

CryoBench



Diverse and Challenging Datasets for the Heterogeneity Problem in Cryo-EM



Minkyu Jeon



Rishwanth Raghu



Miro Astore



Geoff Woollard



Ryan Feathers



Alkin Kaz



Sonya Hanson



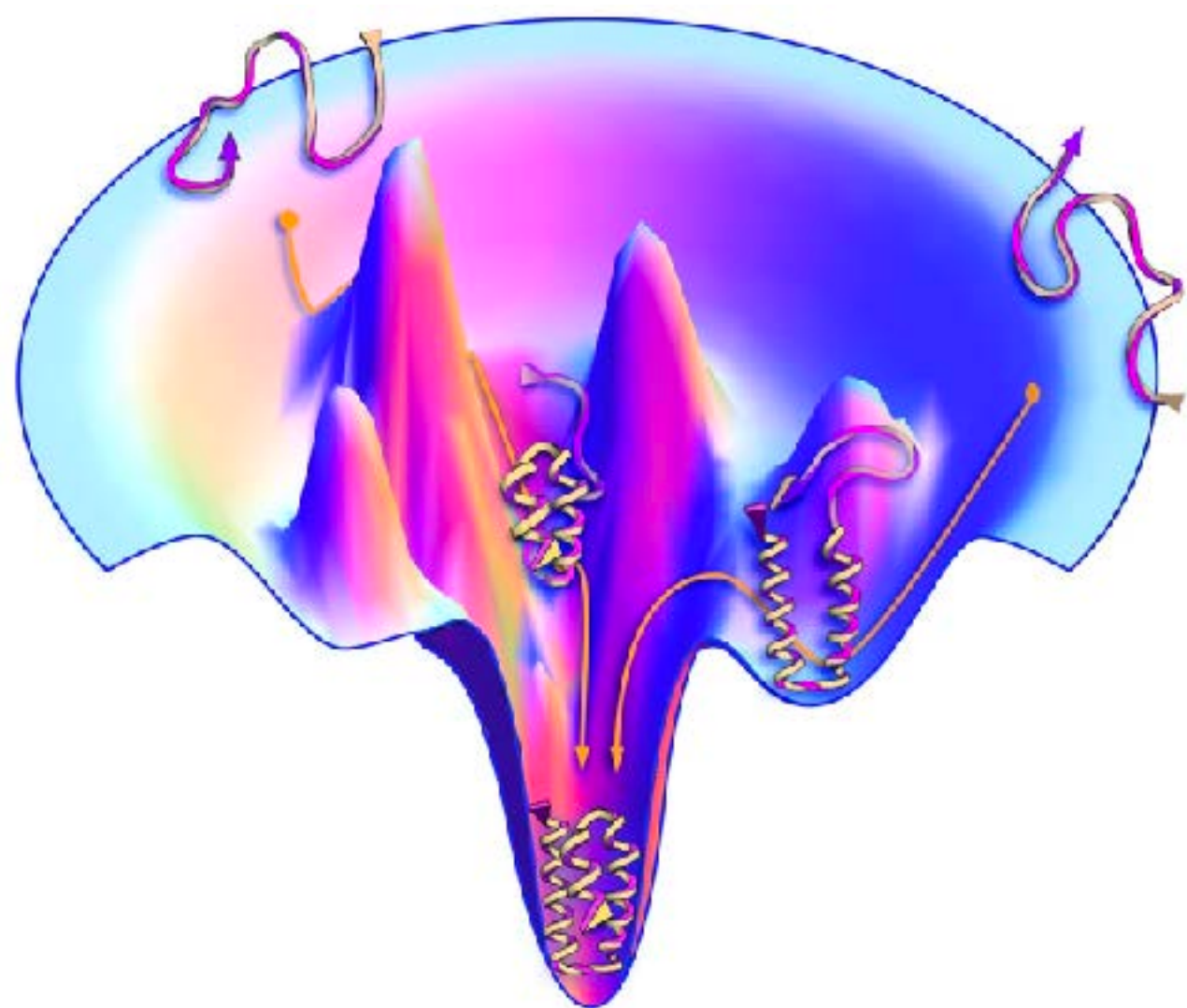
Pilar Cossio



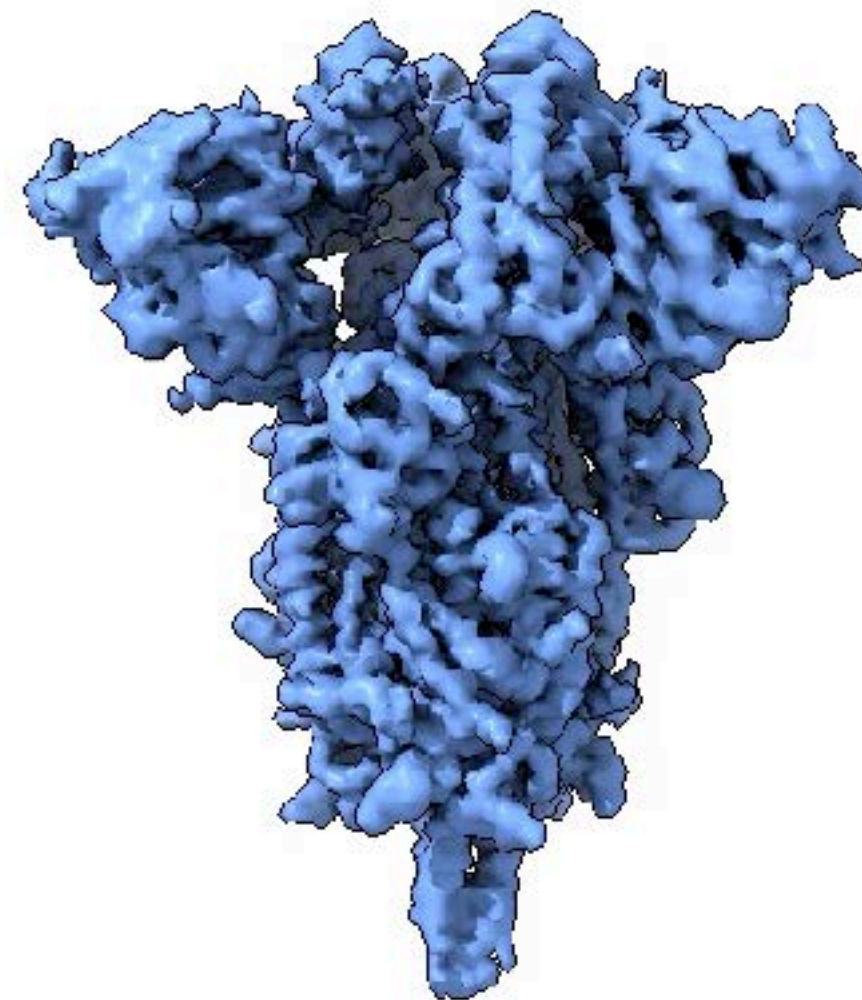
Ellen Zhong

What is biomolecular heterogeneity?

