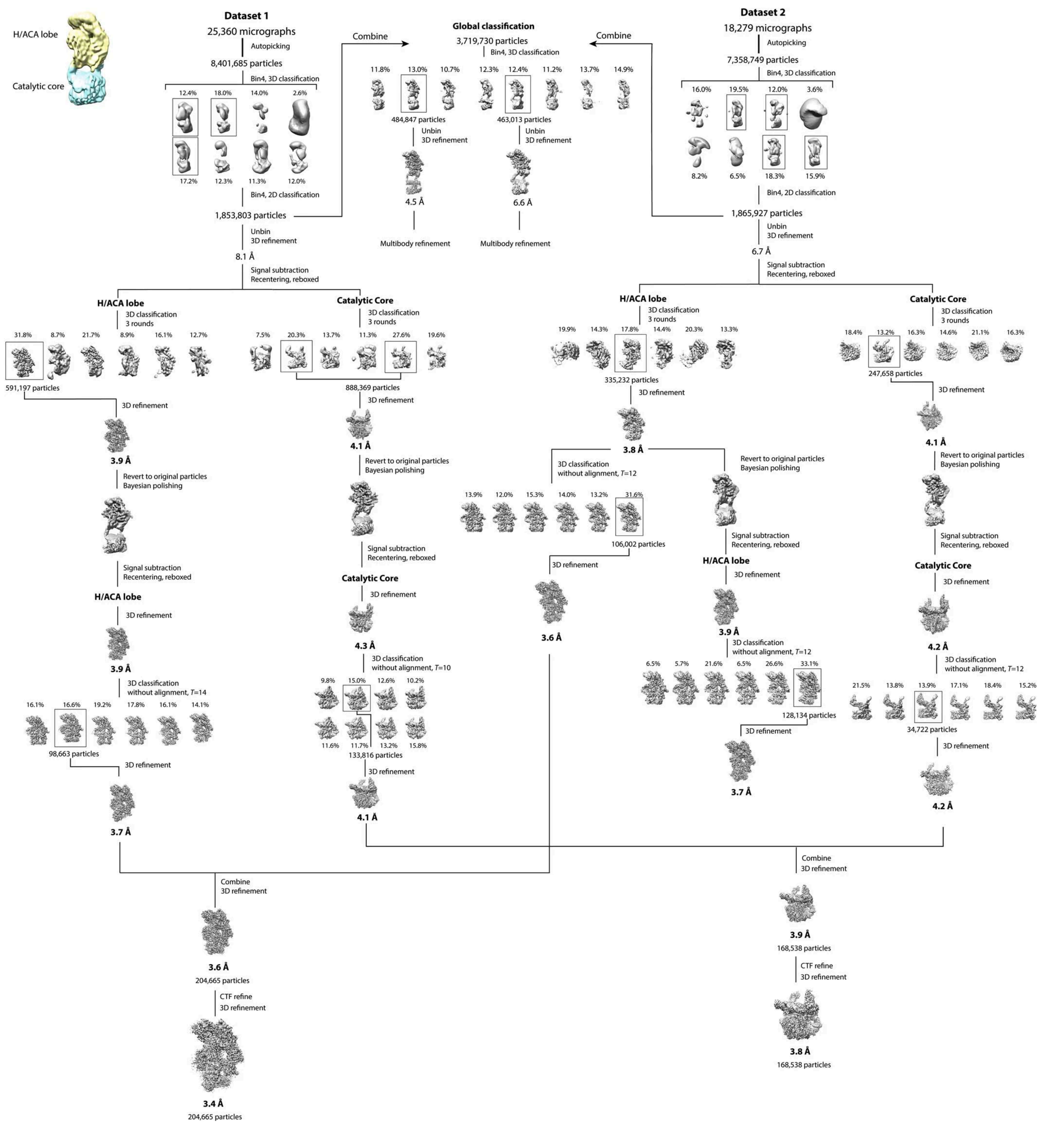
We can again use the E-M algorithm:

E-Step: $$w_i(k) = P(a_i = k | V_1^{(t)}, \ldots, V_K^{(t)}) = \frac{P(X_i | V_k^{(t)})}{\sum_{k' \in [K]} P(X_i | V_{k'}^{(t)})}$$

M-Step: $$V_k^{(t+1)} = \arg\min_V \sum_{i=1}^{n} w_i(k) || X_i - C_i * \mathcal{P}_{\phi_i} V ||^2$$

Practical Challenges in 3D Classification

- The discrete mixture model is not well-suited for continuous forms of heterogeneity
- Higher the number of classes *K*, the fewer the number of particles contributing to each volume
- Optimization especially sensitive to initial volumes
- *Ad hoc* processing pipelines involving iterative, hierarchical procedures with sweeps over *K*



Ghanim et al. Nature 2021

Roadmap

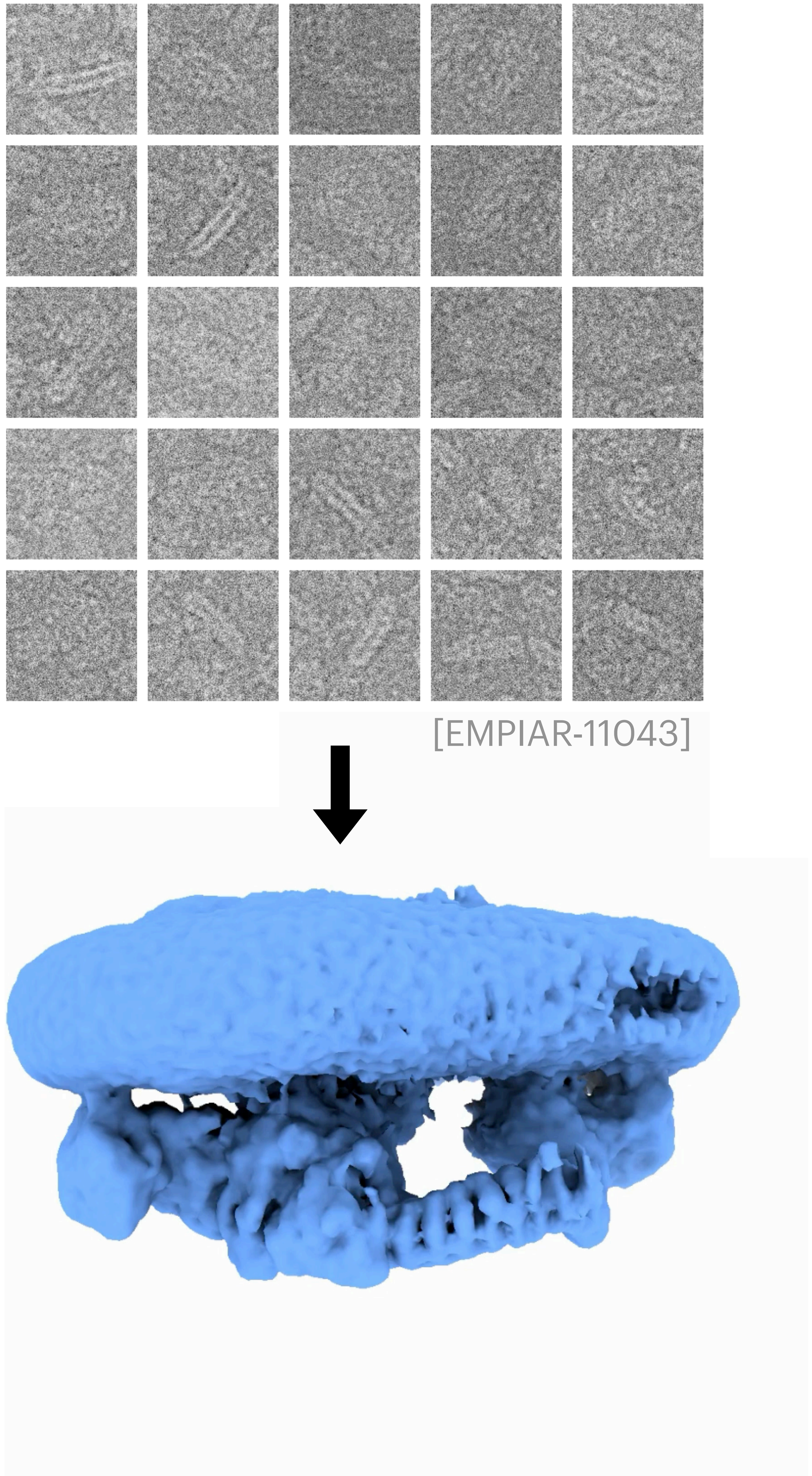
- Overview of the cryo-EM pipeline
- Classical cryo-EM reconstruction
  - Homogeneous reconstruction
  - Multiclass reconstruction
- Modern heterogeneous reconstruction
  - **CryoDRGN**

Heterogeneous Reconstruction

Image Formation Model

$$X_i = C_i * P_{\phi_i} \underbrace{V_\theta(z_i)}_{\text{Distribution of volumes}} + \eta_i$$

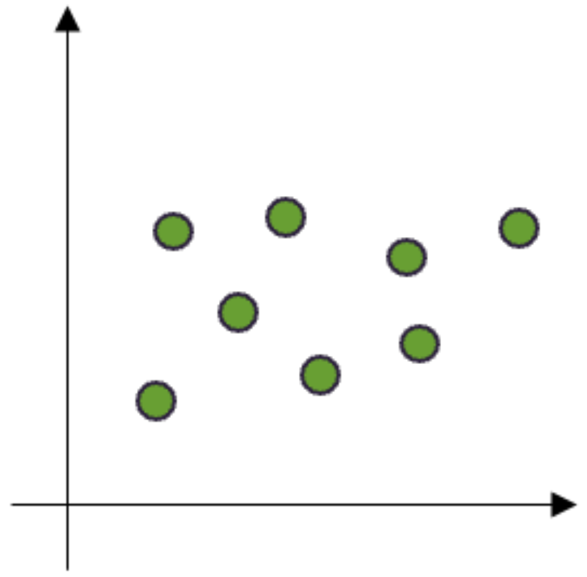
Heterogeneous reconstruction problem:
Find optimal volume distribution V_θ
given the data


[EMPIAR-11043]

