Advance Bioinformatics 1

ARTICLES
https://doi.org/10.1038/s41592-020-01049-4

nature methods

Check for updates

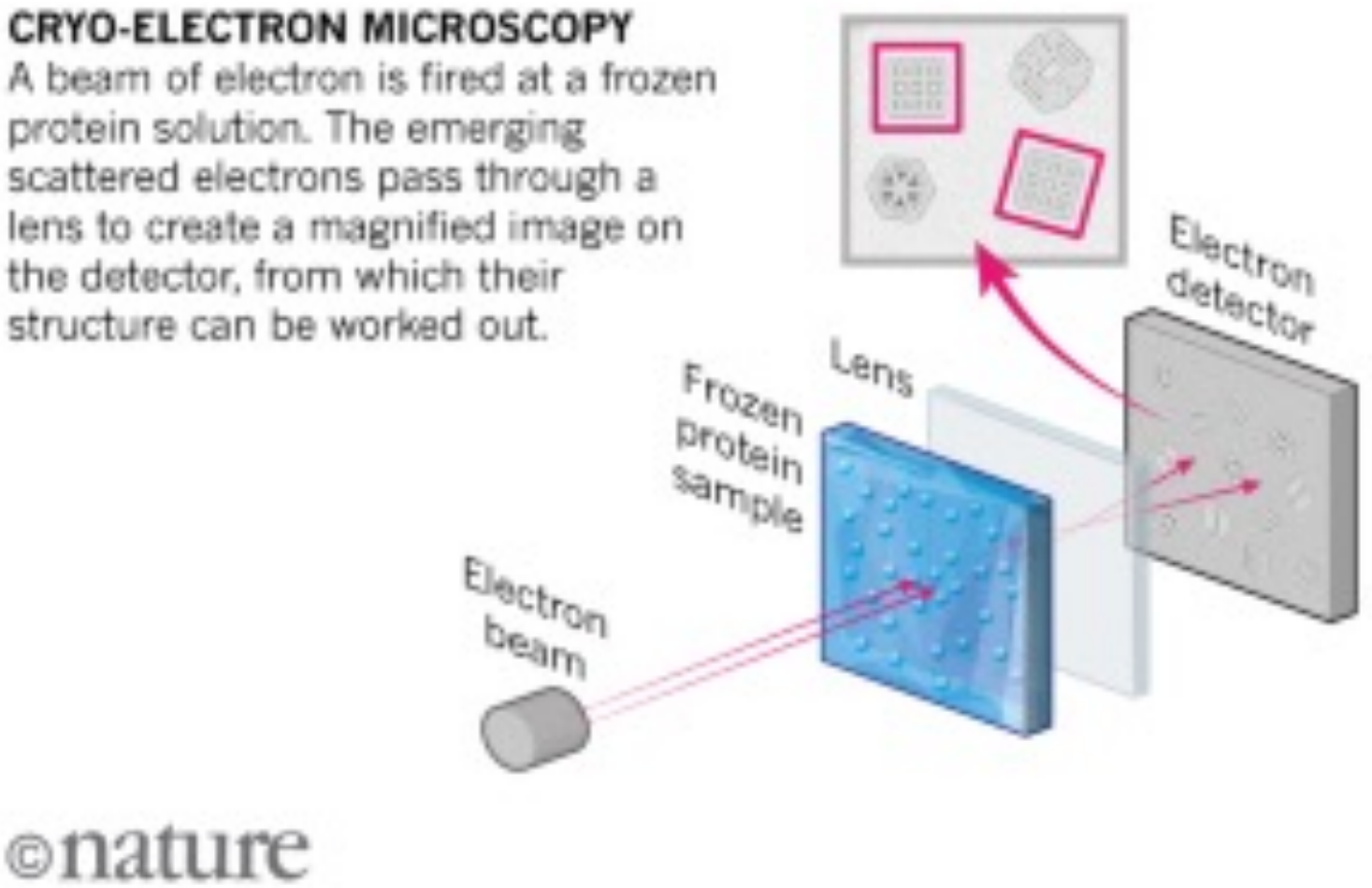
CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong[1,2], Tristan Bepler[1,2], Bonnie Berger[2,3] ✉ and Joseph H. Davis[1,4] ✉

2021. 10. 13
Junsun Park

Nature Methods | VOL 18 | FebruarY 2021 | 176–185

What is Cryo-electron microscopy (Cryo-EM)?



Light microscope

https://www.leica-microsystems.com/products/light-microscopes/

Nature **volume 550**, page167 (2017)

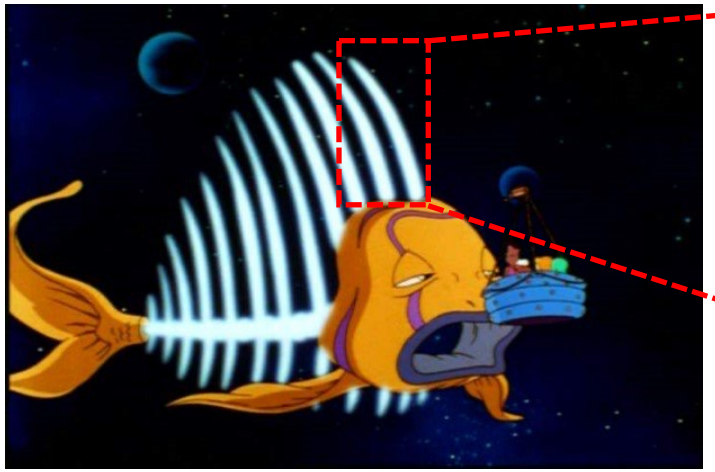
Cryo-electron microscopy technique is a powerful tool for structural biology

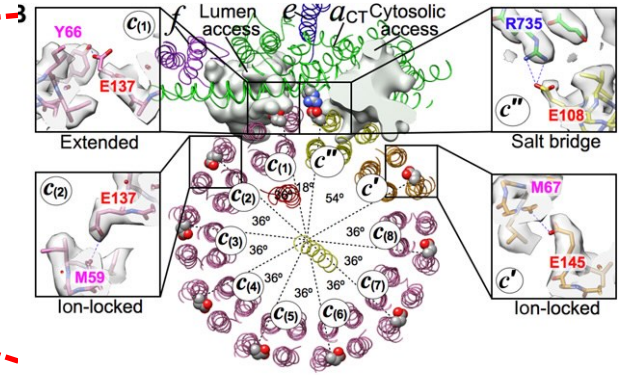
Cryo-electron microscope

+

Biological samples (cells, proteins)

=





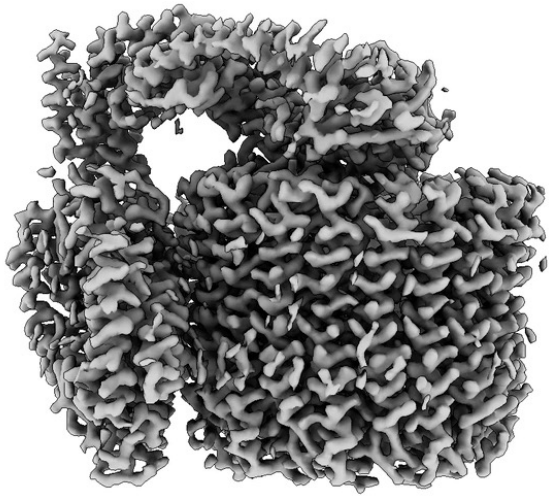
Structural information

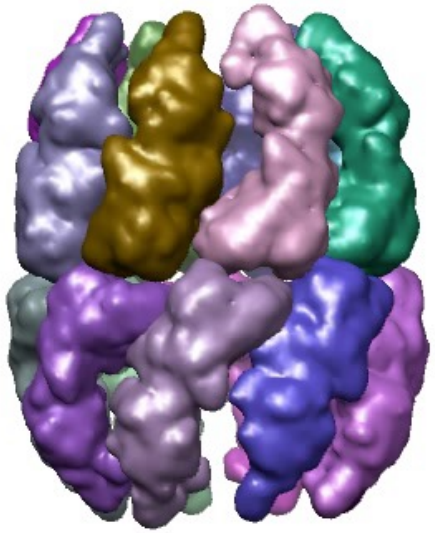
Images from
https://www.thermofisher.com/kr/ko/home/electron-microscopy/products/transmission-electron-microscopes/krios-g4-cryo-tem.html

Science advances, 2021; DOI: 10.1126/sciadv.abb9605

Interpretable electron density maps of biological samples

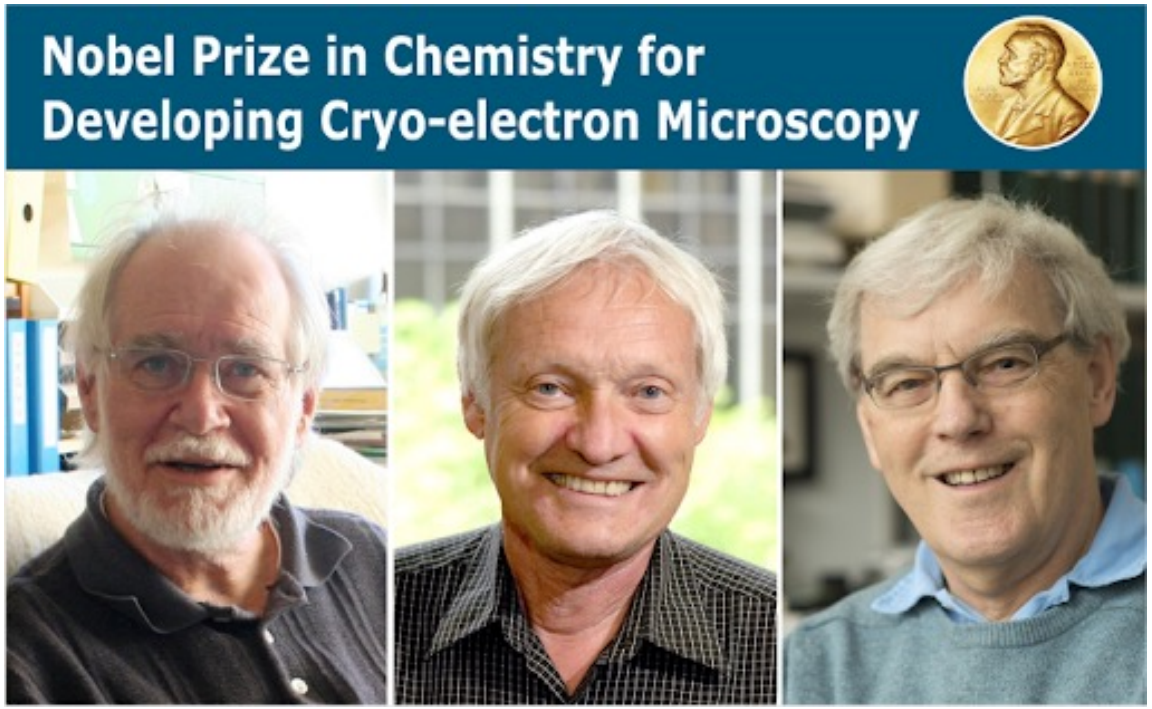
There is a close relationship between the structure of protein and its function