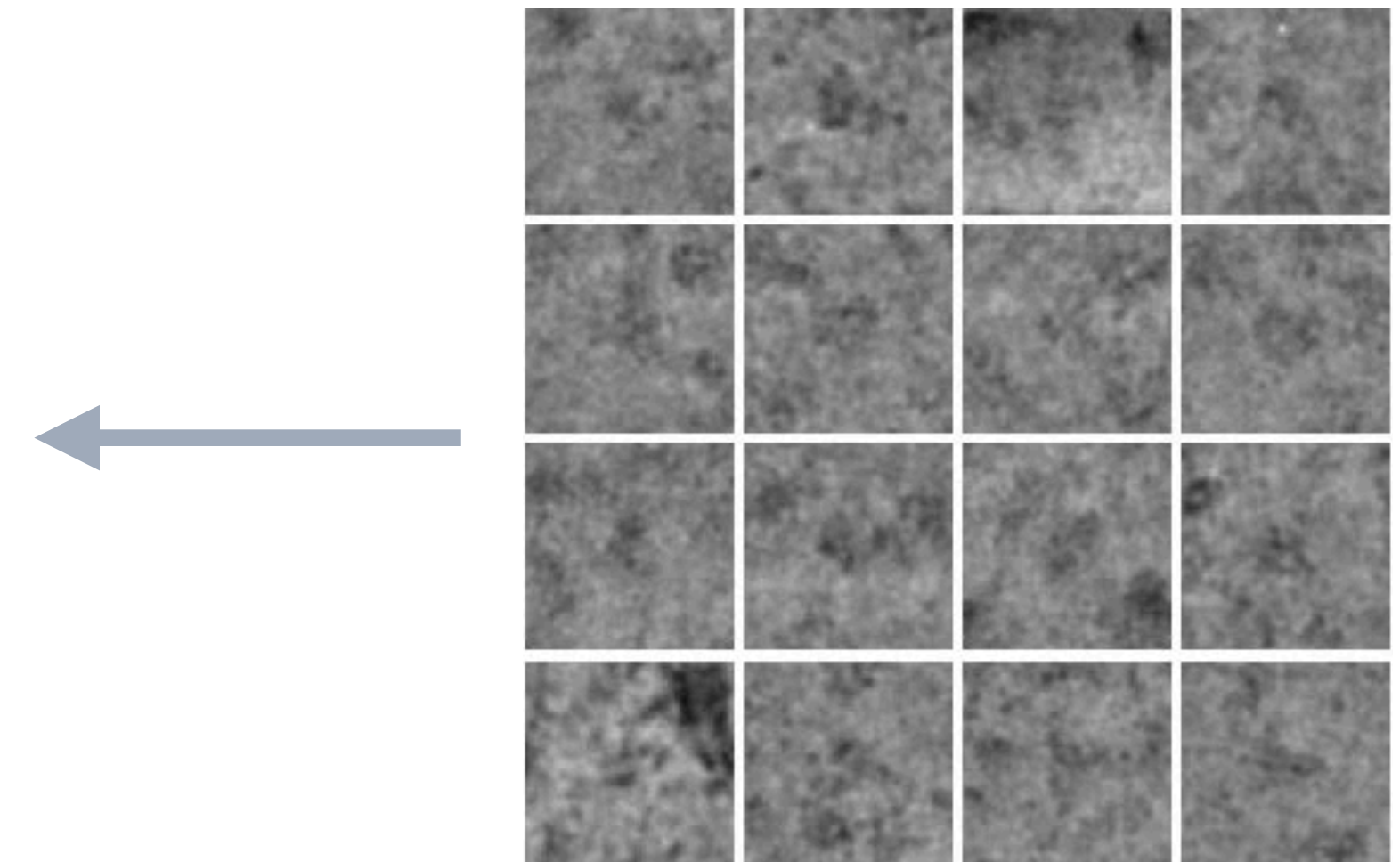
- Protein and other biomolecules form large, dynamic complexes that carry out essential biological functions.
- Existing tools for modeling structure, such as AlphaFold, are limited in their ability to predict different conformations or compositional states.
- **Cryo-electron microscopy (cryo-EM)**, in contrast, is a technique providing a unique opportunity to study biomolecules in near-native conformational states *from experimented data*.