Latent Variable Models for Heterogeneous Reconstruction

$$\{V_\theta(z), z \in Z\}$$

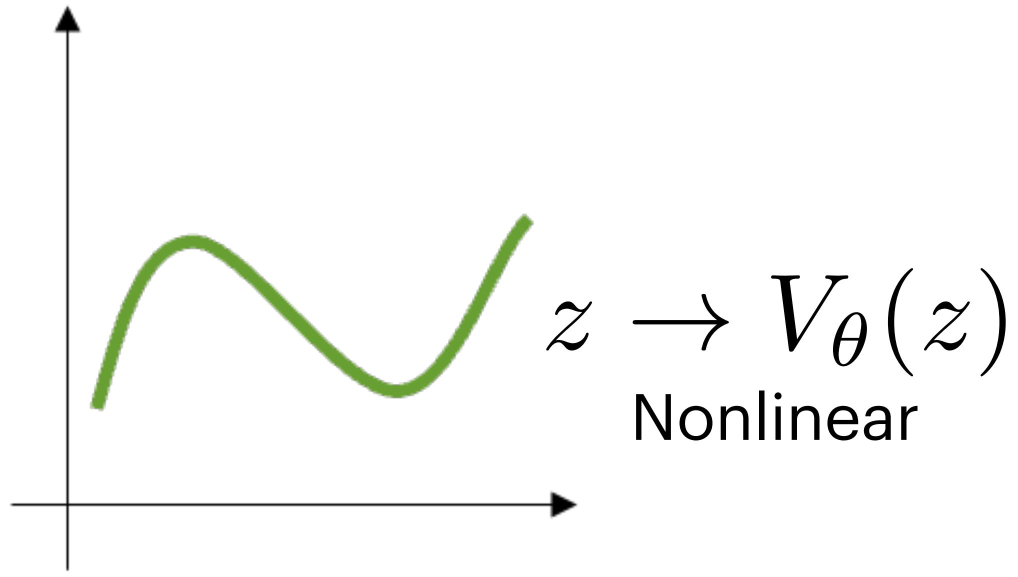
$$Z = \{1,\ldots,K\}$$



Discrete Set of Voxel Grids

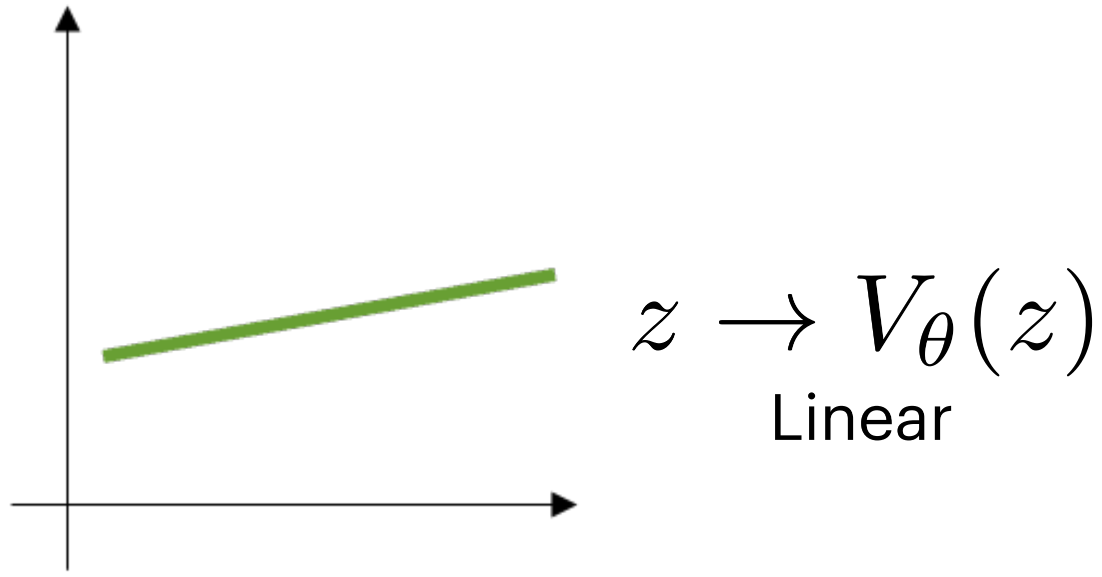
RELION [1], CryoSPARC [2]

$$Z = \mathbb{R}^D$$



Non Linear Parameterizations [8, 9]

$$Z = \mathbb{R}^D$$



Linear Combination of Voxel Grids

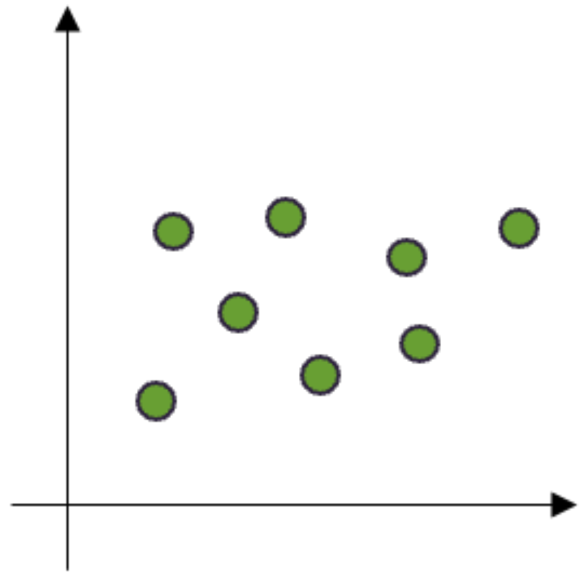
3DVA, RECOVAR,
Multibody

Slide credit: Axel Levy

Latent Variable Models for Heterogeneous Reconstruction

$$\{V_\theta(z), z \in Z\}$$

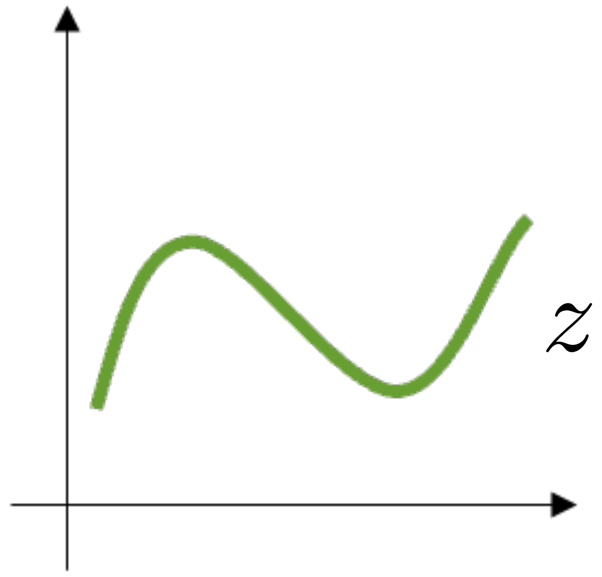
$$Z = \{1,\ldots,K\}$$



Discrete Set of Voxel Grids

RELION [1], CryoSPARC [2]

$$Z = \mathbb{R}^D$$



$z \to V_\theta(z)$
Nonlinear

Non Linear Parameterizations [8, 9]

Neural Representations

CryoDRGN
(Zhong, Nat. Met., 2021)

Explicit: Multi-CryoGAN,
OPUS-DSD

Deformable GMMs

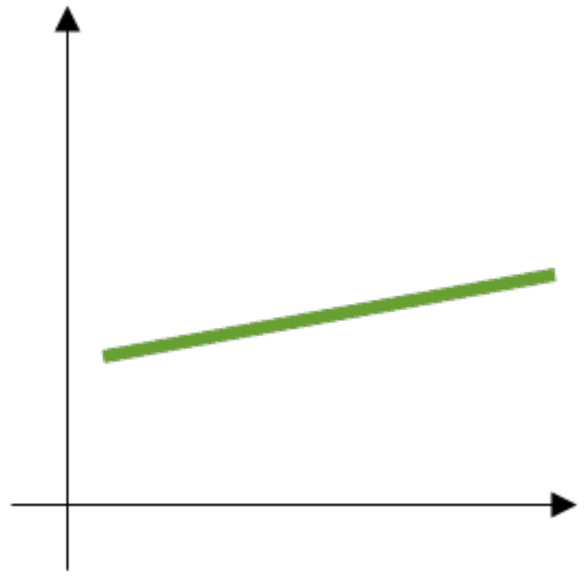
E2GMM
(Chen, Nat. Met., 2021)

HEMNMA, CryoSTAR
Dynamight, Cryofold, ...

Flow Field

3DFLEX
(Punjani, Nat. Met., 2023)

$$Z = \mathbb{R}^D$$



$z \to V_\theta(z)$
Linear

Linear Combination of Voxel Grids

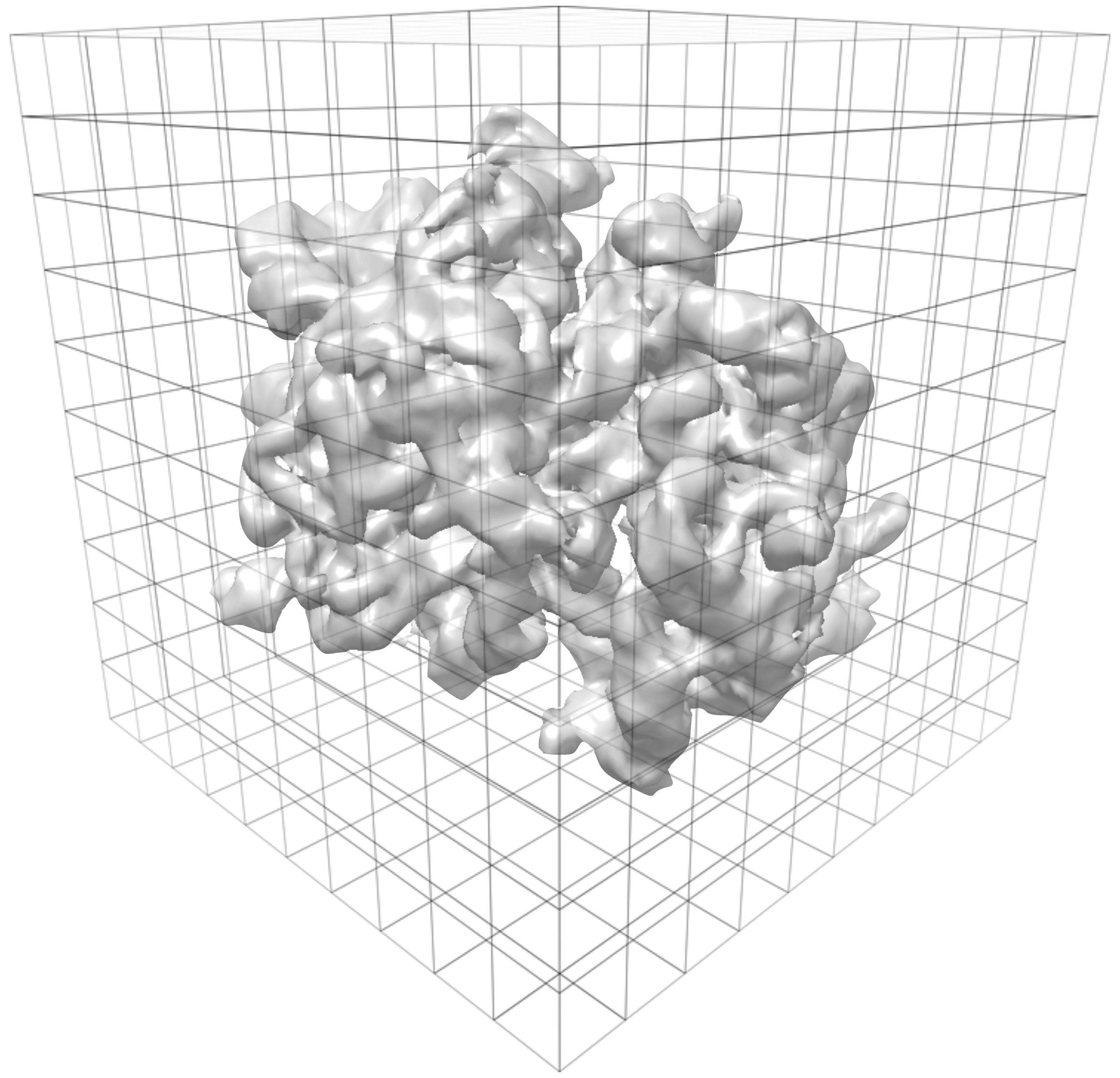
3DVA, RECOVAR,
Multibody

Slide credit: Axel Levy

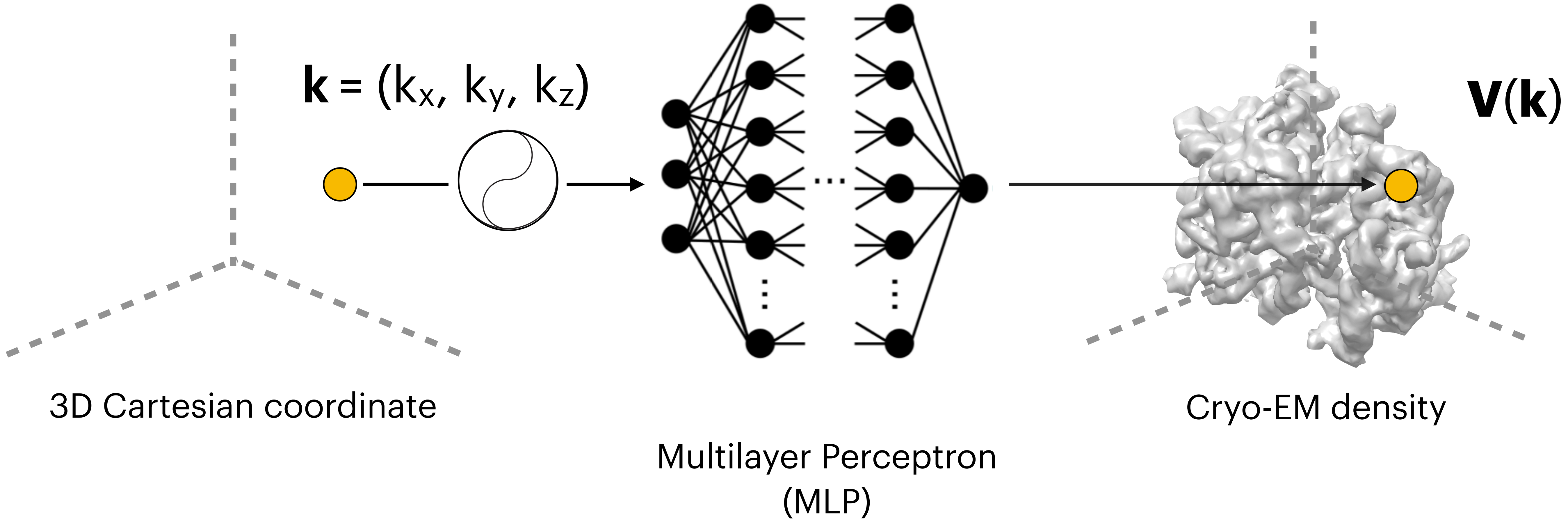
Coordinate-based neural networks for 3D volumes