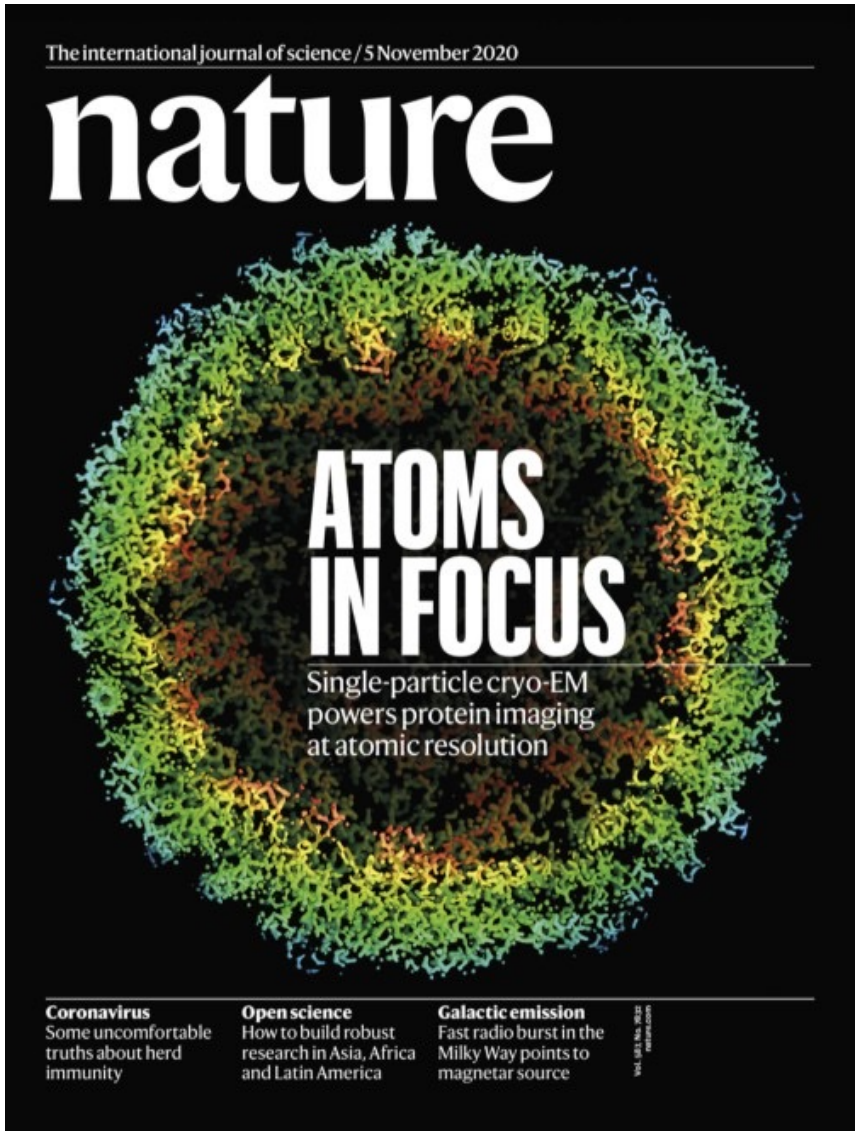




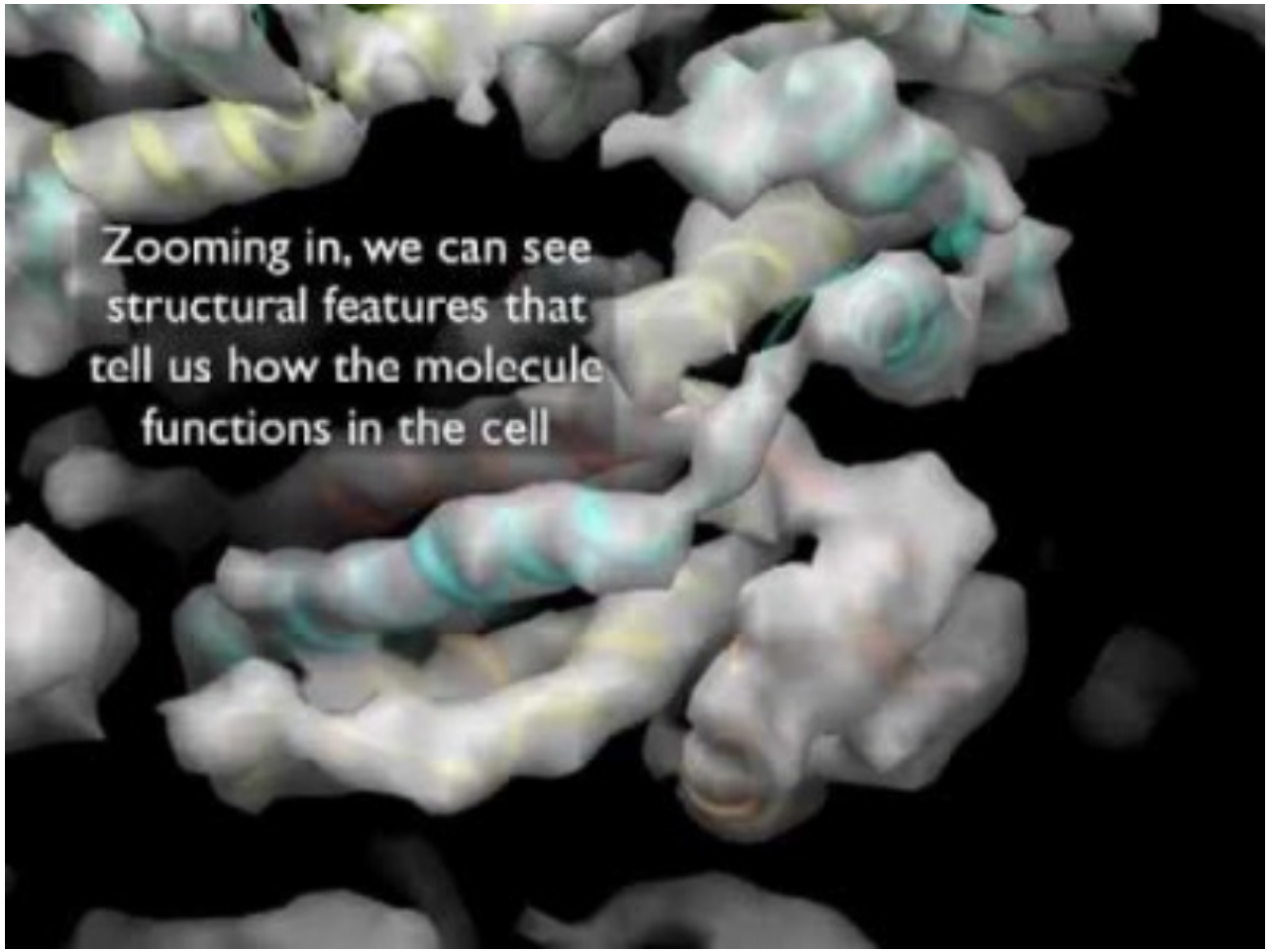
Science advances, 2021; DOI: 10.1126/sciadv.abb9605

Single particle cryo-EM now can visualize protein structure at an atomic resolution





The workflow of single particle analysis cryo-EM

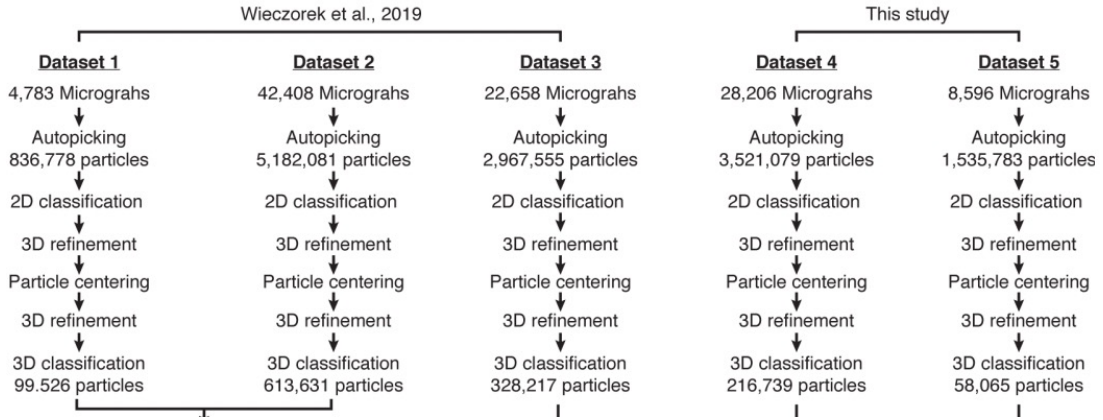


https://www.youtube.com/watch?v=BJKkC0W-6Qk

Averaging individual particles in random orientation and build a 3D model from the projection image!

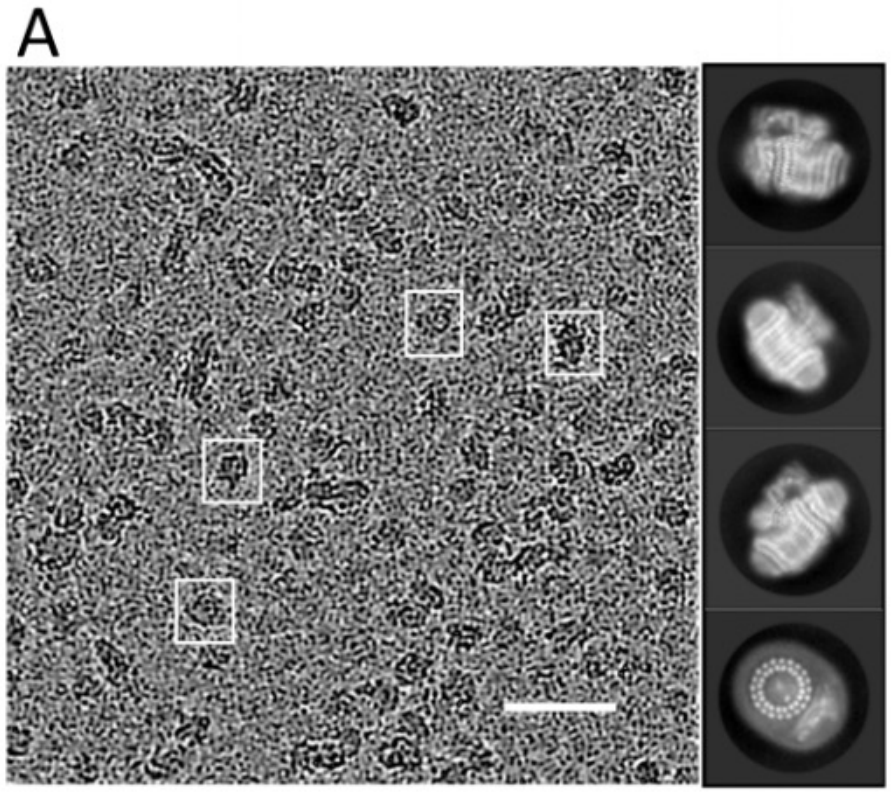
The challenges of single particle Cryo-EM data processing