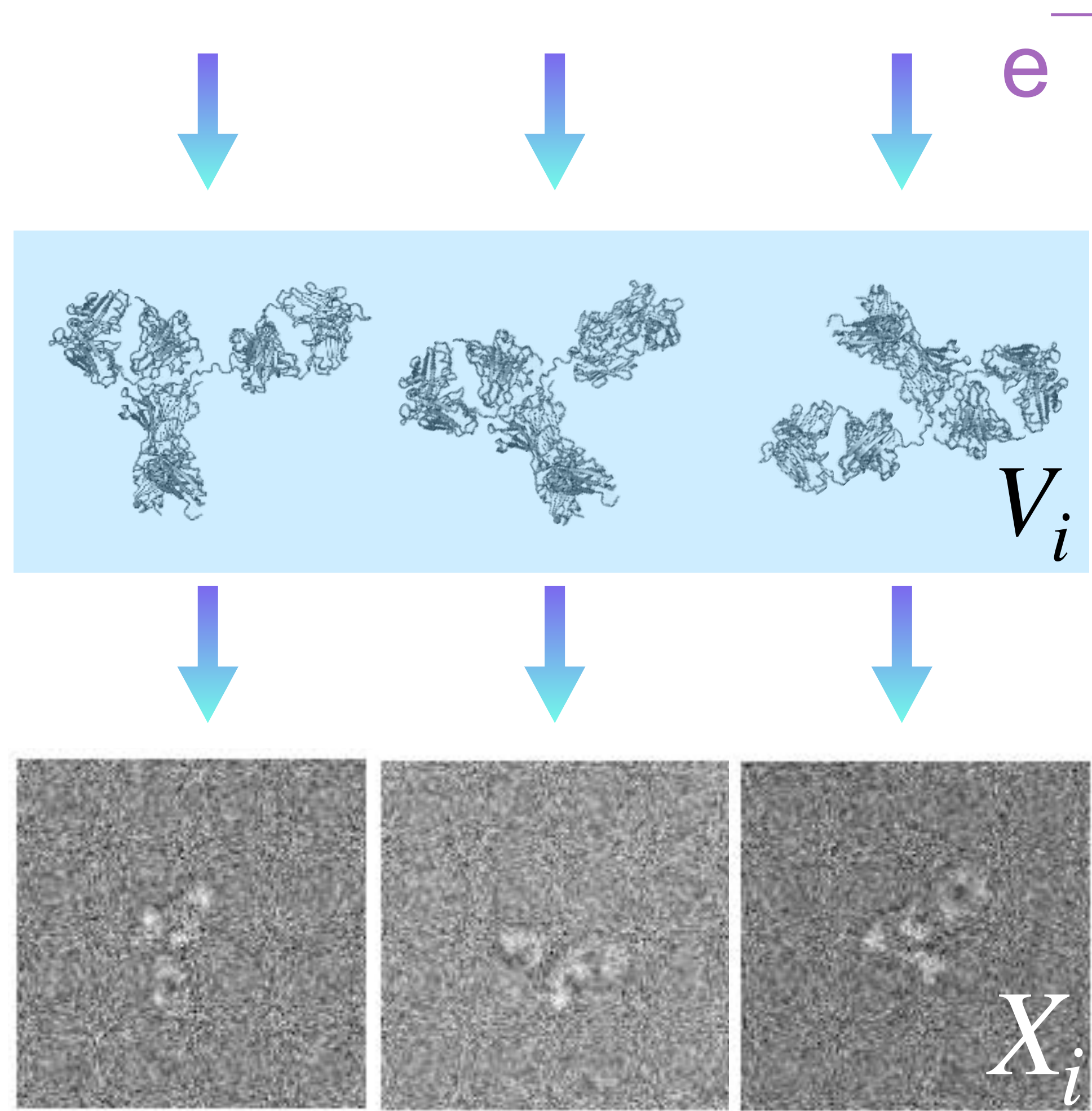
Dill & Maccallum, Science 2012



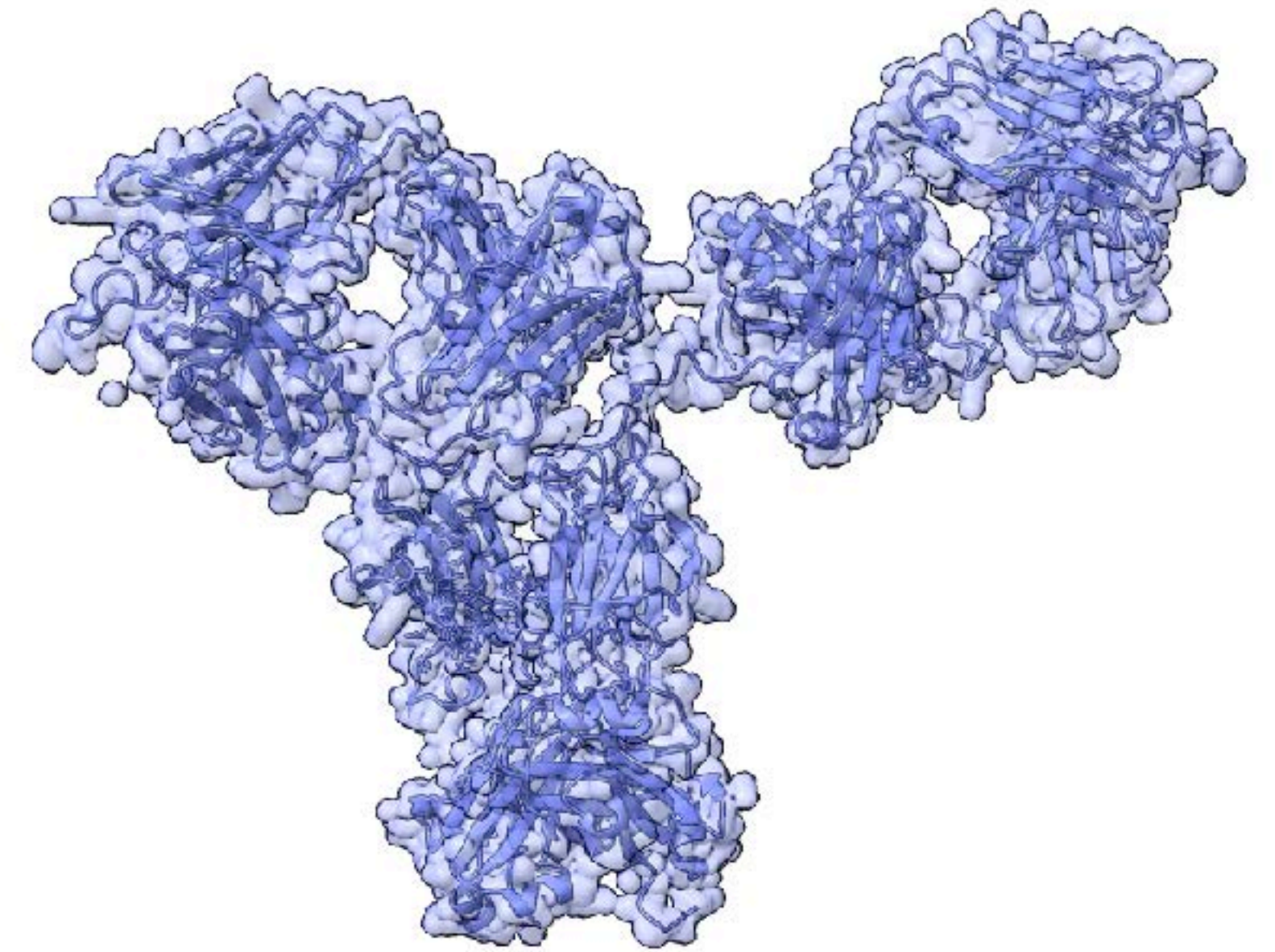
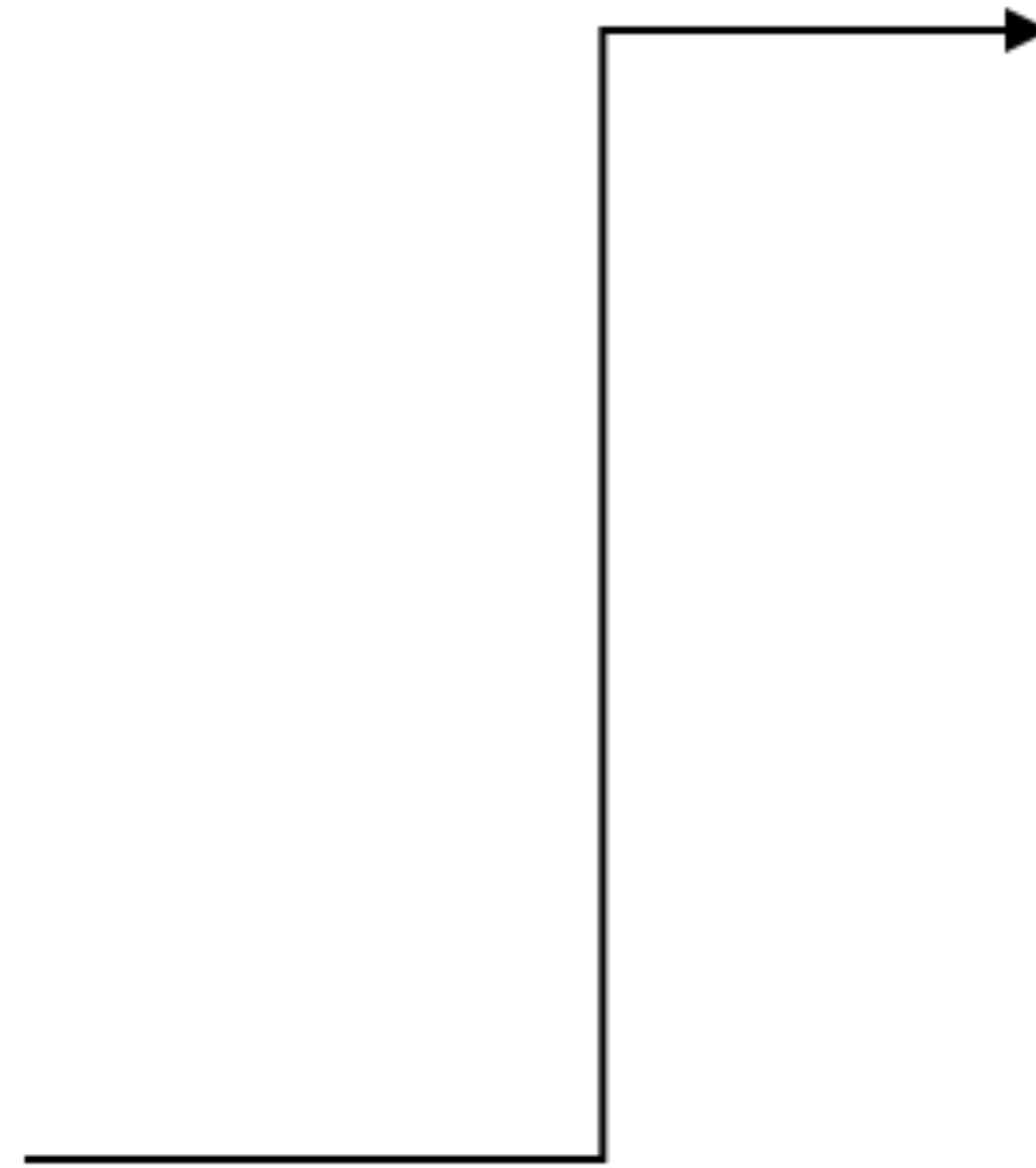
SARS-CoV-2 Spike Protein
Dataset: Walls et al 2020



Cryo-EM and 3D reconstruction



$10^4 - 10^7$ images
SNR ≈ 1 to 5%



$$V : \mathbb{R}^3 \rightarrow \mathbb{R}$$

Electron Scattering Potential

Reconstruction as an inference problem

Image Formation Model

$$Y_i = C_i * P_{\phi_i} V_i + \eta_i$$

CTF Pose Gaussian noise

Projection

The diagram shows the equation $Y_i = C_i * P_{\phi_i} V_i + \eta_i$ enclosed in a rounded rectangle. Above the rectangle is the text 'Image Formation Model'. Below the rectangle, vertical lines connect terms to labels: C_i to 'CTF', P_{ϕ_i} to 'Pose', and η_i to 'Gaussian noise'. A horizontal line spans the width of the rectangle, with the label 'Projection' centered below it.

$$\phi_i \in \text{SO}(3) \times \mathbb{R}^2$$
$$\eta_i \sim \mathcal{N}$$

Goal: Estimate V and poses $\{\phi_i\}$ typically with maximum likelihood techniques