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



$\mathbf{k} = (k_x, k_y, k_z)$

$\mathbf{V}(\mathbf{k})$

3D Cartesian coordinate

Multilayer Perceptron (MLP)

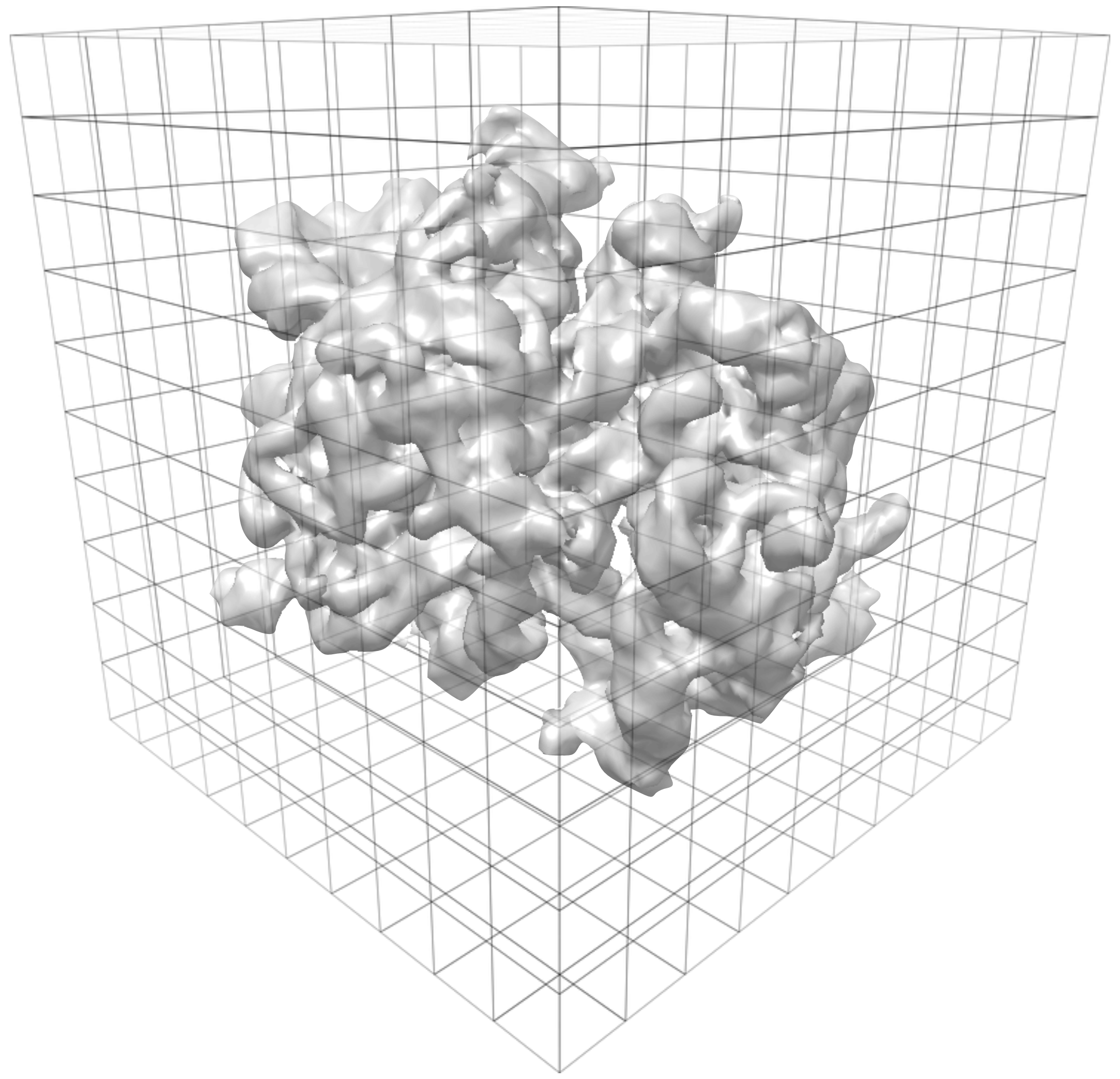
Cryo-EM density

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

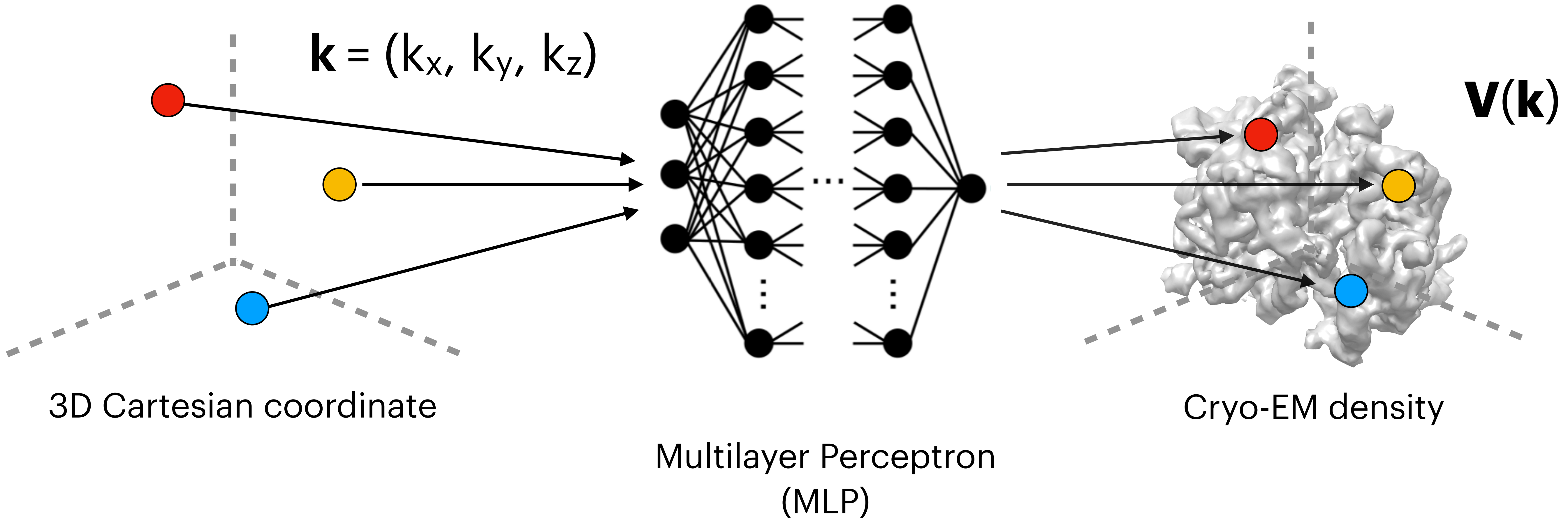
Coordinate-based neural networks for 3D volumes