- Large data set of raw images (up to 20TB these days)
- Ill-posed problem, (we don't know the answer: the original protein shape, or the pose)
- Low signal to noise level
- Few millions of unique protein particle images: the inherent heterogeneity (non-linear hidden variables)



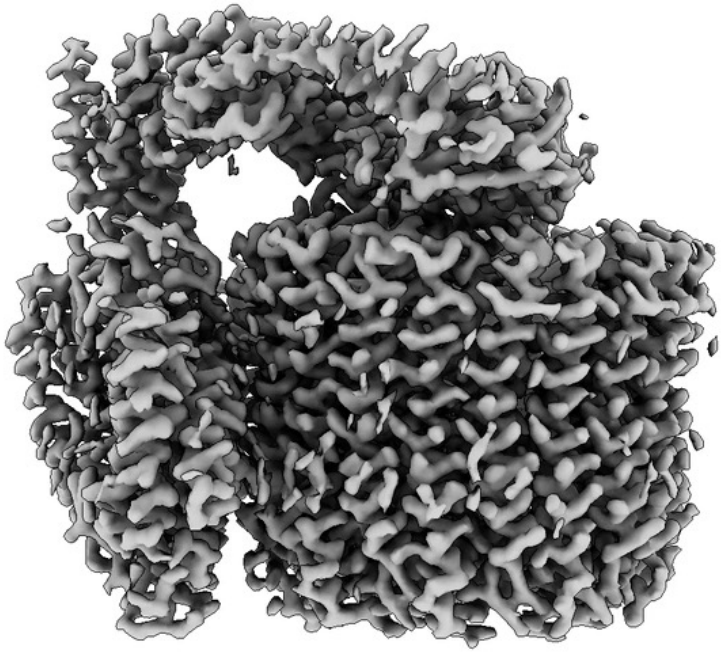
Cell reports, Volume 31, Issue 13, 30 June 2020, 107791

Large data is now used for reconstruction of 3D electron density map



4,083 raw Micrographs (~4TB)

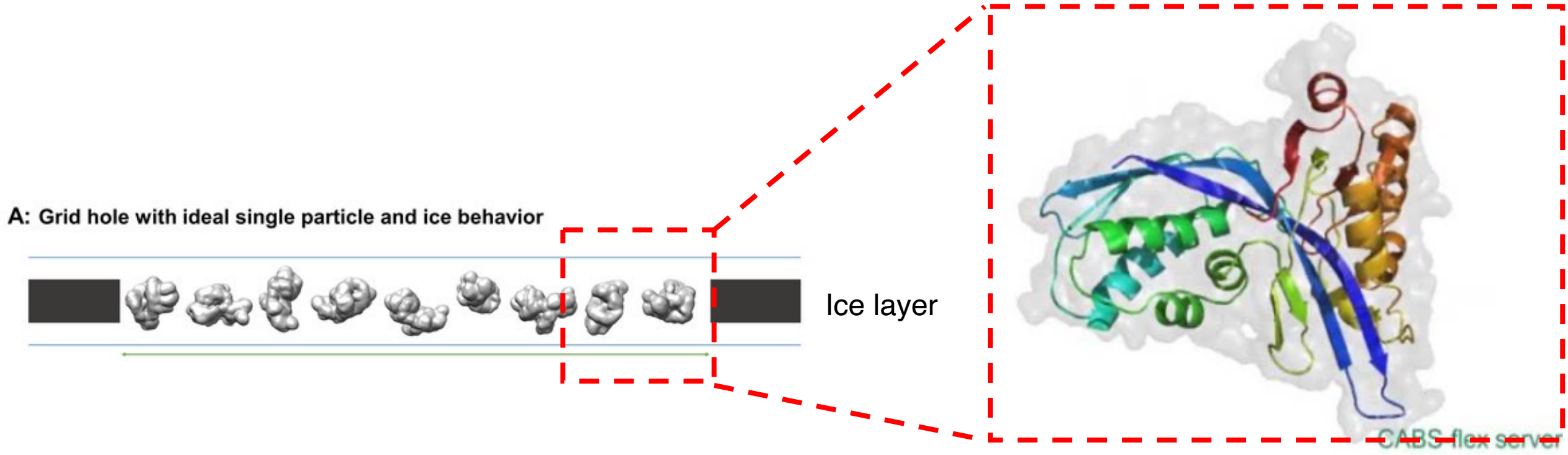
~ 700K individual protein images (few hundreds GB)



3D Cryo-EM map

Science advances, 2021; DOI: 10.1126/sciadv.abb9605

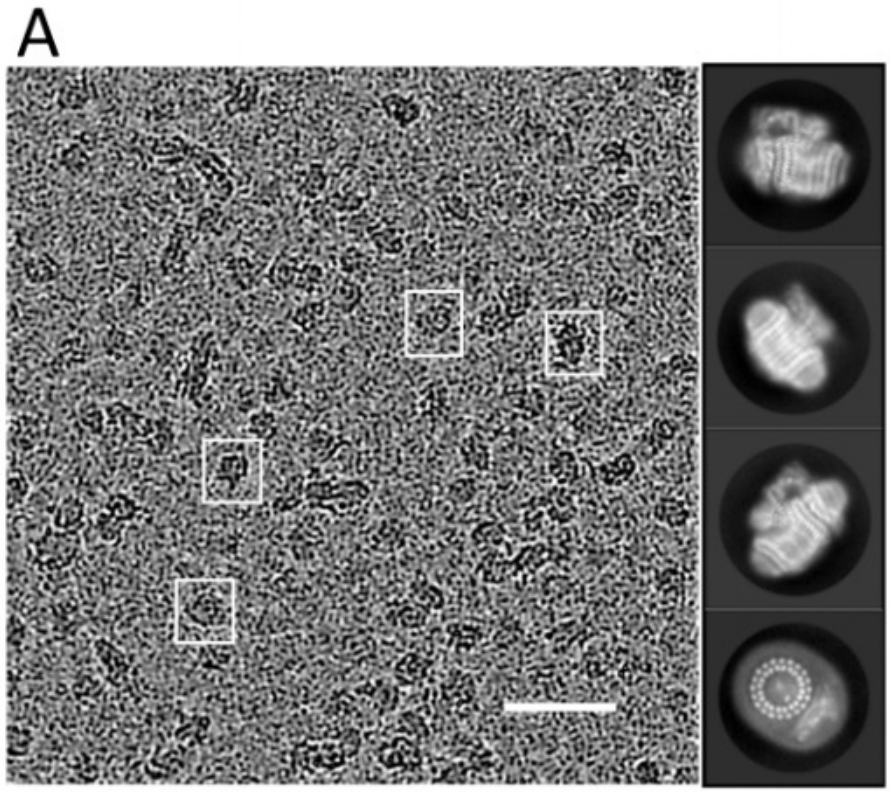
Proteins are flexible. When proteins are embedded in the ice layer very quickly, inherent structural movement or heterogeneity can be captured!



https://www.youtube.com/watch?v=iTKWwkMIn6U&t=1s

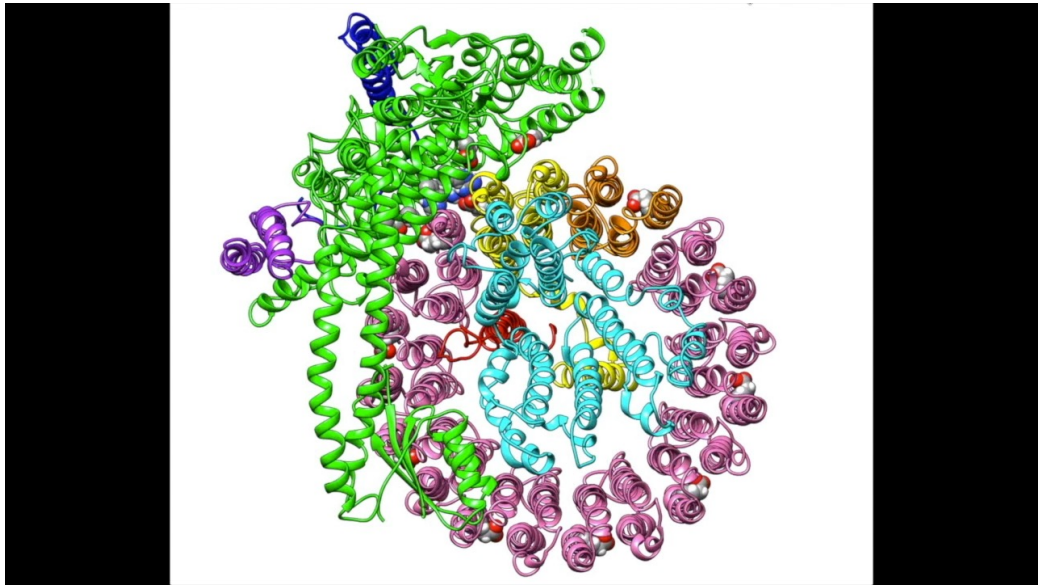
eLife, 2018;7:e34257 DOI: 10.7554/ELIFE.34257

Large data is now used for reconstruction of 3D electron density map revealing heterogeneity of native protein



4,083 raw Micrographs (~4TB)

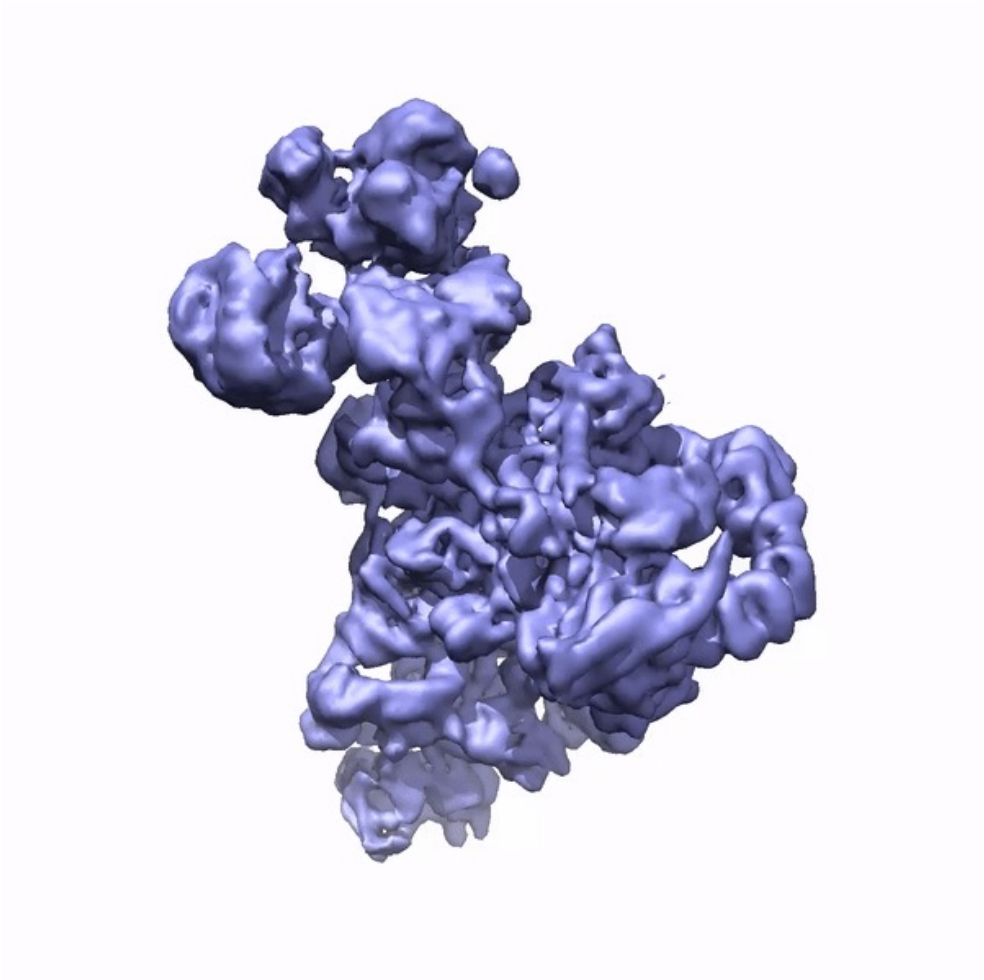
~ 700K individual protein images (few hundreds GB)