Motivation

1. Reconstruction of molecular movies is now possible

3DVA



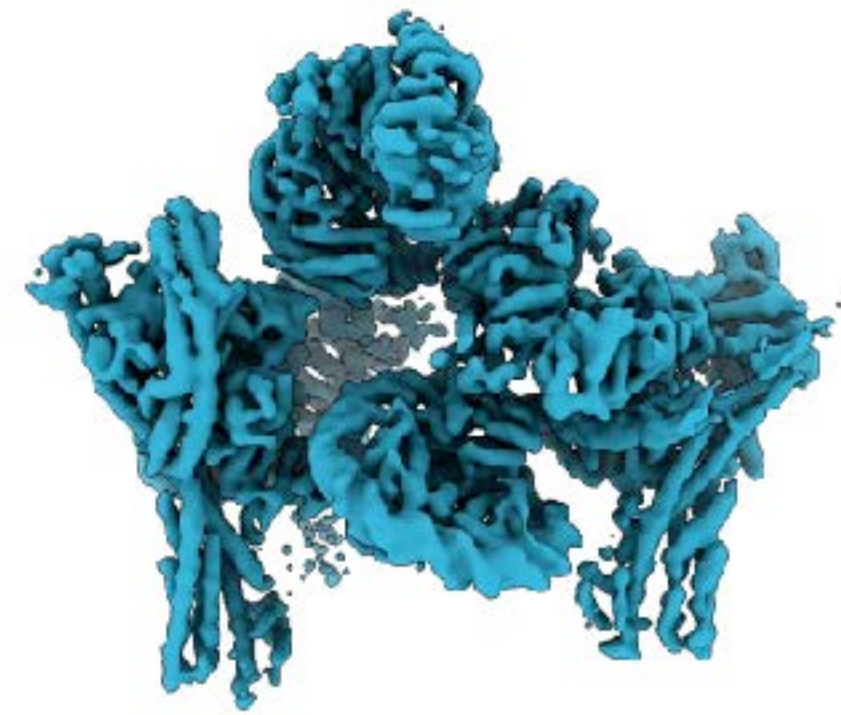
Na_v 1.7 ion channel
[EMPIAR-10261]

3DFlex



$\alpha V\beta 8$ integrin
[EMPIAR-10345]

DynaMight



Inner kinetochore
[EMPIAR-11910]

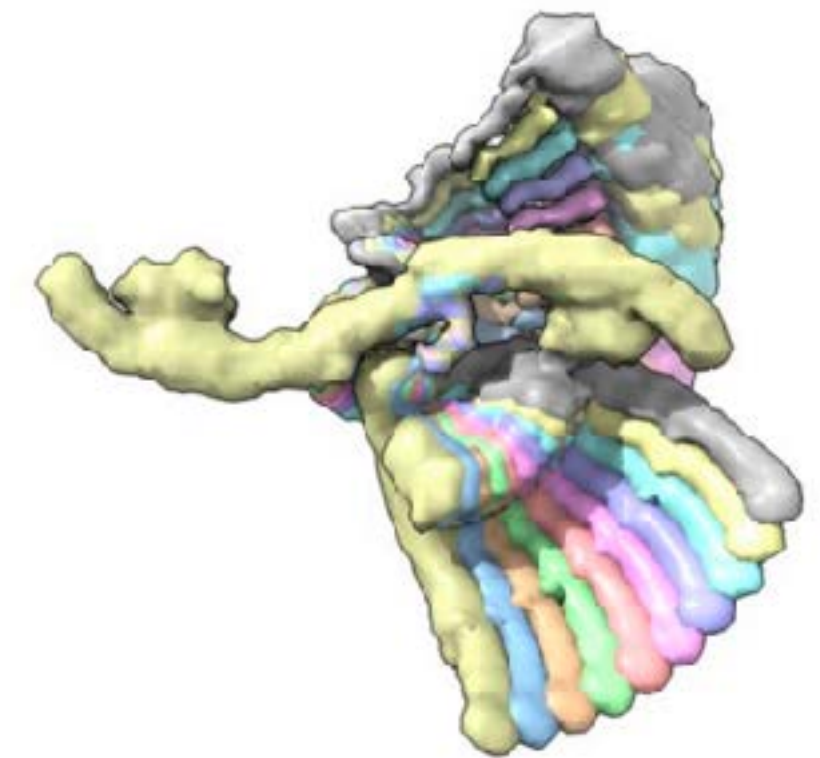
CryoDRGN



Pre-catalytic spliceosome
[EMPIAR-10180]

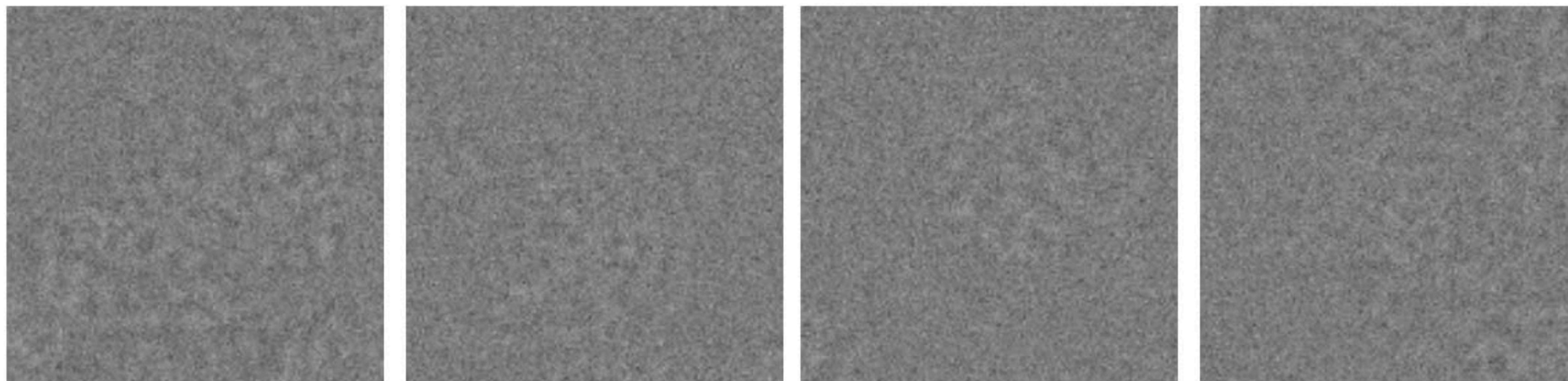
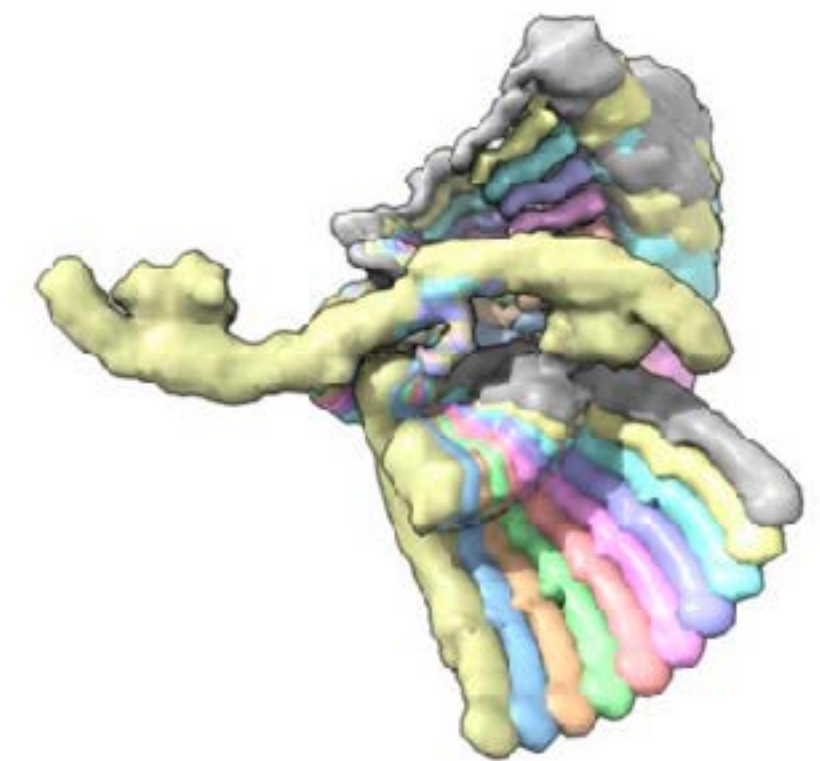
Motivation