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

Traditional algorithms



cryoDRGN

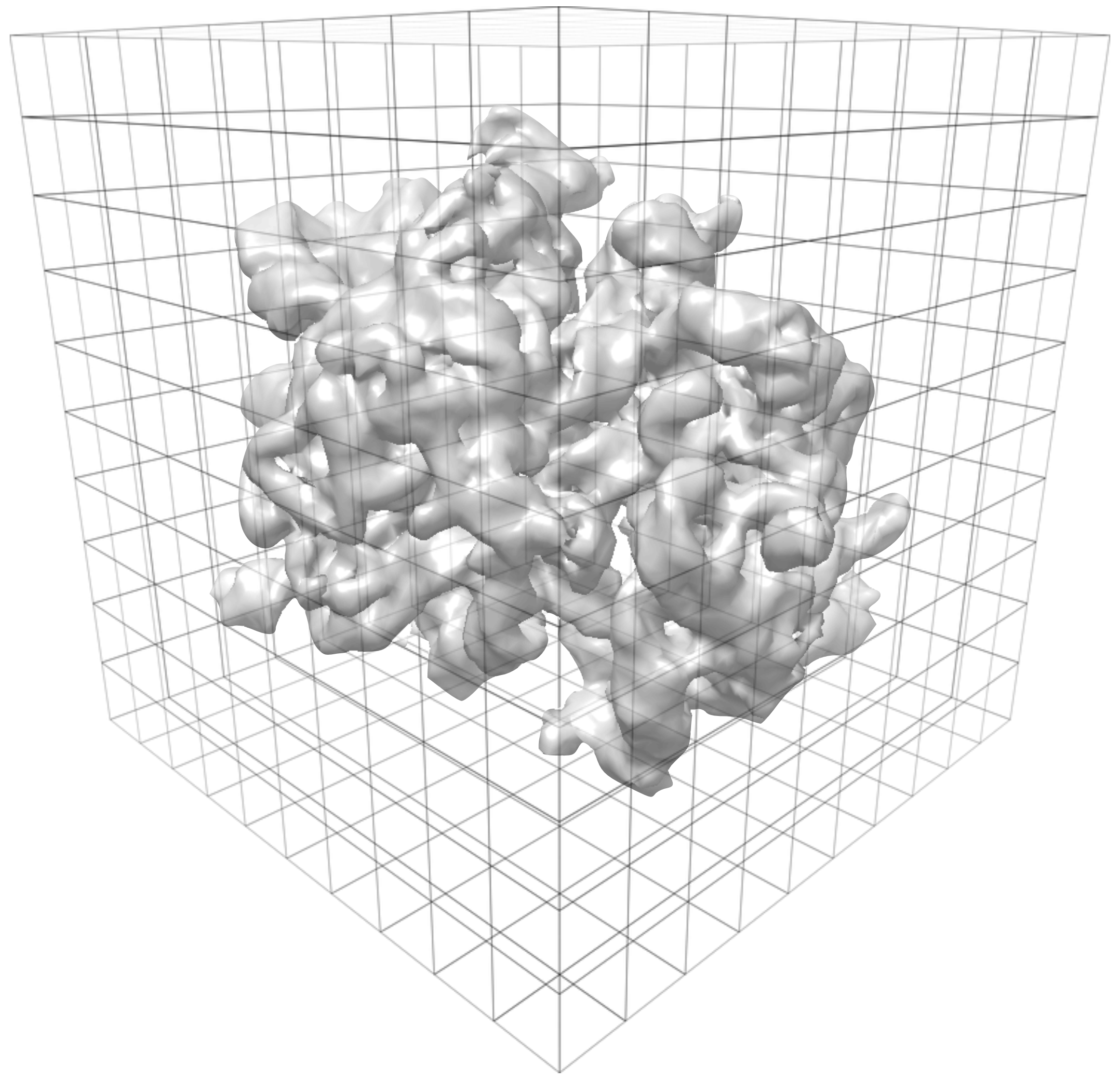


To construct a volume, evaluate the function at different spatial frequencies

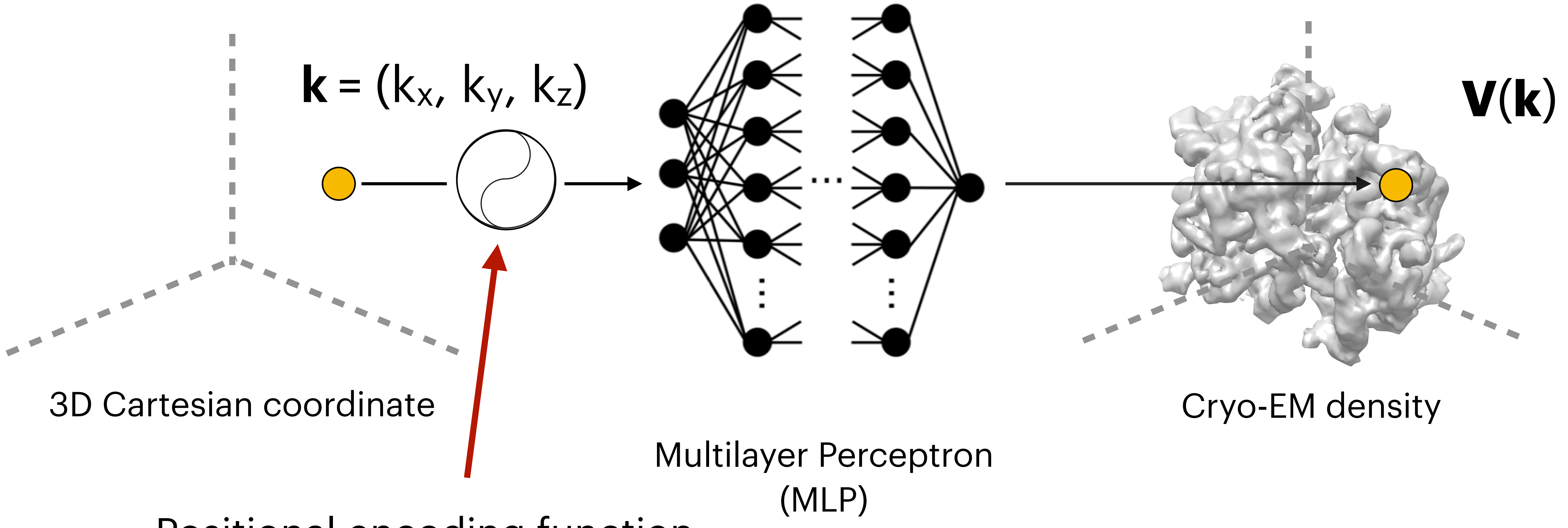
Coordinate-based neural networks for 3D volumes

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

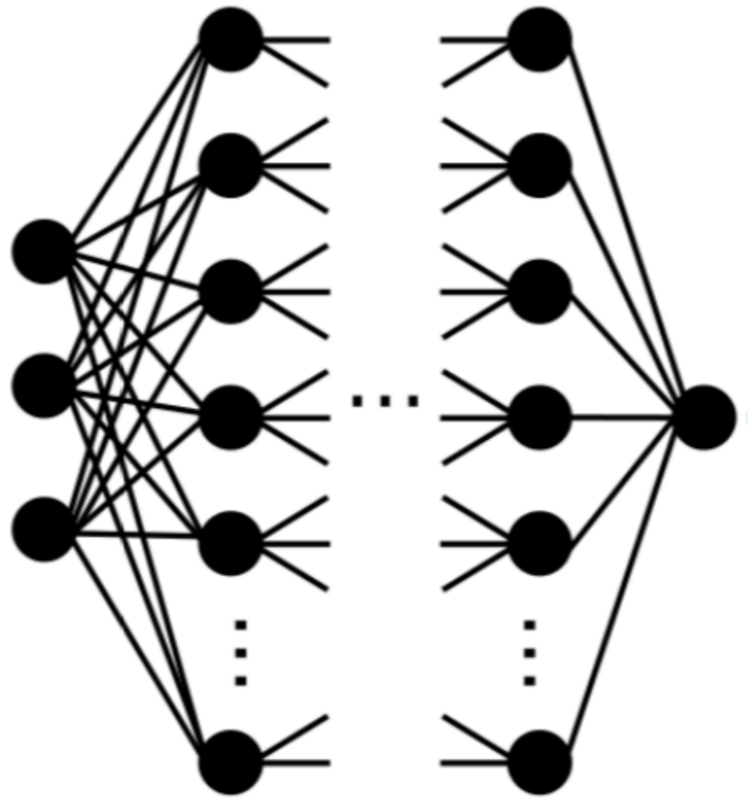
$$pe^{(2i)}(k_j) = sin(k_j D\pi(2/D)^{2i/D}),\ i = 1, ..., D/2; k_j \in k$$

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

Sinusoidal featurization of inputs

A toy example: Use an MLP to learn an image

$$(x,y) \longrightarrow \left(\begin{matrix} \sin(\gamma_0 x) \\ \cos(\gamma_0, x) \\ \cdots \end{matrix} \, , \, \begin{matrix} \sin(\gamma_0 y) \\ \cos(\gamma_0 y) \\ \cdots \end{matrix} \right) \longrightarrow$$