Each 3D Cryo-EM map resembles different conformations

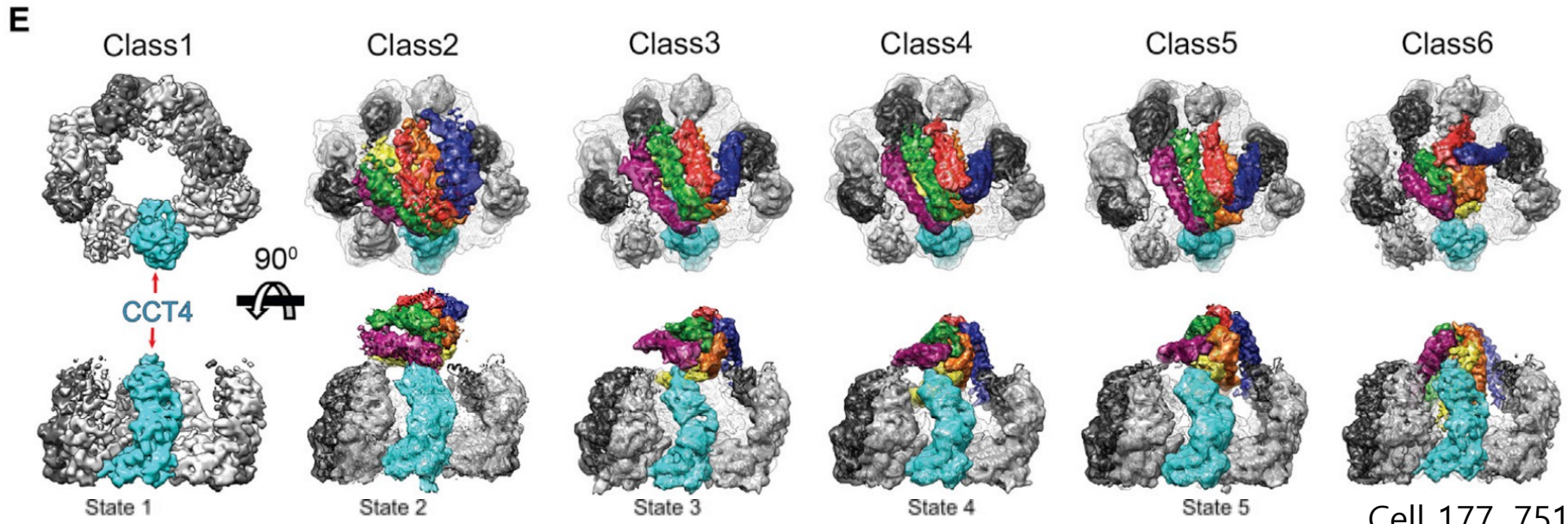
Science advances, 2021; DOI: 10.1126/sciadv.abb9605

Heterogeneity reduces the resolution as it gets averaged out

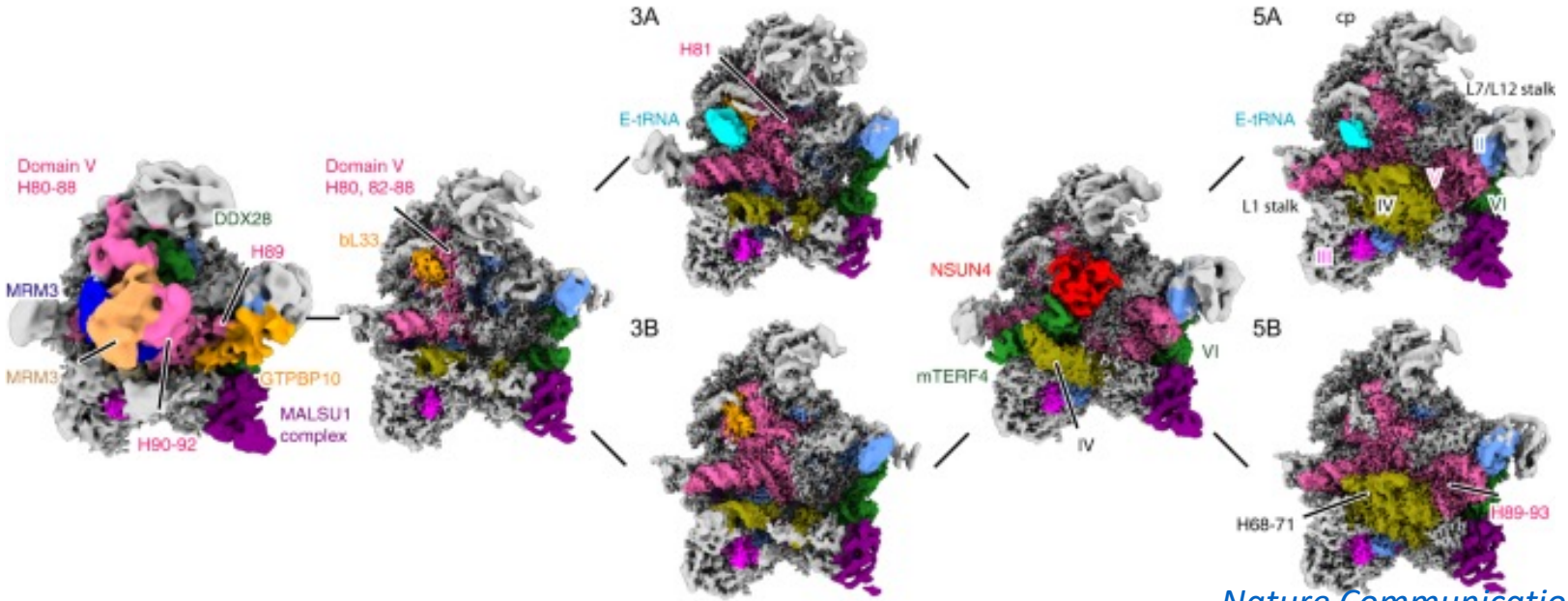


https://guide.cryosparc.com/processing-data/tutorials-and-case-studies/tutorial-3d-variability-analysis-part-one

Revealing heterogeneity or dynamics provides more insights on biological reactions