1. Reconstruction of molecular movies is now possible
2. Methods often use simple toy motions for validation and comparison with other approaches



Motivation

1. Reconstruction of molecular movies is now possible
2. Methods often use simple toy motions for validation and comparison with other approaches
3. **No ground truth** exists for real data; Evaluation currently requires **benchmarking-by-eye**



CryoBench : Contributions

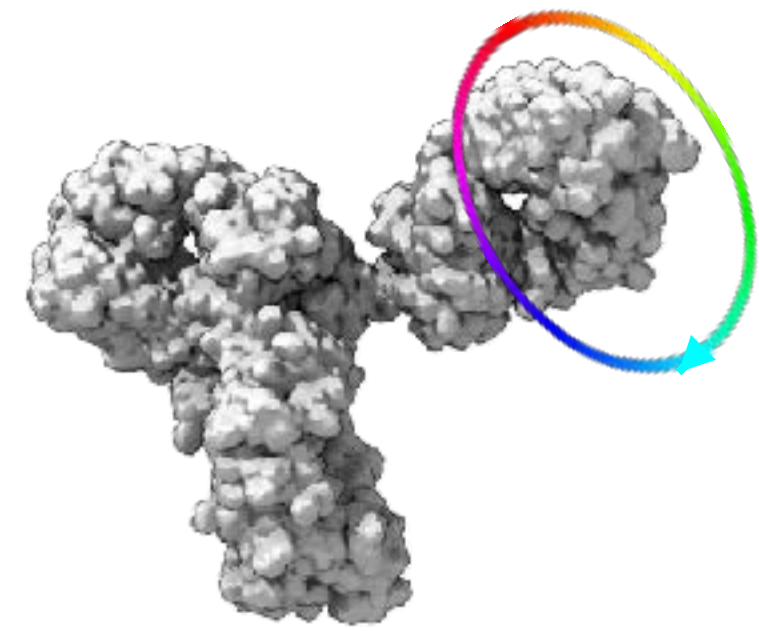
1. Design new **synthetic datasets** with challenging forms of heterogeneity to motivate new tasks and methods development
2. Introduce **metrics** for quantitative comparison of methods for heterogeneity reconstruction
3. **Benchmark** existing state-of-the-art methods

CryoBench ❄️🪑: Datasets

Conformational Heterogeneity

Compositional Heterogeneity

Diagnostic

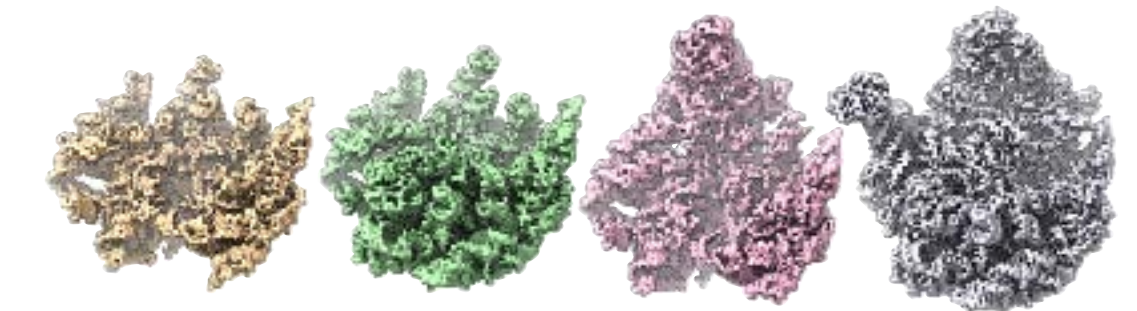
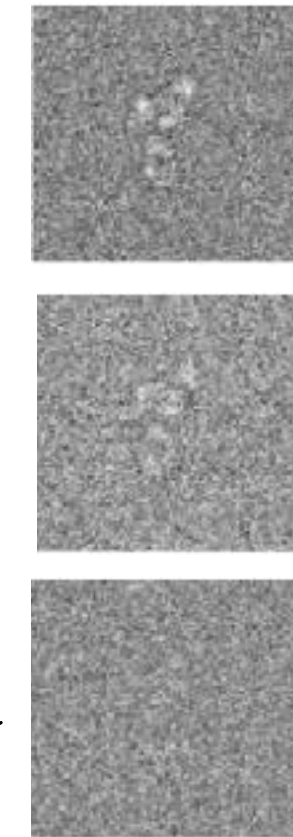