$$\longrightarrow (r,g,b)$$

Ground truth

No sinusoidal encoding

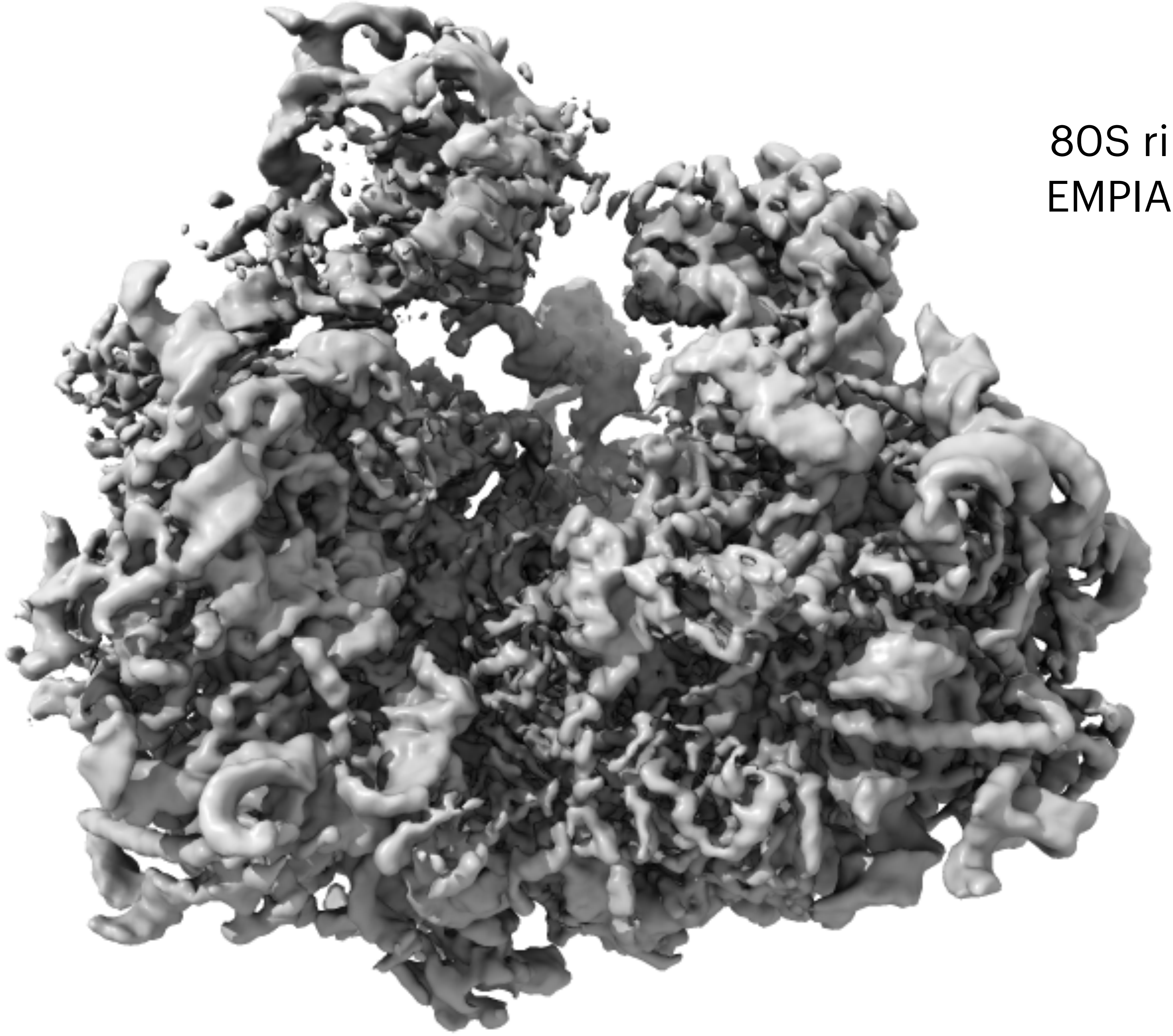
With sinusoidal encoding

Also see: Tancik et al, NeurIPS 2020

Neural fields can learn high resolution maps

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

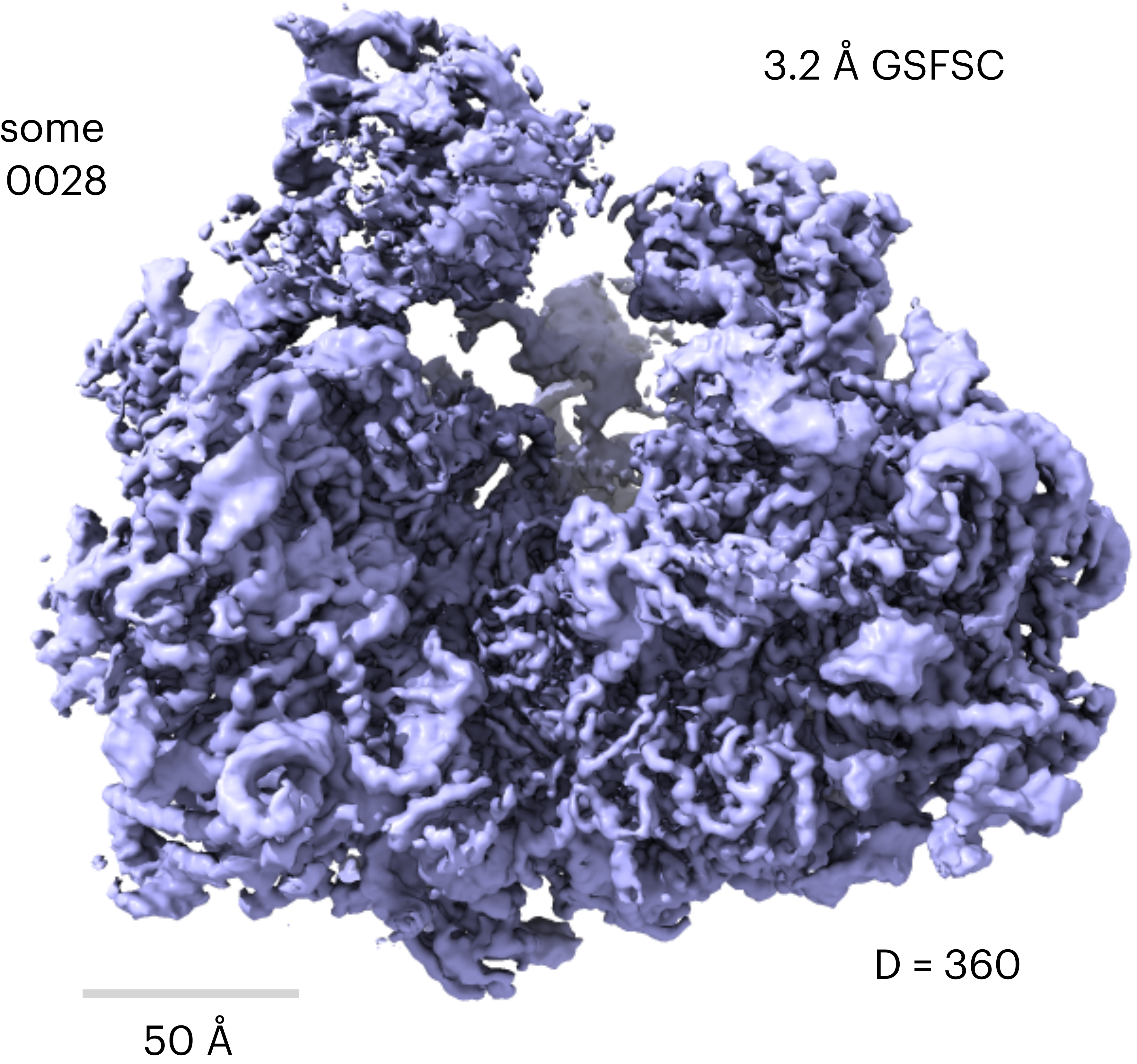
Neural network representation



1024x10 architecture
50 epochs

80S ribosome
EMPIAR 10028

Voxel-based representation



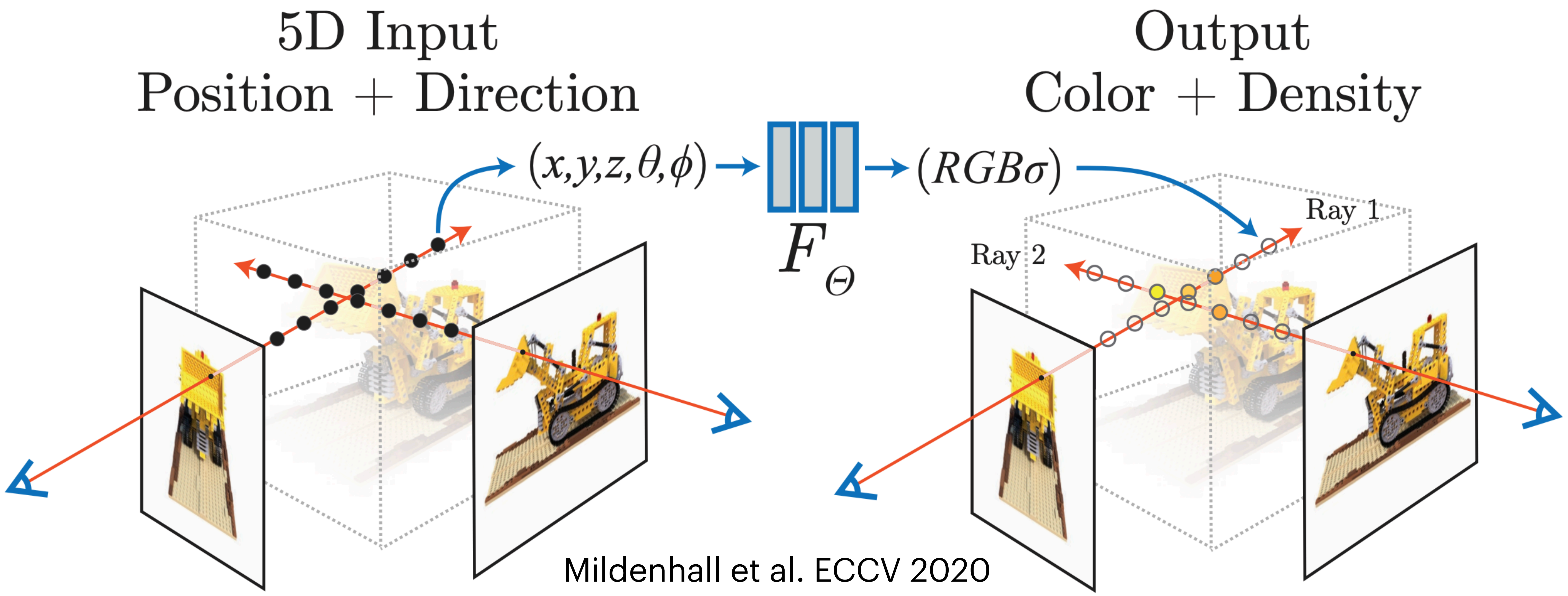
3.2 Å GSFSC

D = 360

Zhong et al., Nature Methods 2021

Neural fields show broad applicability

Neural Radiance Fields



5D Input
Position + Direction

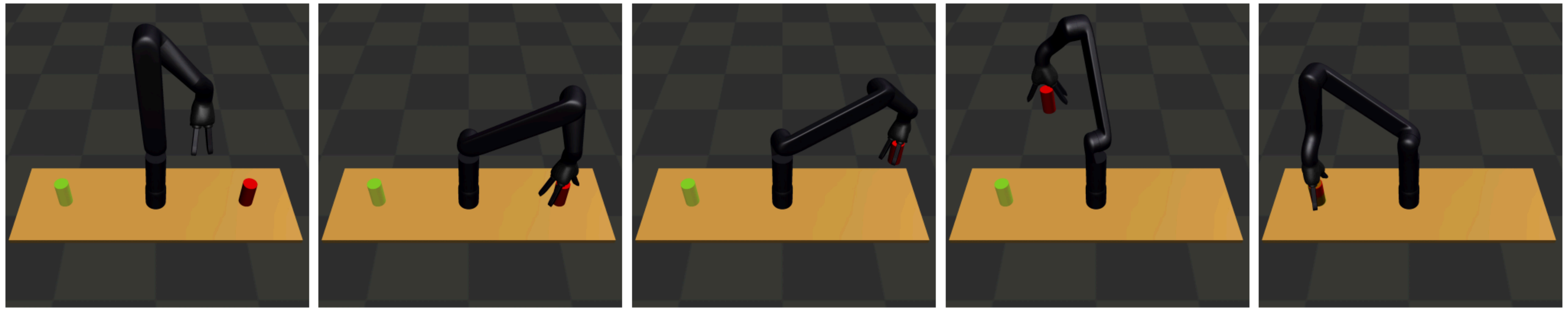
Output
Color + Density

Mildenhall et al. ECCV 2020

Scene Representation

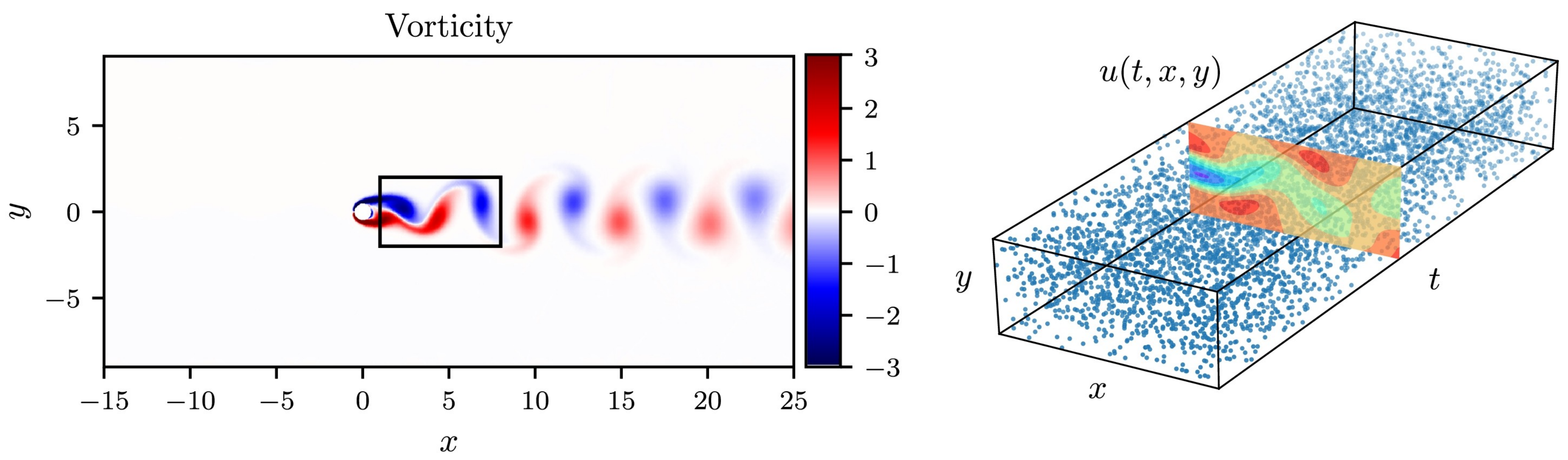
Martin-Brualla et al. CVPR 2021

Robotic Planning



Sutanto et al. CoRL 2020

Physics

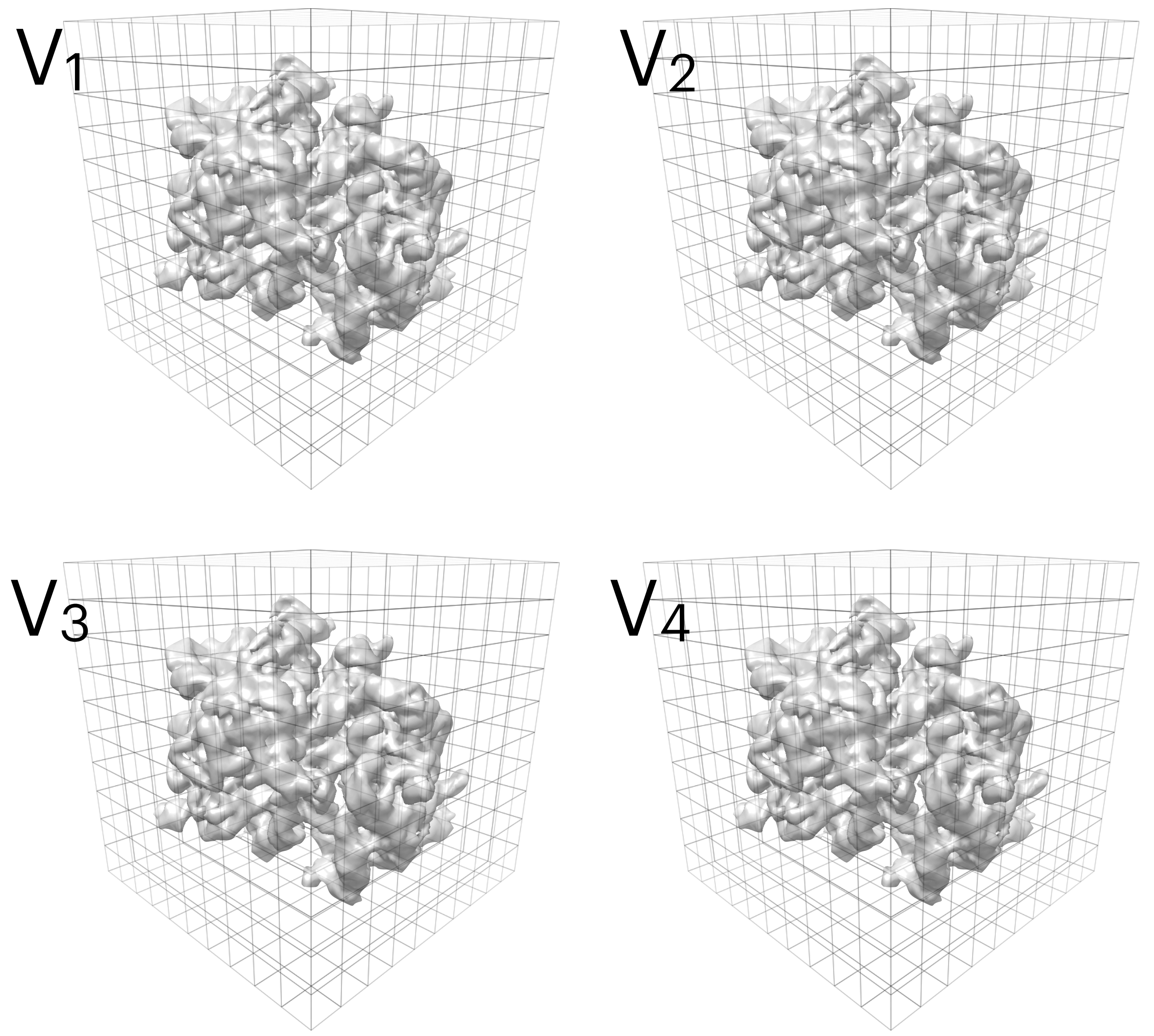