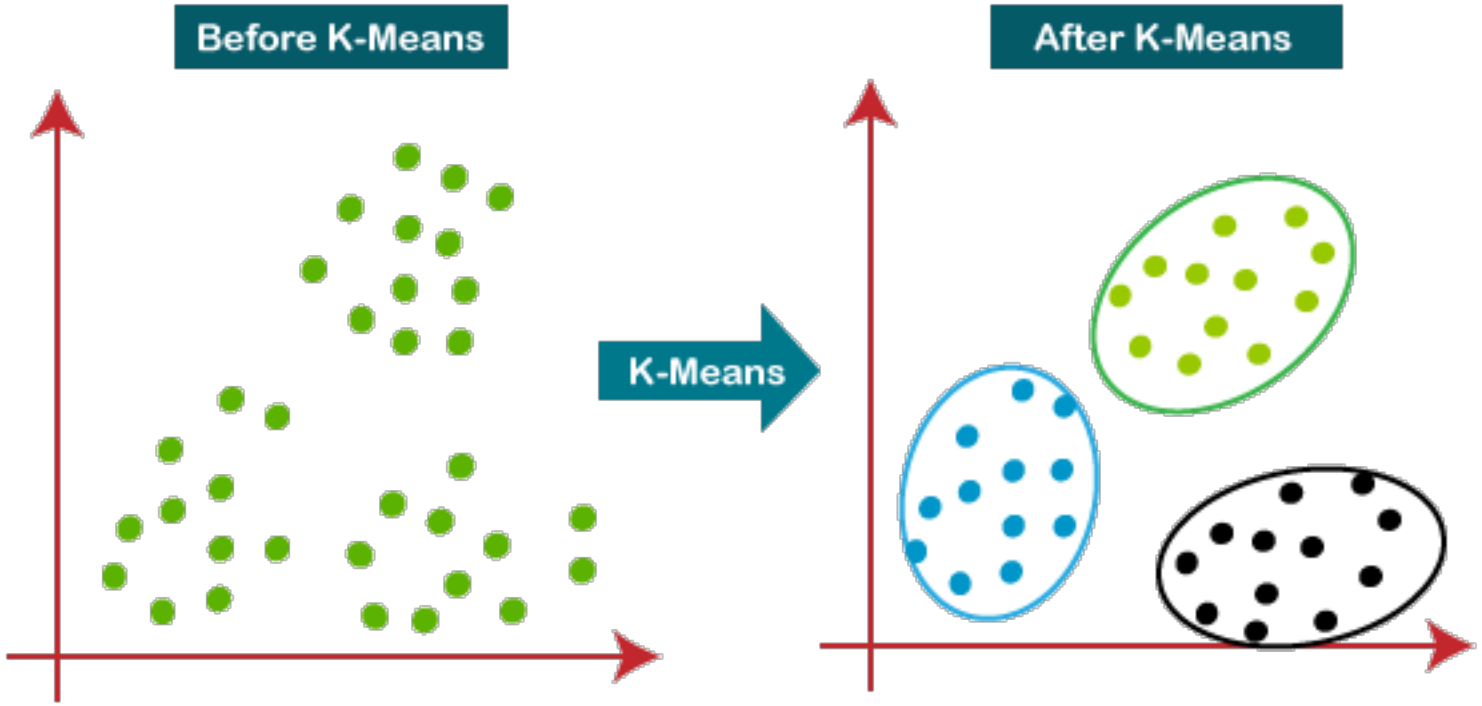


Cell 177, 751–765, April 18, 2019



Nature Communications **volume 12**, Article number: 4544 (202

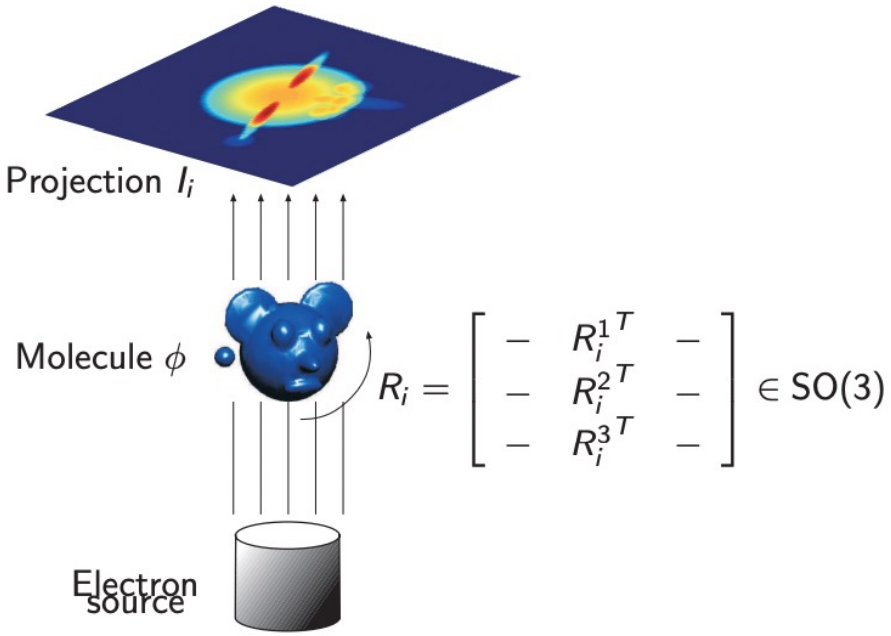
To deal with heterogeneity, a discrete classification (K means clustering) was used for the traditional procedure.



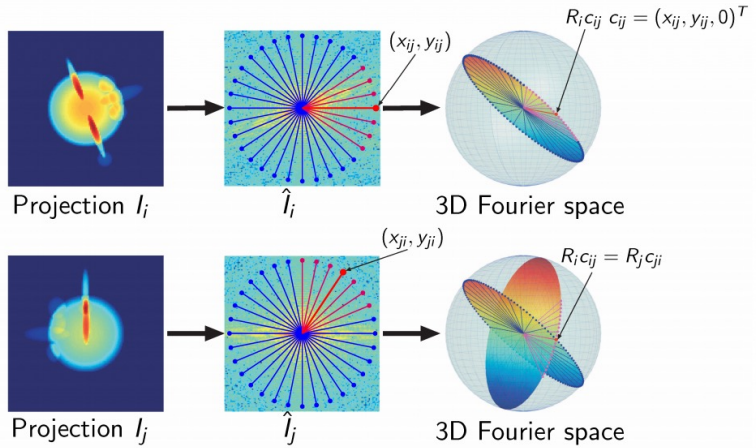
But how many Ks? How to prevent biased results?
What if there is a continuous heterogeneity?

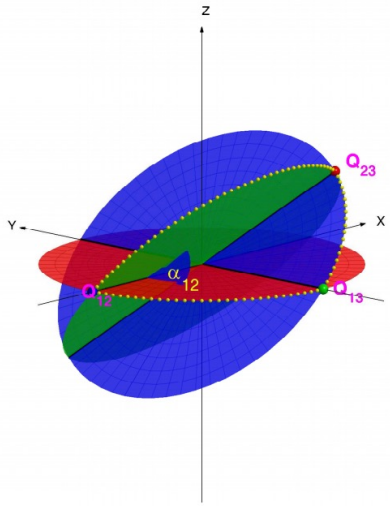
Conventional method for the 3D reconstruction

Fourier slice theorem



$$R_i = \begin{bmatrix} - & R_i^{1\,T} & - \\ - & R_i^{2\,T} & - \\ - & R_i^{3\,T} & - \end{bmatrix} \in SO(3)$$





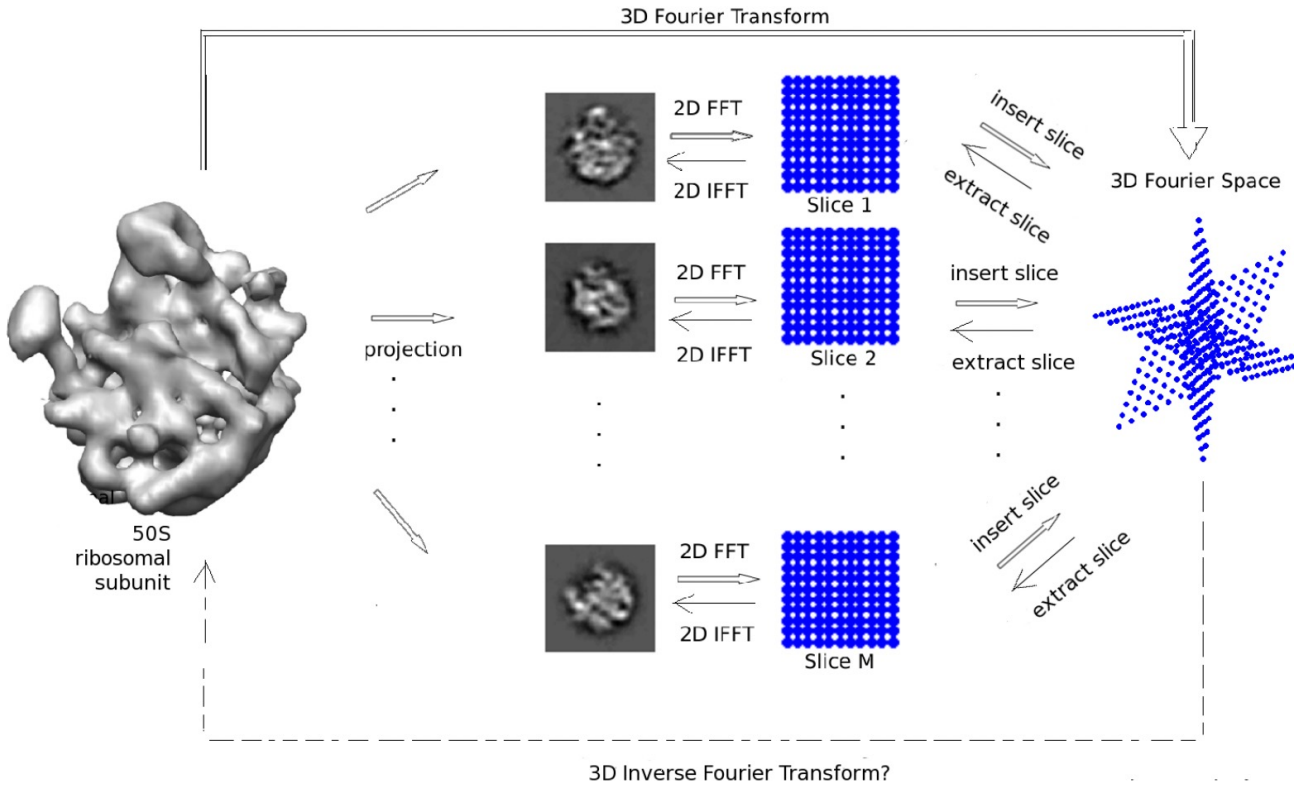
Infer the 3D densities of the voxel array!

Then using statistical weighting for each particle and iterate!

arXiv:1803.06714v1, 2018

Conventional method for the 3D reconstruction

Fourier slice theorem

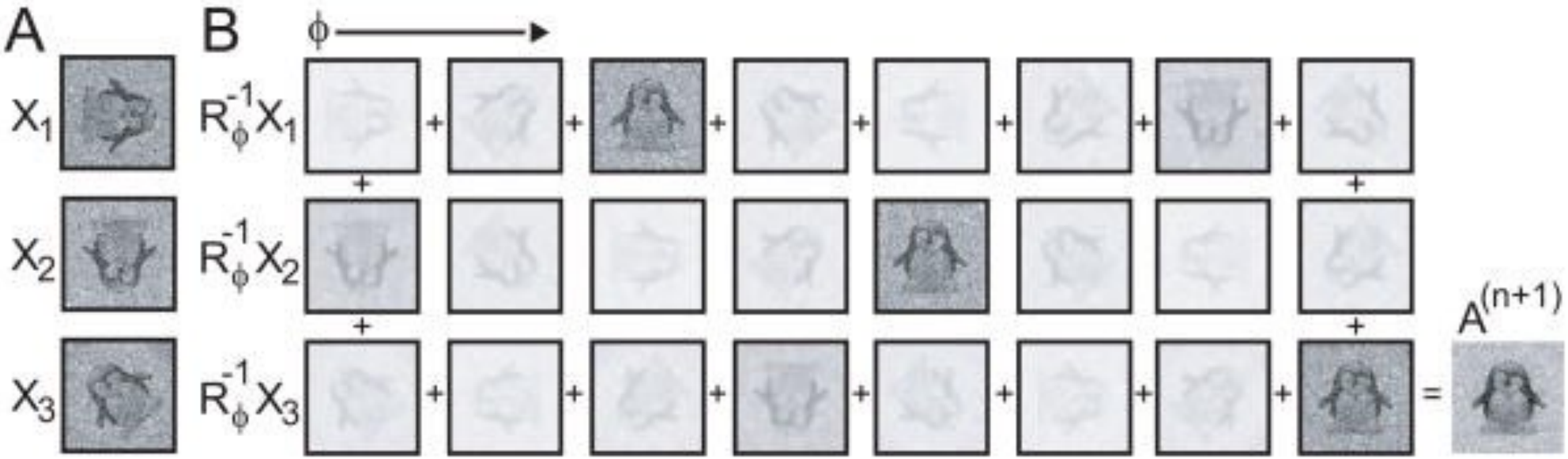


Finding the projection vectors by starting from reference image and iterate

arXiv:1803.06714v1, 2018

Conventional method for the 3D classification (dealing with heterogeneity)

Conventional method for the 3D classification



Bayesian approach for the alignment and finding the accurate angle of each particle!

Methods Enzymol. 2010; 482: 263–294.

There were many approaches to model the continuous nature of flexible molecules.