IgG-1D

IgG-1D

IgG-1D-noisier

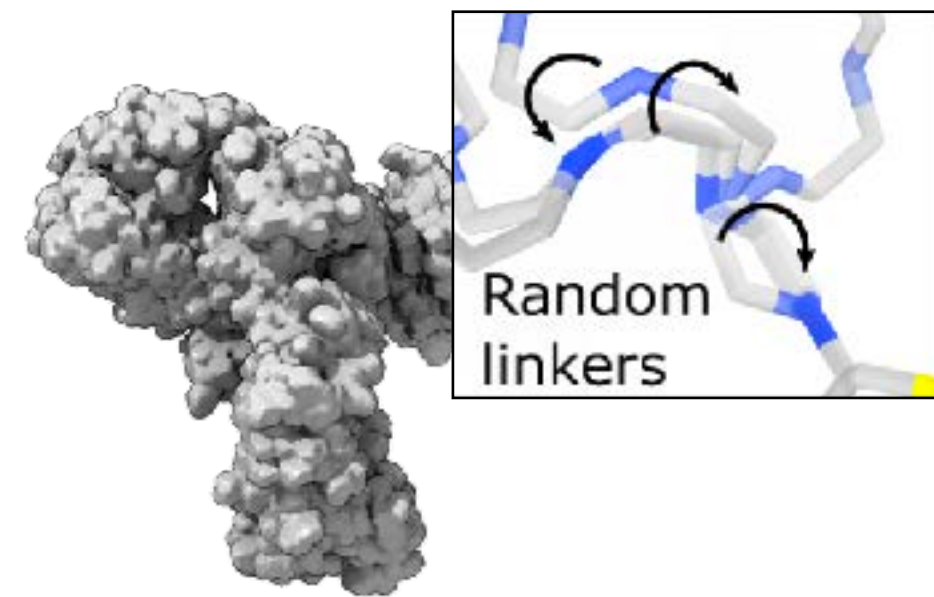
IgG-1D-noisiest



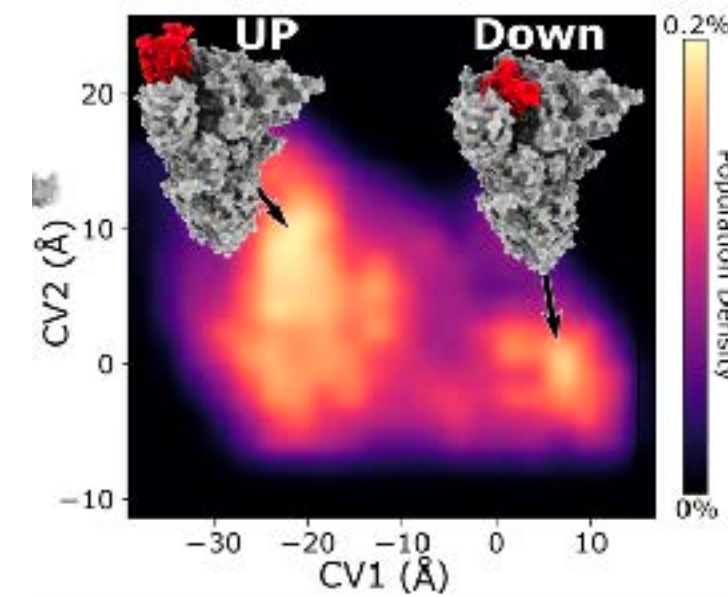
16 structures

Ribosome

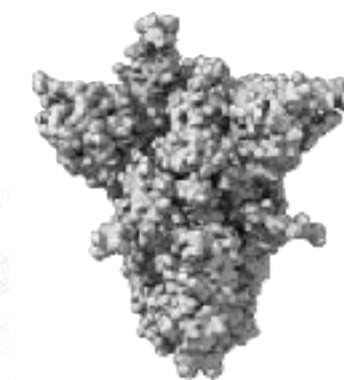
Challenging



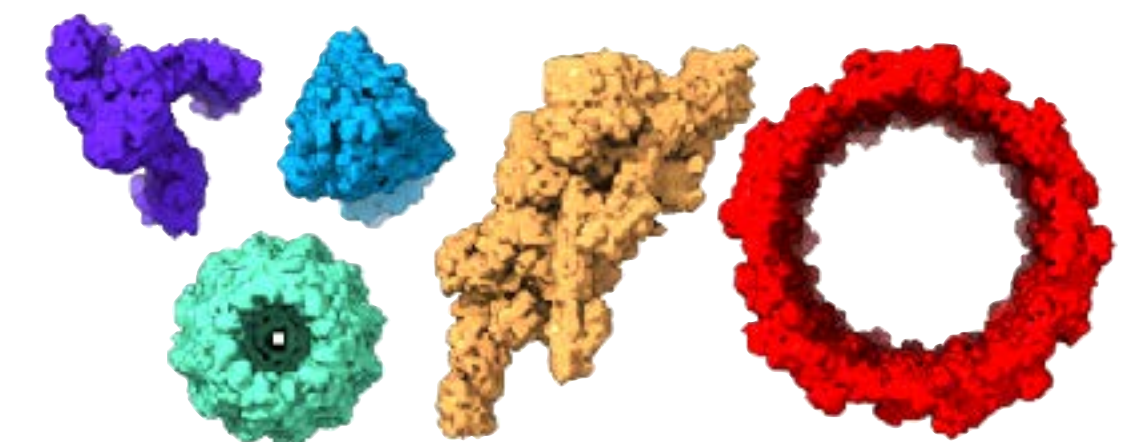
IgG-RL



Spike-MD



~46k structures



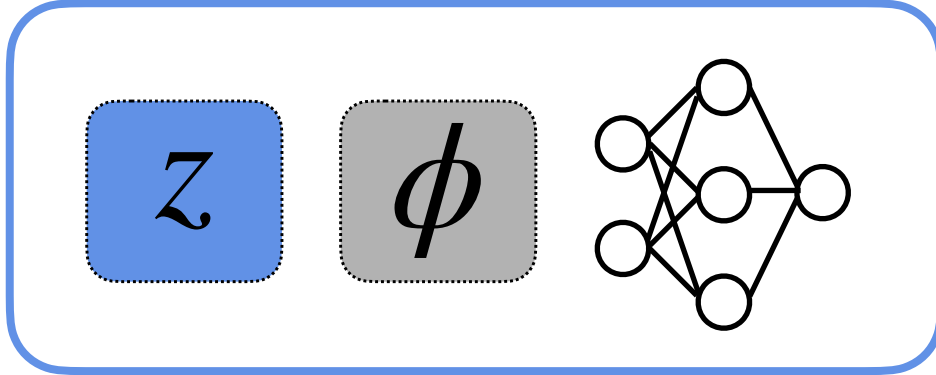