Raissi et al. JCP 2019

Latent variable models for heterogeneous structures

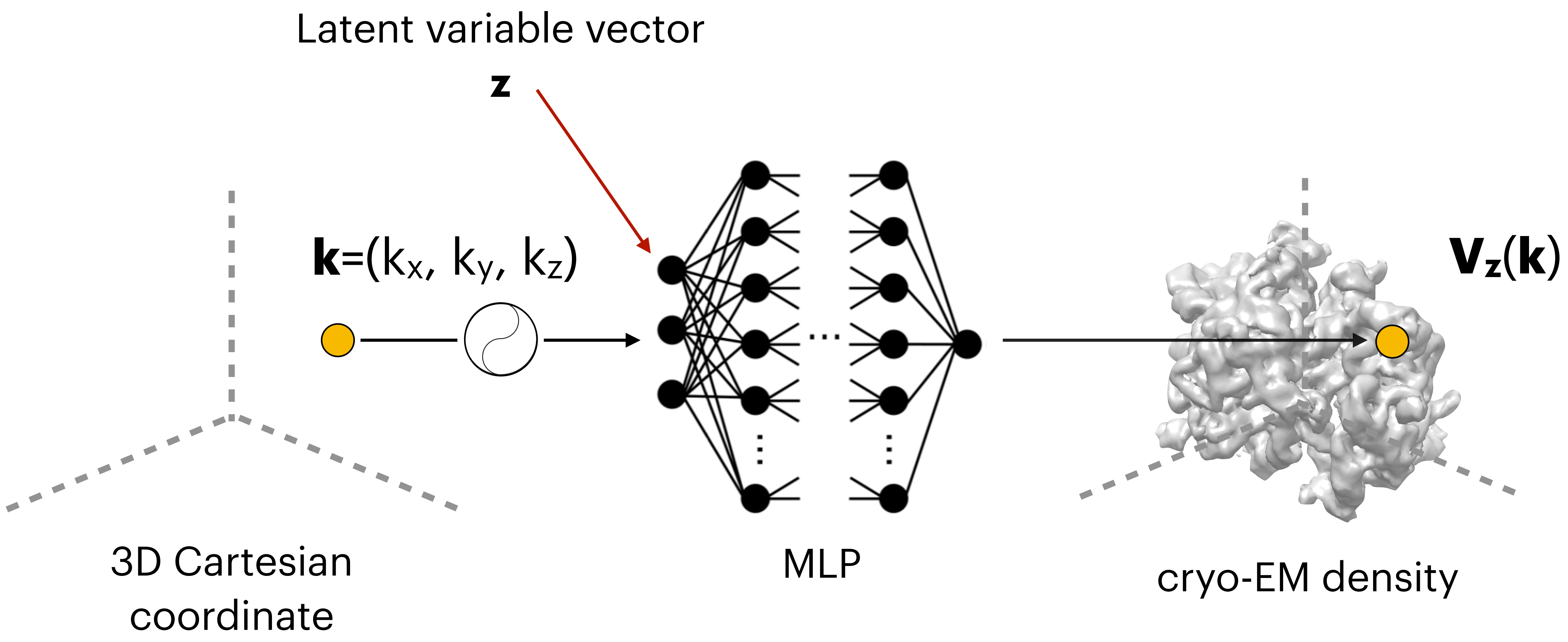
Multiclass reconstruction

V_z, where $z \in \{1, ..., K\}$



cryoDRGN❄️🐉

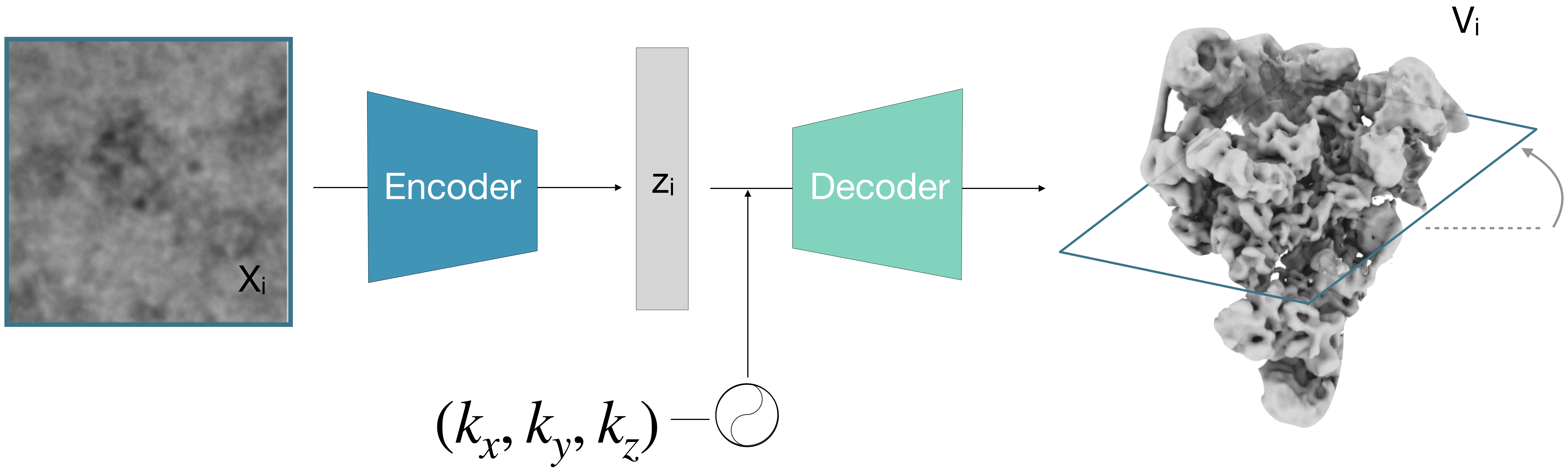
$$z \rightarrow V_\theta(z), z \in \mathbb{R}^D$$



CryoDRGN Architecture

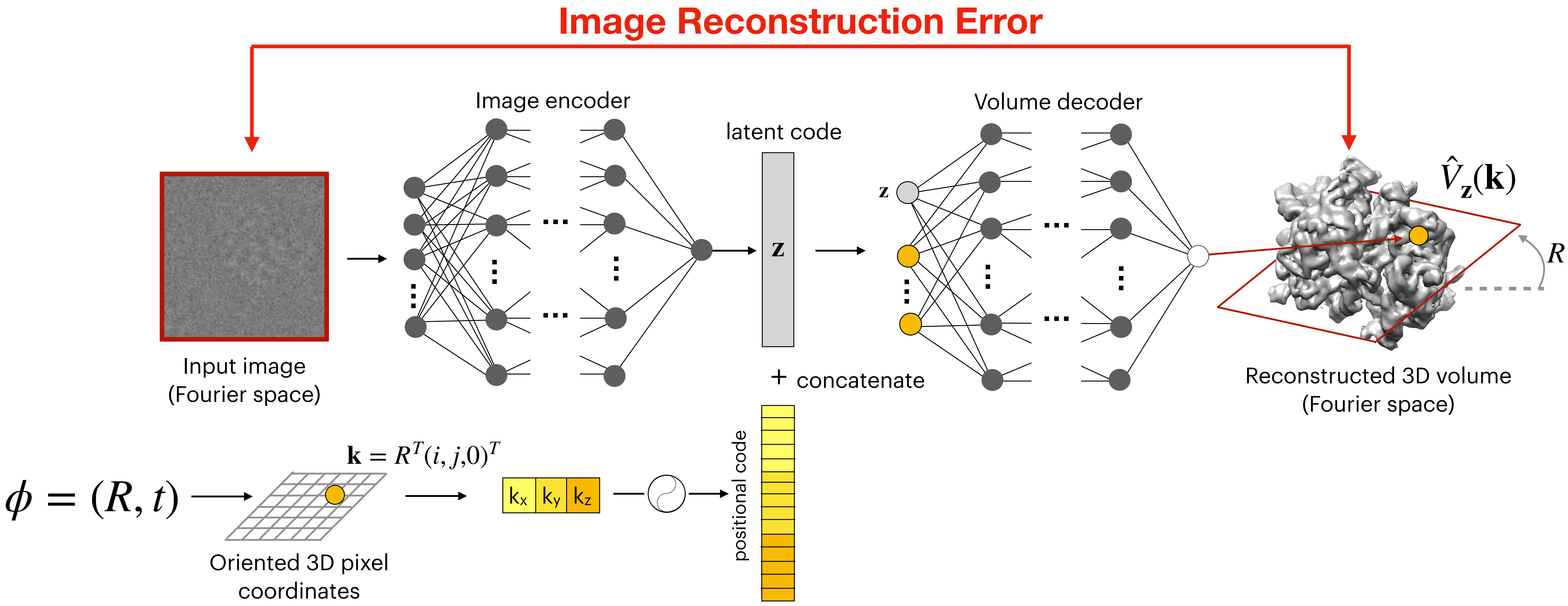
Variational autoencoder

- Given a particle, the **encoder** predicts a latent variable z describing the particle’s conformation
- Given z and spatial frequencies, the **decoder** outputs cryo-EM densities



Zhong *et al*. ICLR 2020. Nature Methods 2021

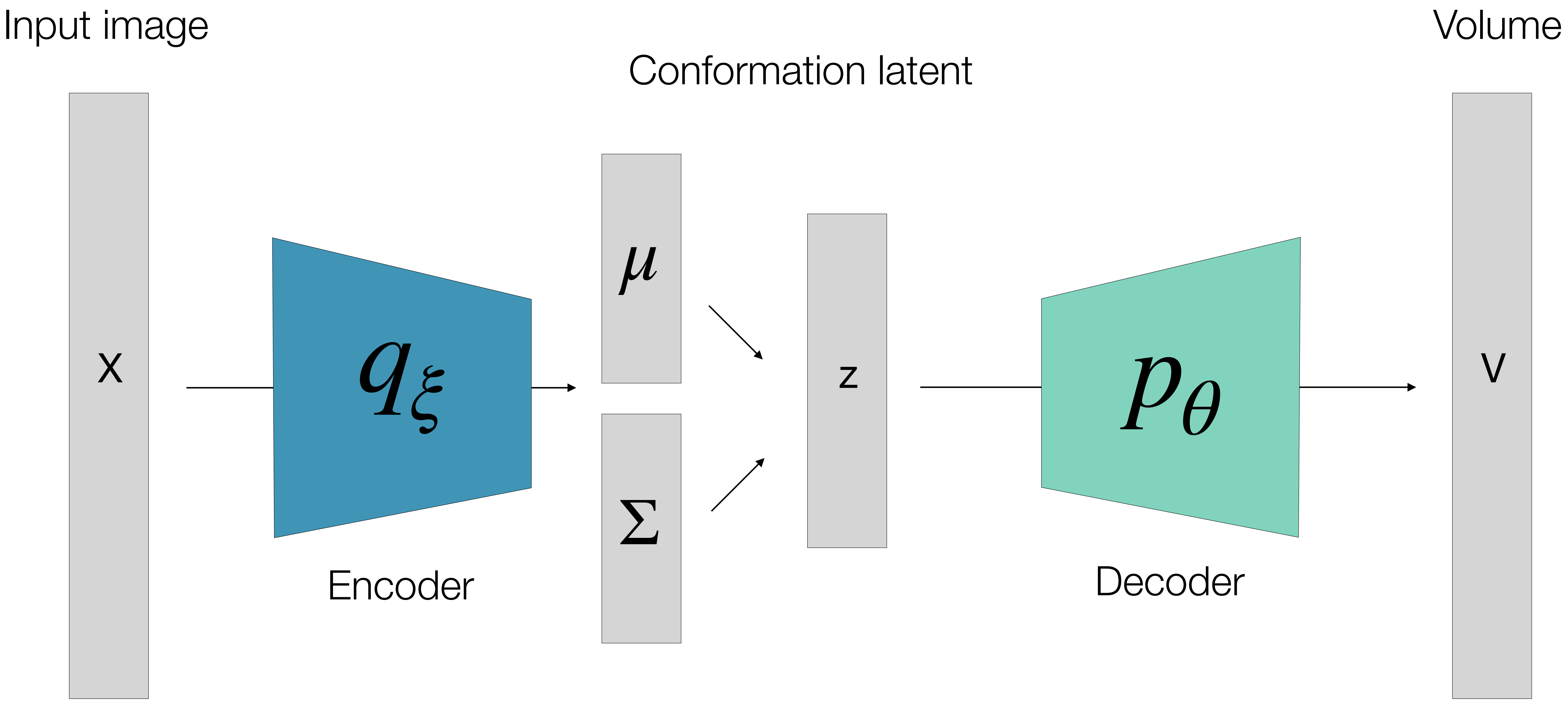
CryoDRGN Training



- Input images are oriented central slices of the reconstructed volume by the Fourier slice theorem
- Training loss is mean squared error between input image and reconstructed image (plus a KL-divergence term)