1. User defined rigid bodies based refinement
eLife **7**, e36861 (2018)

2. Principle component analysis
Journal of Structural Biology Volume 213, Issue 2, June 2021, 107702

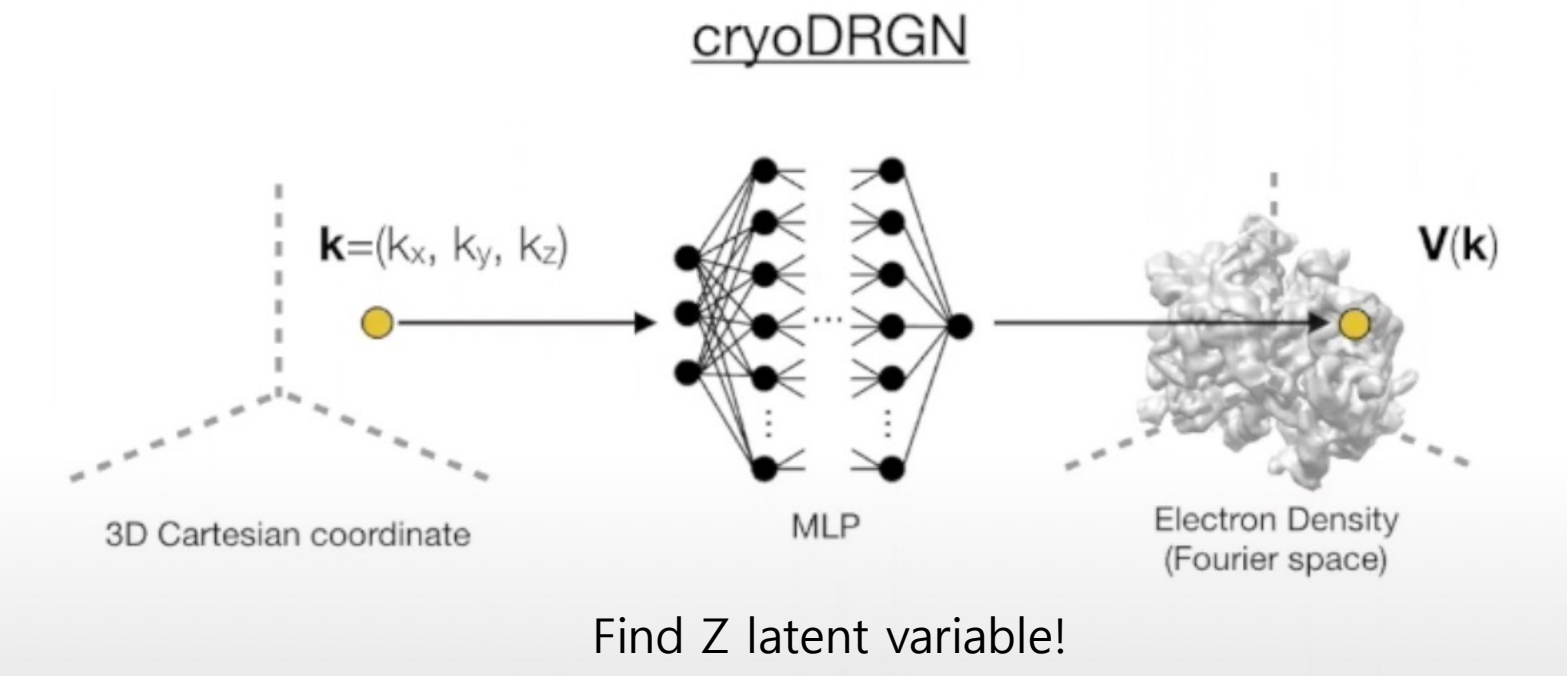
3. Manifold embedding approach
Methods **100**, 61–67 (2016)

4. Other algorithms..

→ Produces artifacts due to the linear subspace model (1,2) or computationally heavy cost

Heterogeneity can be modeled in non-linear manner (high dimensional manifold embedding)

Cryodrgn uses neural network-based reconstruction to build an electron density map



Find Z latent variable!

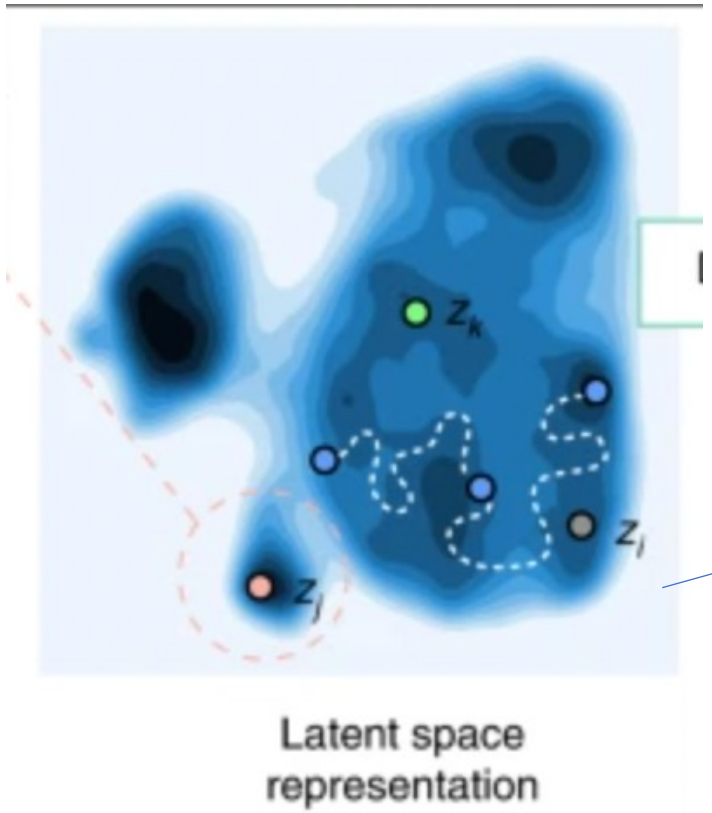
$$pe^{(2i)}\left(k_j\right) = \sin\left(k_j D\pi\left(\frac{2}{D}\right)^{\frac{i}{\frac{D}{2}-1}}\right), i = 0, \ldots, \frac{D}{2} - 1; k_j \in k$$

$$pe^{(2i+1)}\left(k_j\right) = \cos\left(k_j D\pi\left(\frac{2}{D}\right)^{\frac{i}{\frac{D}{2}-1}}\right), i = 0, \ldots, \frac{D}{2} - 1; k_j \in k$$

arXiv:1909.05215, 2020

Cryodrgn also learns innate heterogeneity of protein by using Deep learning (the first neural network-based approach for cryo-EM reconstruction)

Find Z latent variable : position on a conformational energy landscape

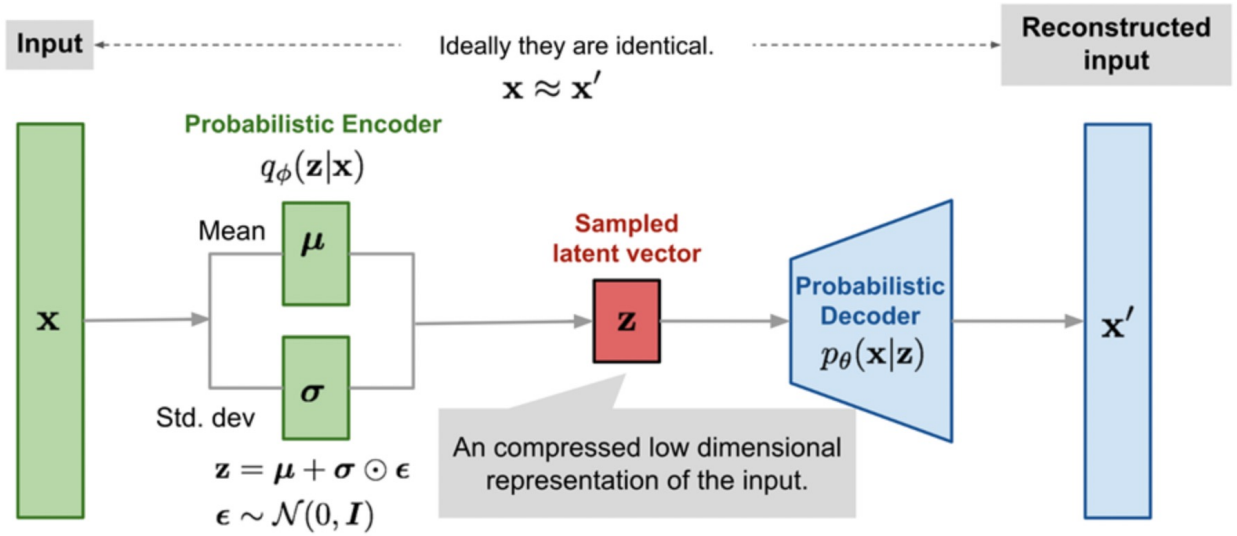


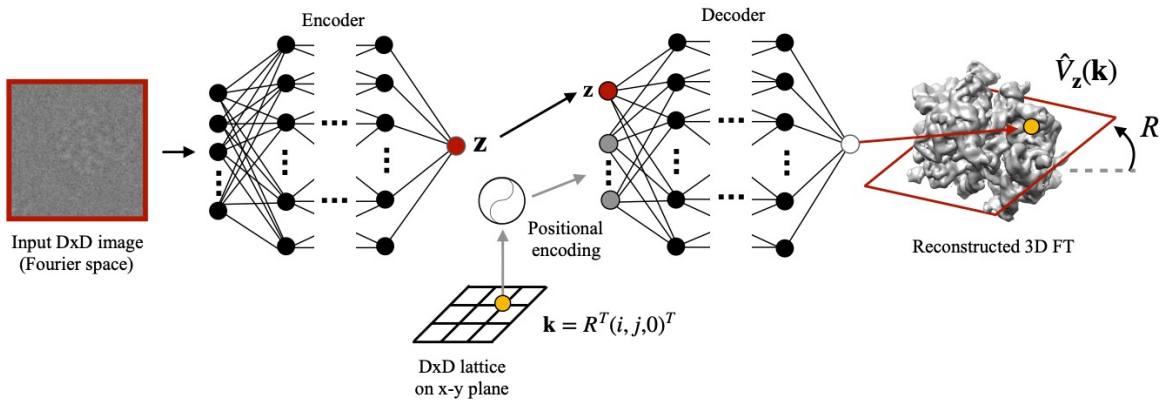
Assumption: the heterogeneous structures can be embedded within a continuous, low-dimensional manifold in the latent space. (dimensionality specified by the user)

Goal: learn the generative model from a continuous latent space to 3D volumes

Nature Methods **volume 18**, pages176–185 (2021)

Cryodrgn architecture adopts encoder-decoder system using two neural network structure