Zhong *et al*. ICLR 2020. Nature Methods 2021

Amortized Latent Inference



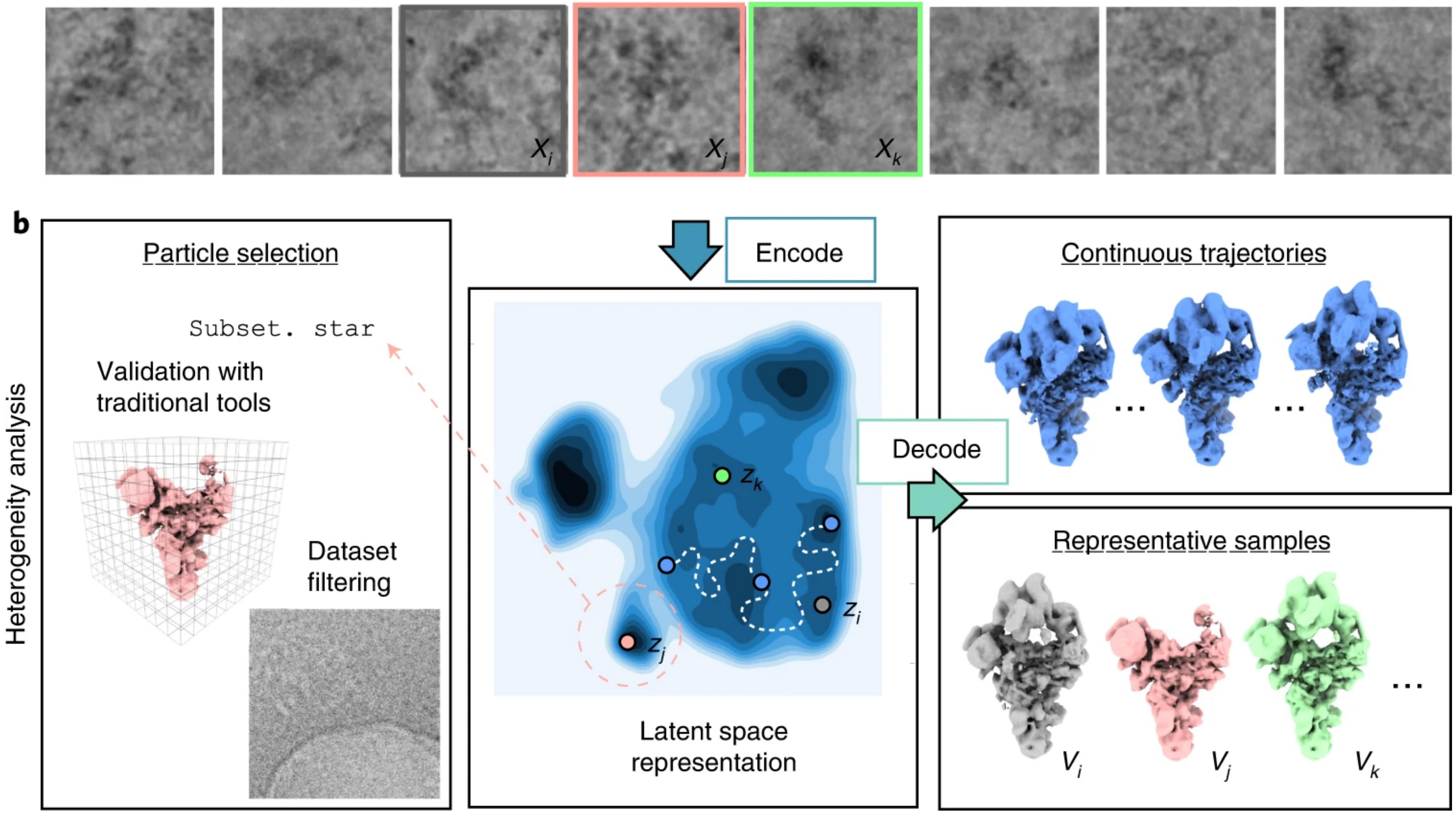
The VAE extends the AE as inference of a probabilistic model — *"a regularized autoencoder"*

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

Zhong *et al.* ICLR 2020. Nature Methods 2021

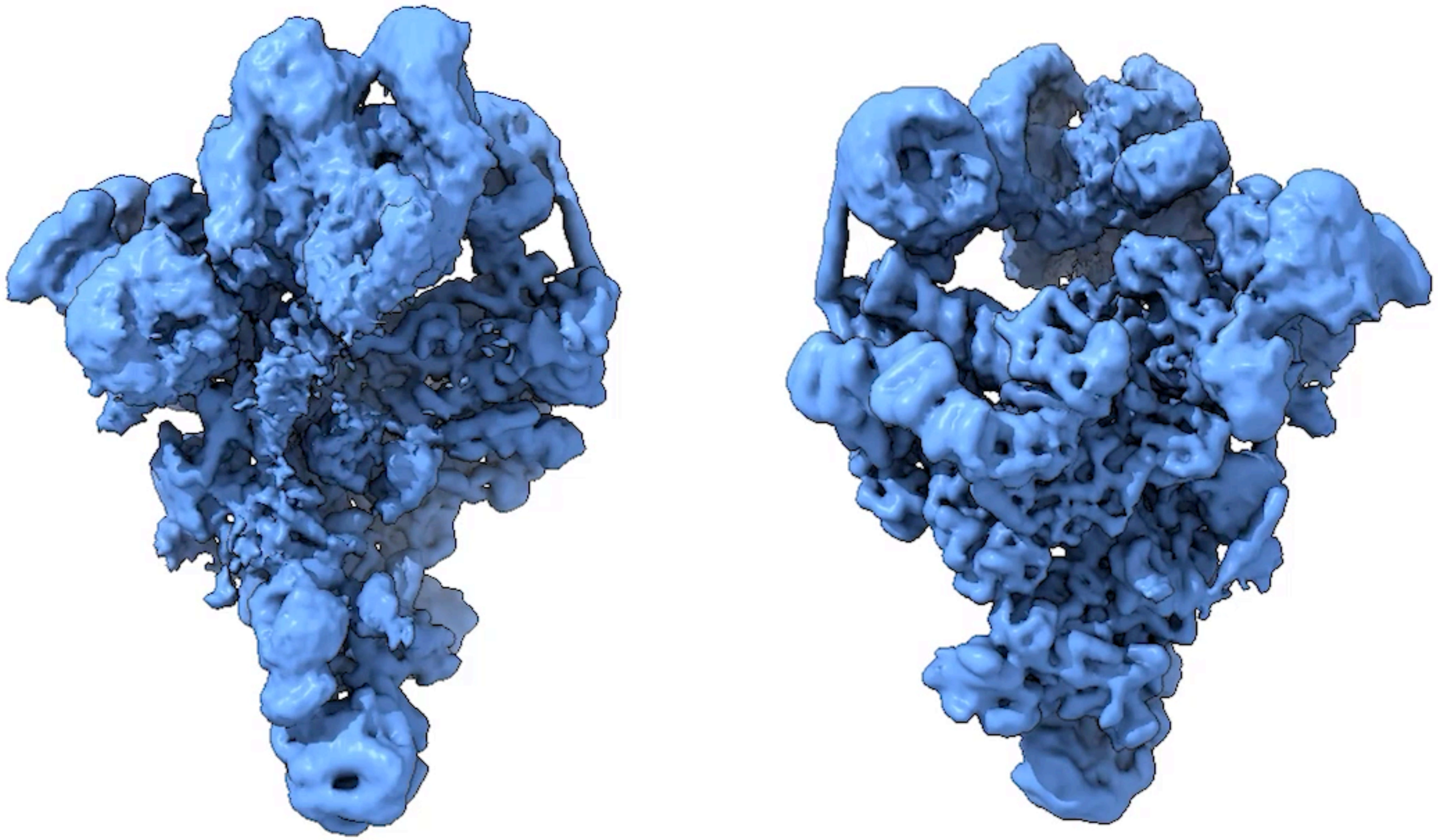
CryoDRGN Inference

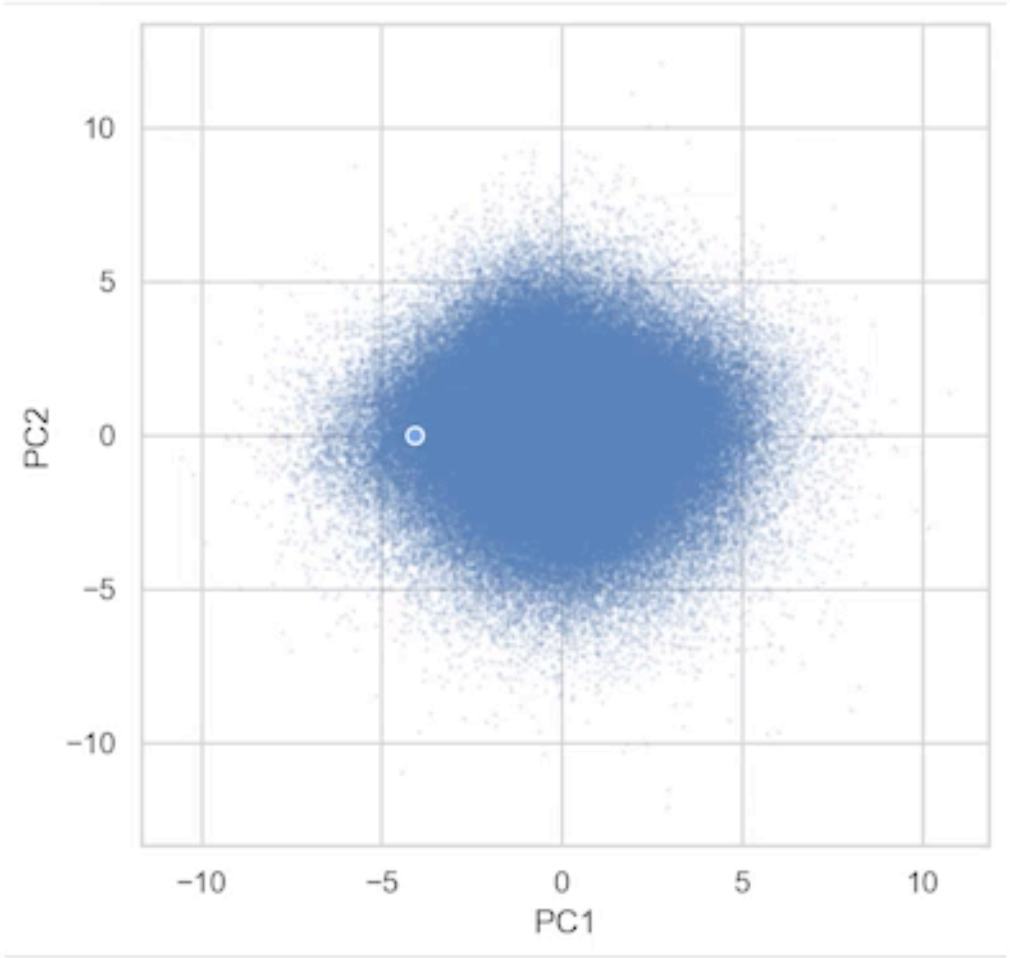
- Use the encoder network to evaluate the latent embedding **z** for each image
- Use the decoder network to generate **V** at different values of **z**



Zhong *et al*. ICLR 2020. Nature Methods 2021

Reconstructing continuous variability of the spliceosome



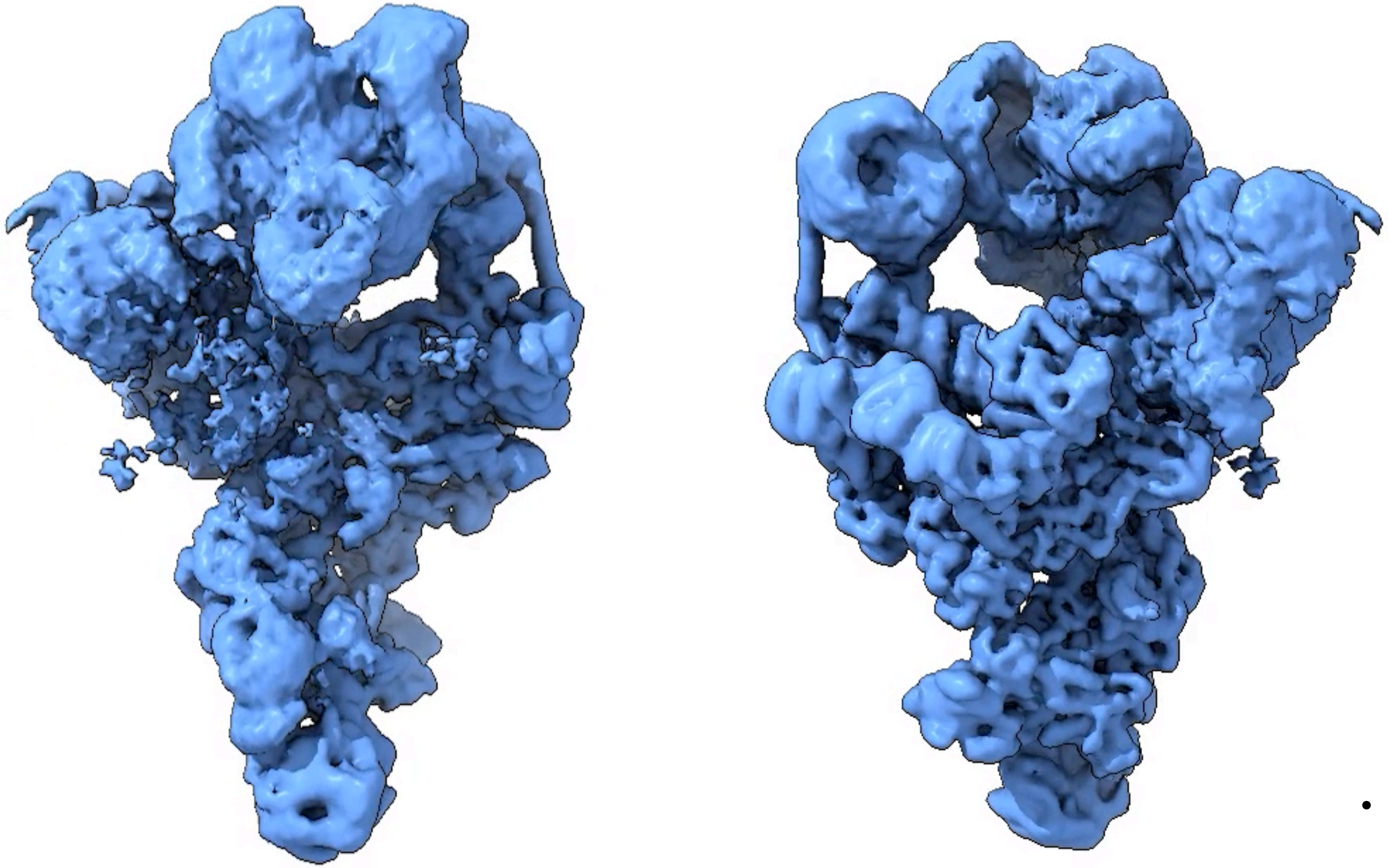


[EMPIAR-10180]

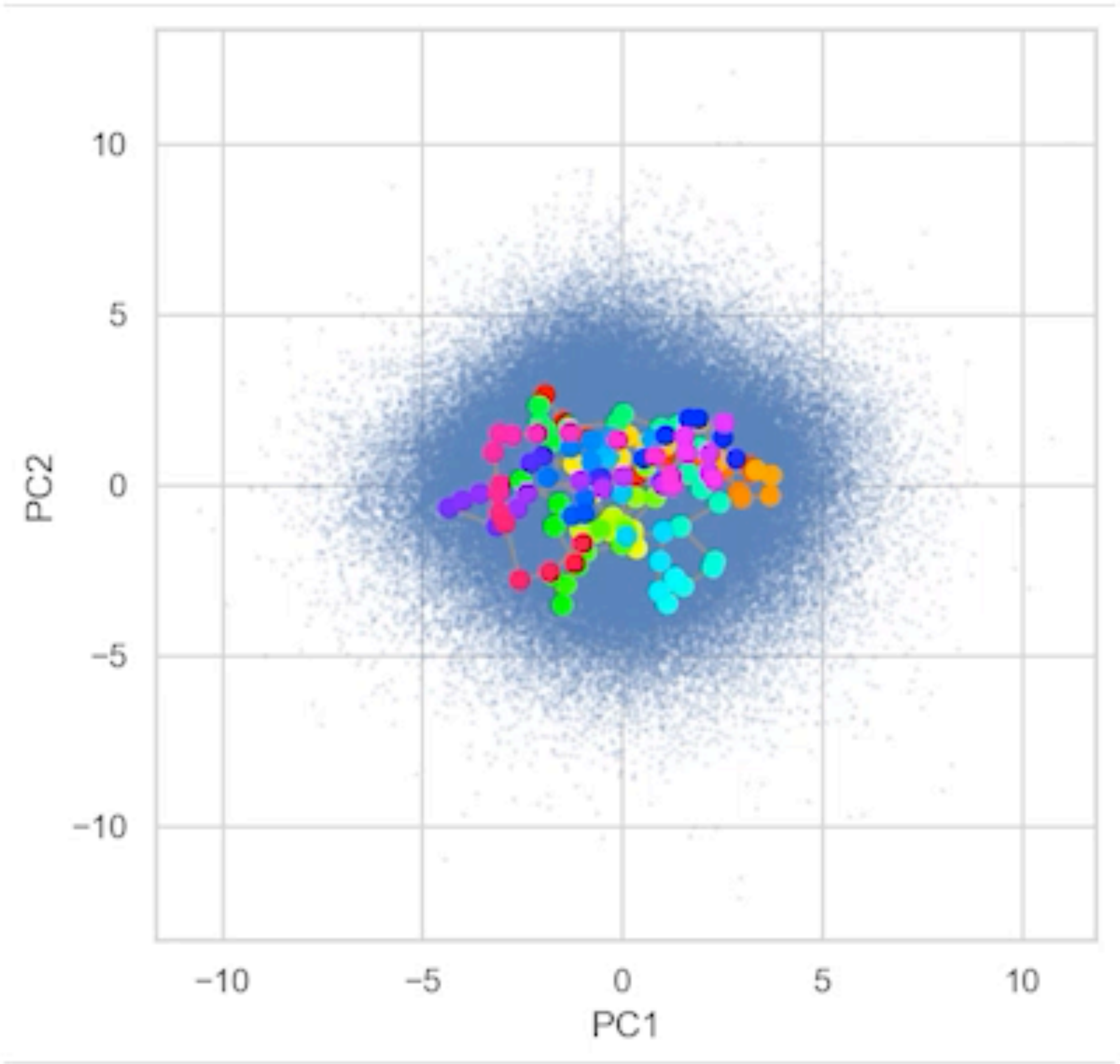
```
cryodrgn pc_traversal
```

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

Generating trajectories with a graph traversal algorithm



`cryodrgn graph_traversal`

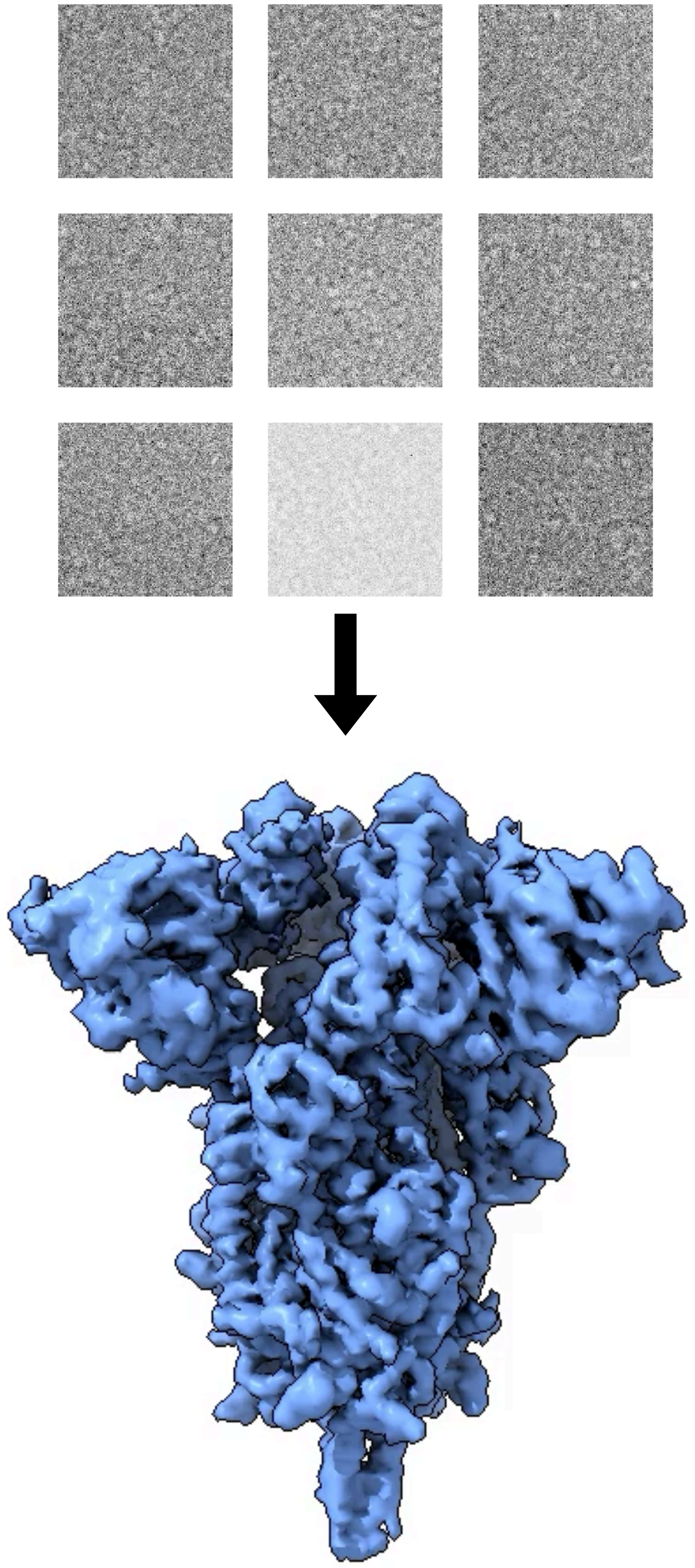


[EMPIAR-10180]

- Explore the learned distribution
- Extended with latent space diffusion models (Kreis, Dockhorn, Li, Zhong 2022)

Summary

- Cryo-EM captures proteins in their “near-native” state, sampled from an “equilibrium” distribution of conformations
- Principled probabilistic techniques underlie cryo-EM reconstruction
- Tools for heterogeneous reconstruction encode different assumptions about the heterogeneity present in the data
- CryoDRGN’s implicit neural representation provides a expressive model of heterogeneity



Zhong *et al*. ICLR 2020. Nature Methods 2021