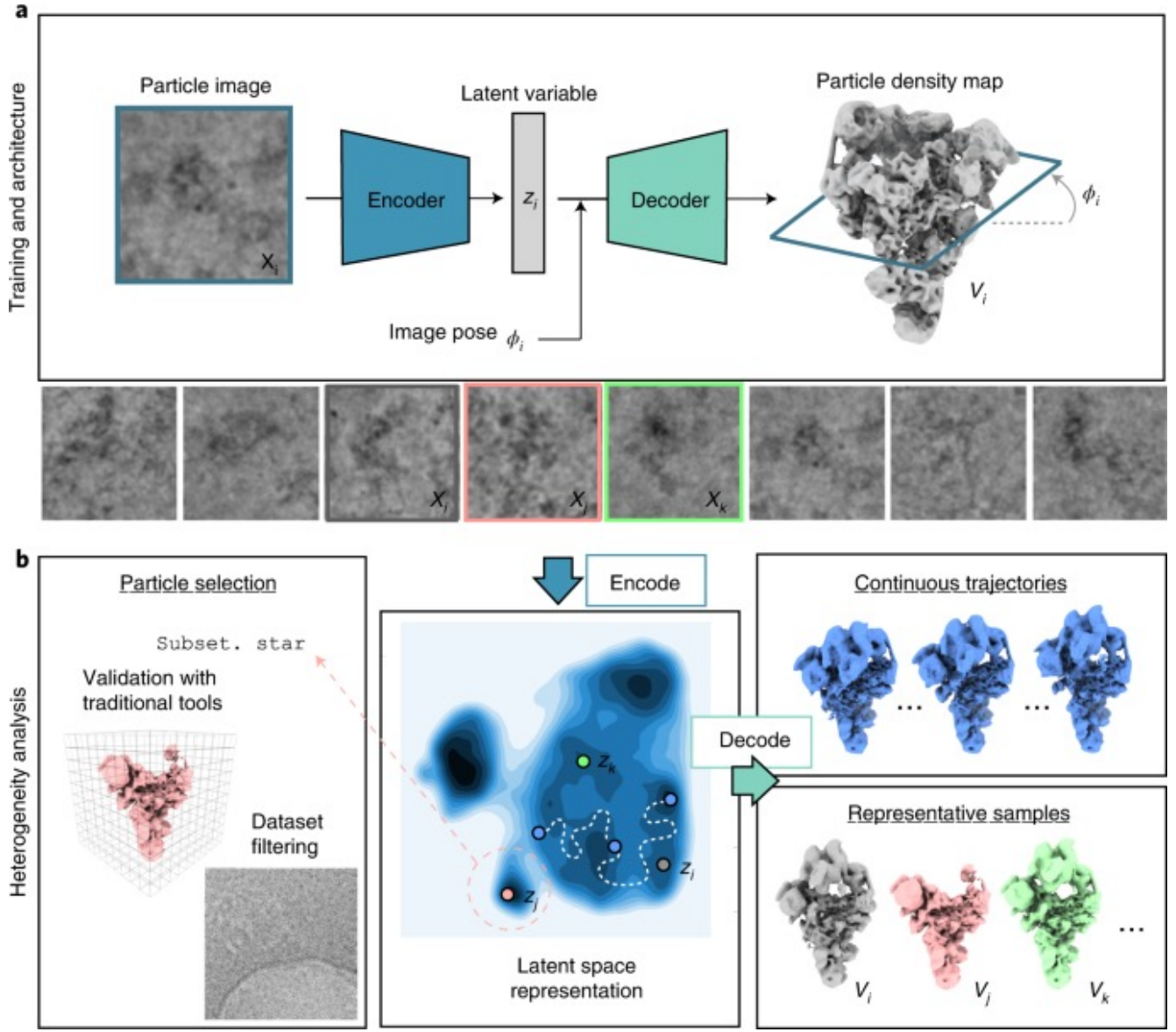
Learning low dimensional z from X (encoder), producing a generative model and infer X′ in a probabilistic manner (decoder)

Learning heterogeneity or the conformational landscape (z) from the consensus map (X) and infer the position in the landscape (X′)

Courtesy of Kwon sohee (from prof Seok Lap)

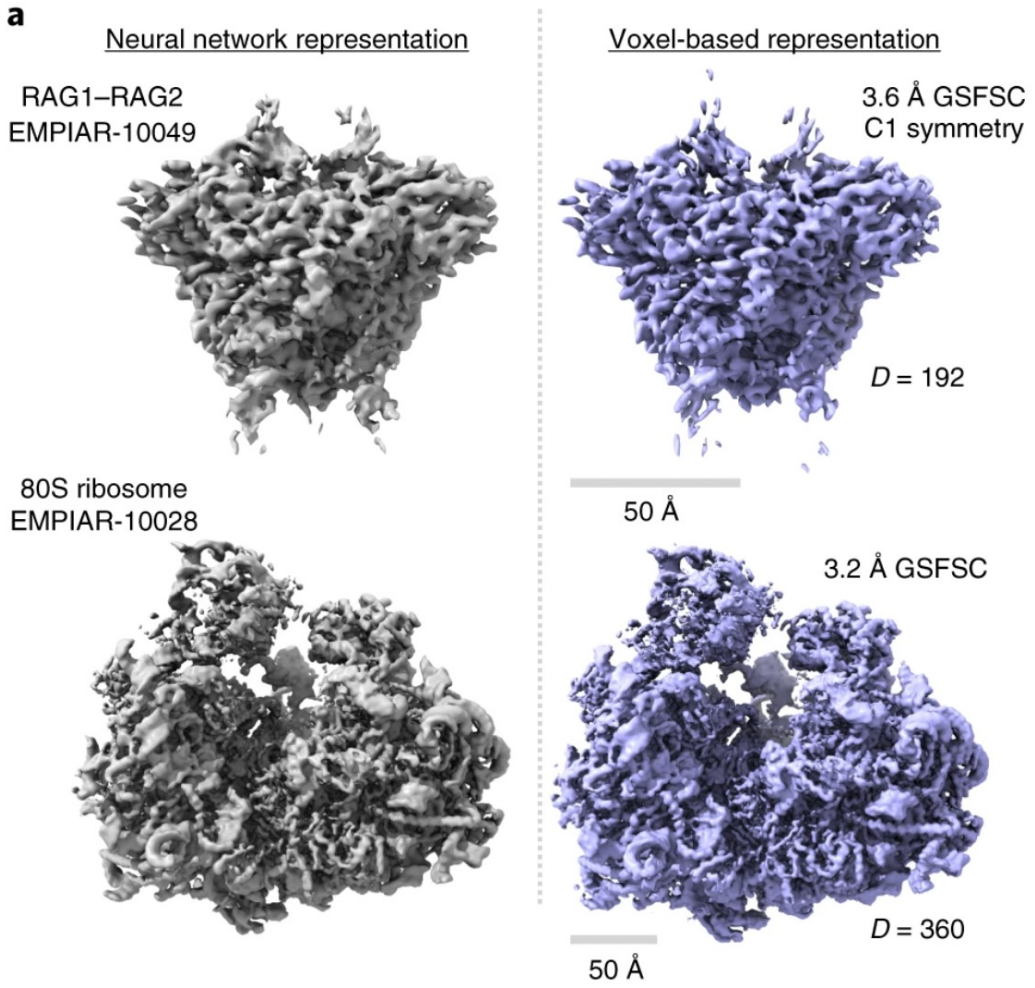
arXiv:1909.05215, 2020

After learning, use the encoder to evaluate the predicted latent z for each image, and use the decoder to generate 3D volume



Nature Methods **volume 18**, pages176–185 (2021)

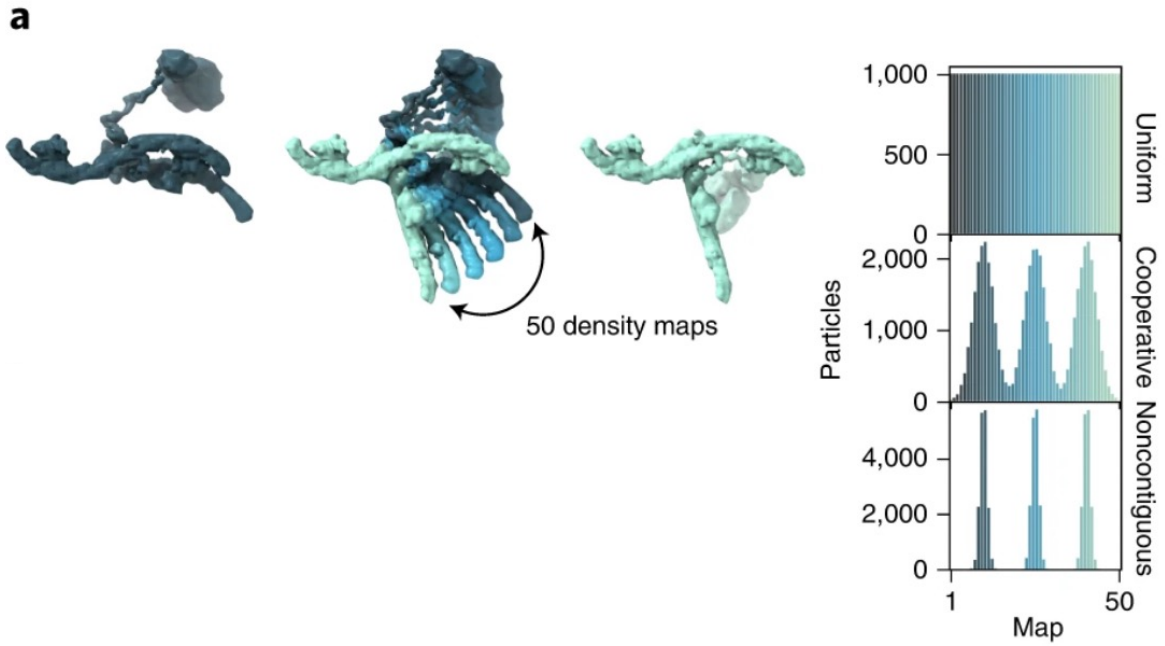
Generative models of decoder was capable of learning density maps from experimental dataset.

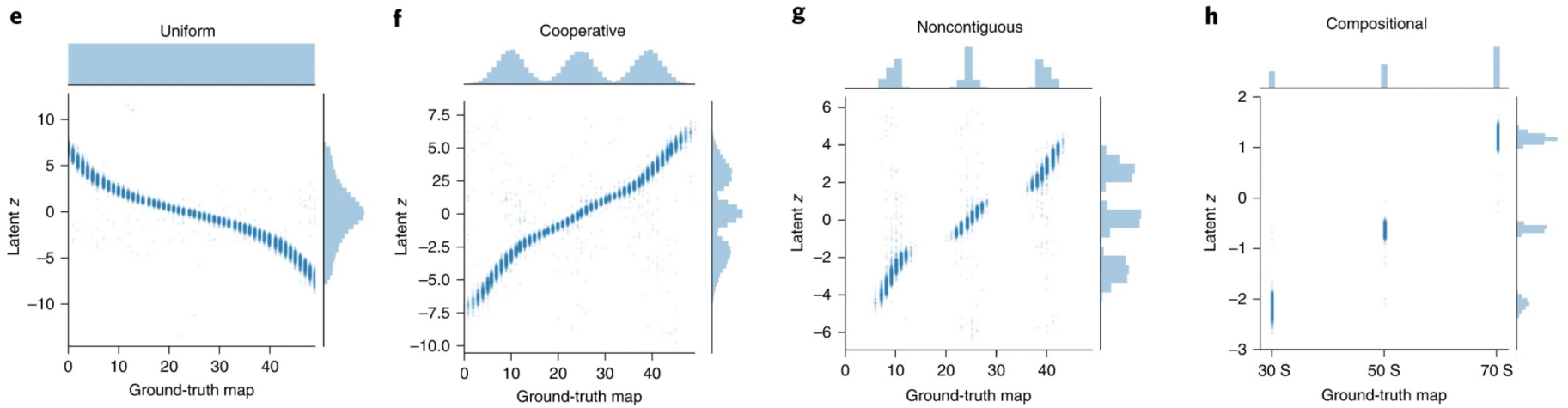


Architecture	Parameters	Training speed (min per 10^5 particles; 1 GPU)		
		$D = 128$	$D = 256$	$D = 360$
128 × 3	148,226	2.9	5.1	10.9
256 × 3	394,757	3.5	6.8	14.3
512 × 3	1,182,722	4.1	10.5	22.3
1,024 × 3	3,938,306	6.7	21.0	43.7
1,024 × 10	11,285,506	16.2	56.1	112.1

Nature Methods **volume 18**, pages176–185 (2021)

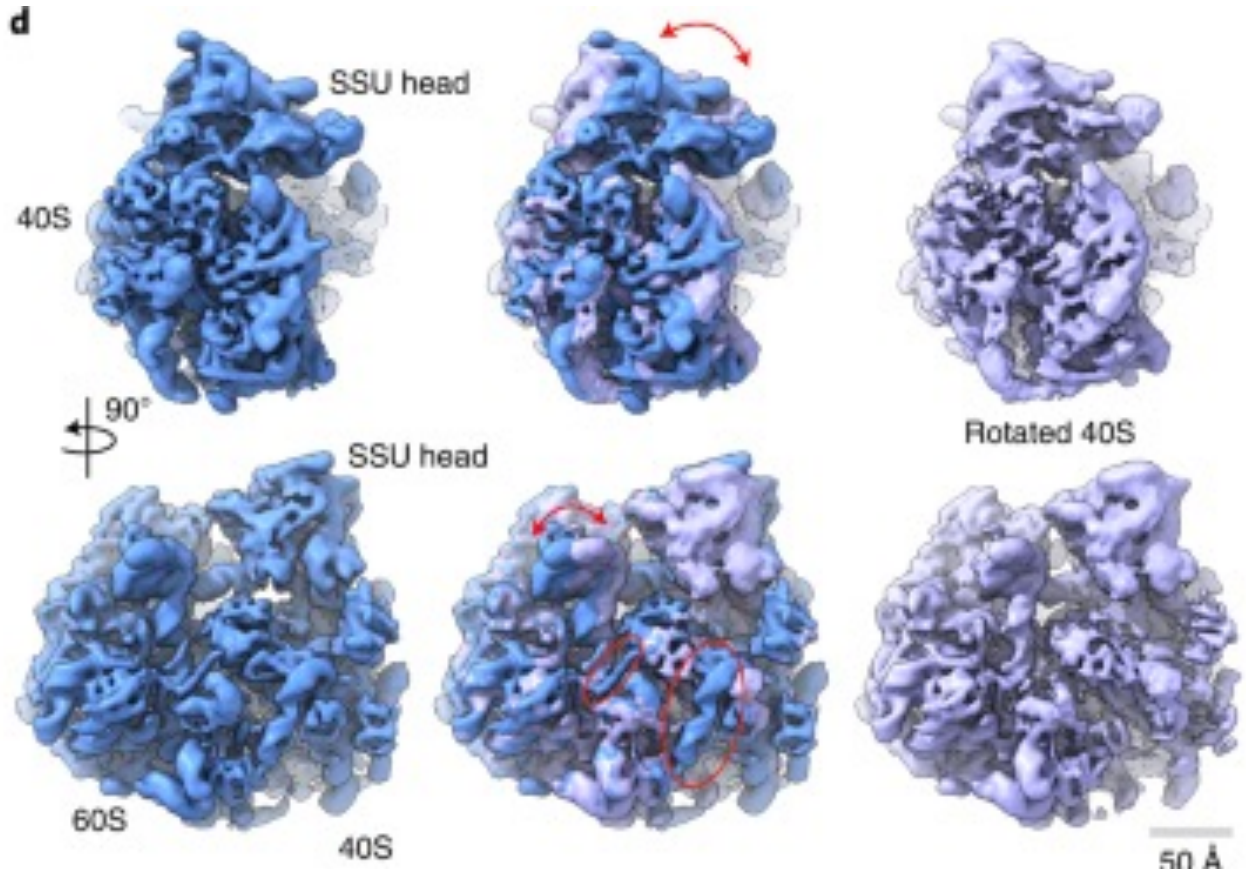
Generative models of encoder was capable of producing a latent space and reconstruction of the continuous heterogeniety.

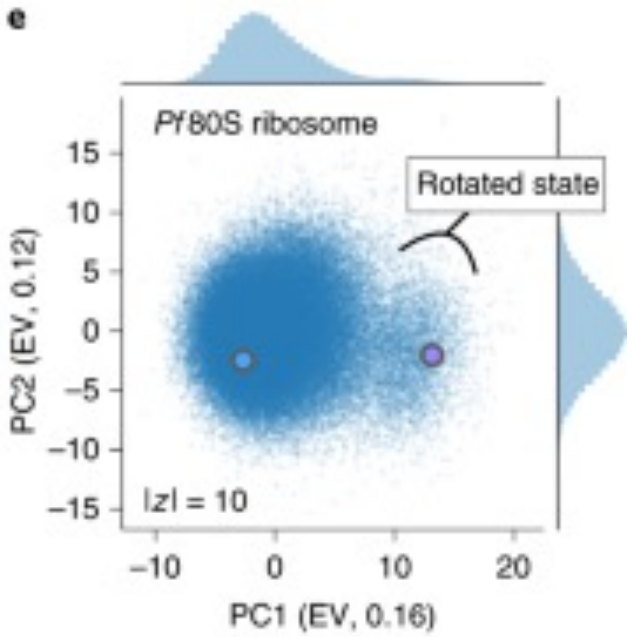




Nature Methods **volume 18**, pages176–185 (2021)

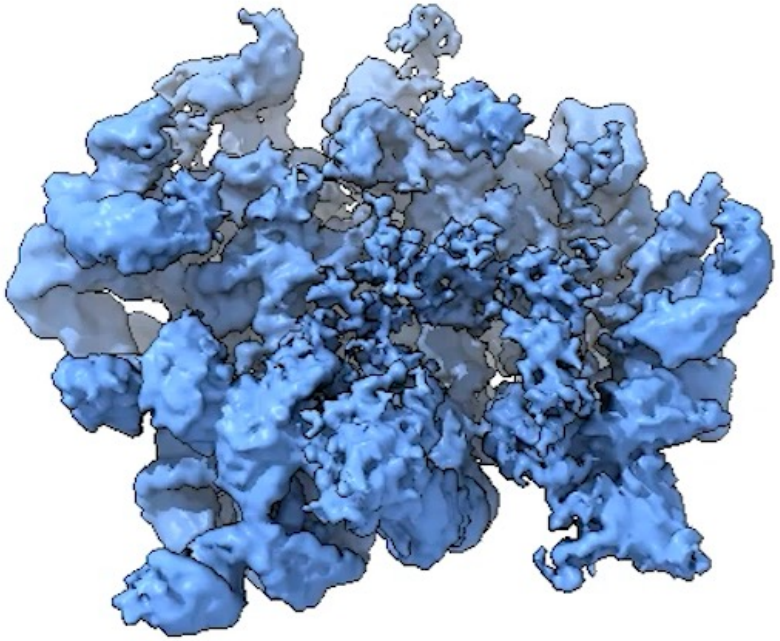
Cryodrgn discovered the innate heterogeneity in protein complex



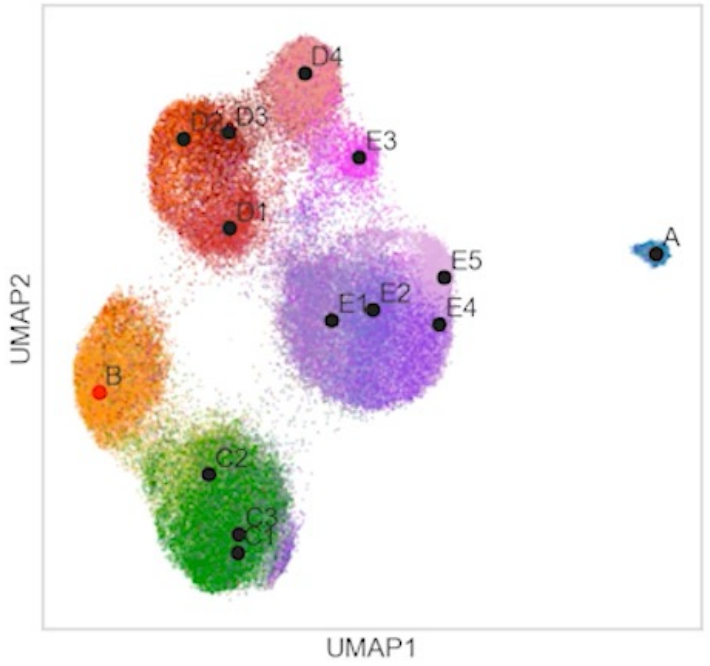


Nature Methods **volume 18**, pages176–185 (2021)

Cryodrgn discovered the innate heterogeneity in protein complex



Assembly path:
B→D1→D2→D3→D4→E3→E5



Nature Methods **volume 18**, pages176–185 (2021)

Many deep learning based programs are now trying to solve the challenges in structural biology

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, [...] Demis Hassabis

Nature **596**, 583–589 (2021) | Cite this article

392k Accesses | **2797** Altmetric | Metrics

nature > nature methods > articles > article

Article | Published: 07 October 2019

Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble & Bonnie Berger

Nature Methods **16**, 1153–1160 (2019) | Cite this article

8172 Accesses | **71** Citations | **129** Altmetric | Metrics

This article has been updated

CORRECTED PROOF

Full-length *de novo* protein structure determination from cryo-EM maps using deep learning

Jiahua He, Sheng-You Huang

Bioinformatics, btab357, https://doi.org/10.1093/bioinformatics/btab357

Published: 12 May 2021 **Article history ▾**

Article | Open Access | Published: 15 July 2021

DeepEMhancer: a deep learning solution for cryo-EM volume post-processing

Ruben Sanchez-Garcia, Josue Gomez-Blanco, Ana Cuervo, Jose Maria Carazo, Carlos Oscar S. Sorzano & Javier Vargas

Communications Biology **4**, Article number: 874 (2021) | Cite this article

1345 Accesses | **3** Citations | Metrics

arXiv:1909.05215, 2020

Summary

- **Cryodrgn** can learn the generative model from a continuous latent space to 3D volumes
- **Cryodrgn** uses encoder-decoder system; encoder can produce a latent space of protein particles; decoder can generate 3D structures of particles.
- **Cryodrgn** can discover inherent heterogeneity of protein structures

CryoDRGN v0.2 usage and workflow

Compute requirements:

- GPUs
- PyTorch https://pytorch.org/

Input requirements:

- Particle images
- High quality pose estimates, e.g. secondary structure features in consensus reconstruction
- Accurate CTF parameters

Total training times were recorded from training on a single Nvidia Tesla V100 32GB memory GPU card on either an Intel Xeon Gold 6130 CPU (2.10GHz, 791GB of RAM) or an IBM Power9 node with 1.2 TB of RAM.

https://cosmic-cryoem.org/tools/cryodrgn/