100 structures

Tomotwin-100

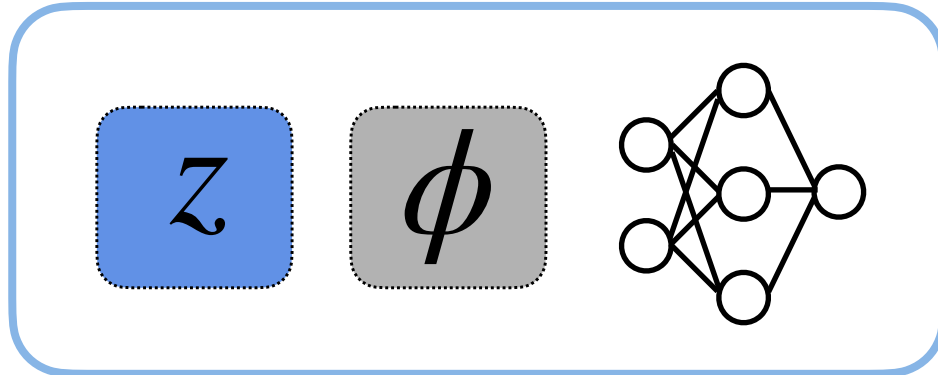
Methods

Neural

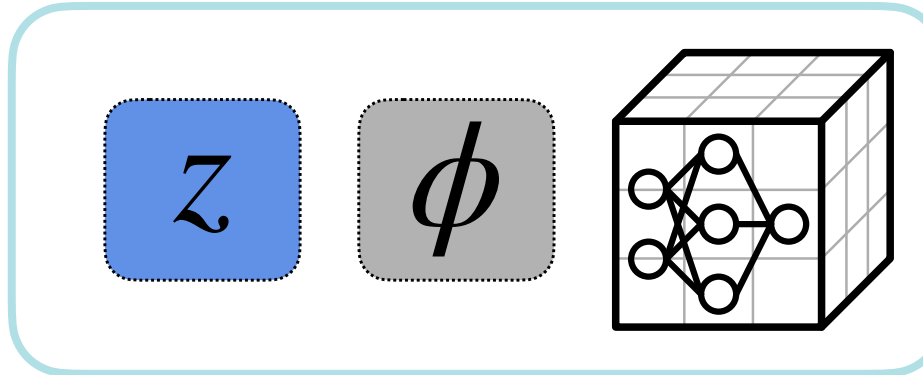
CryoDRGN



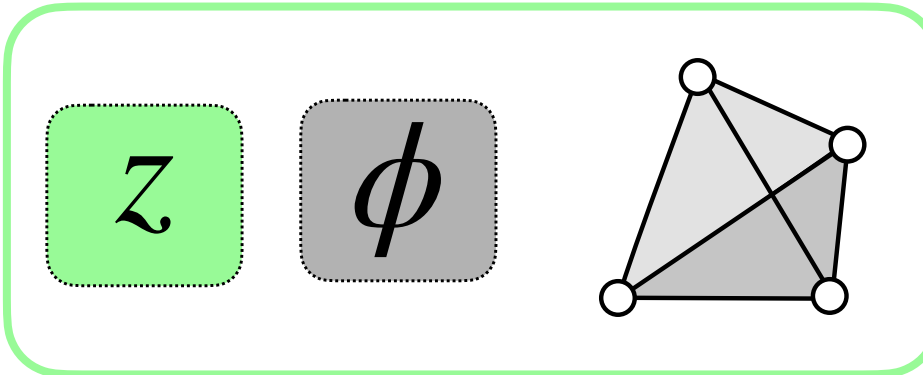
DRGN-AI-fixed



Opus-DSD

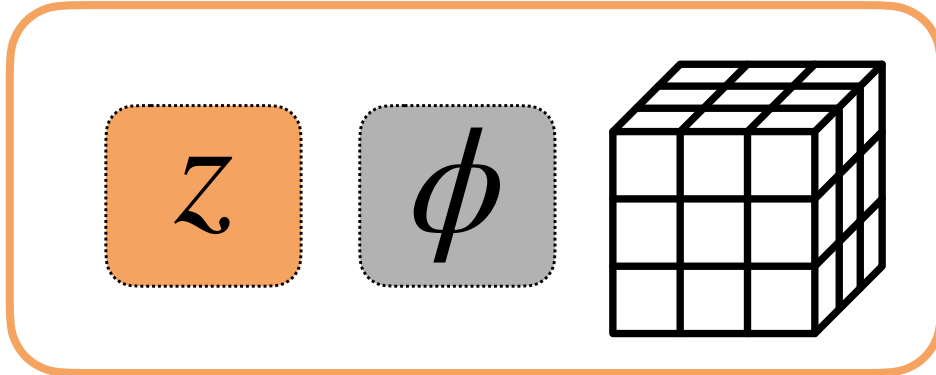


3DFlex

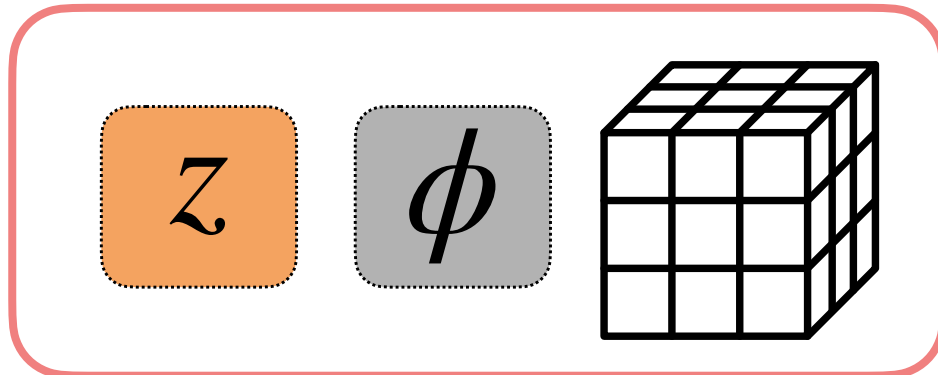


Linear

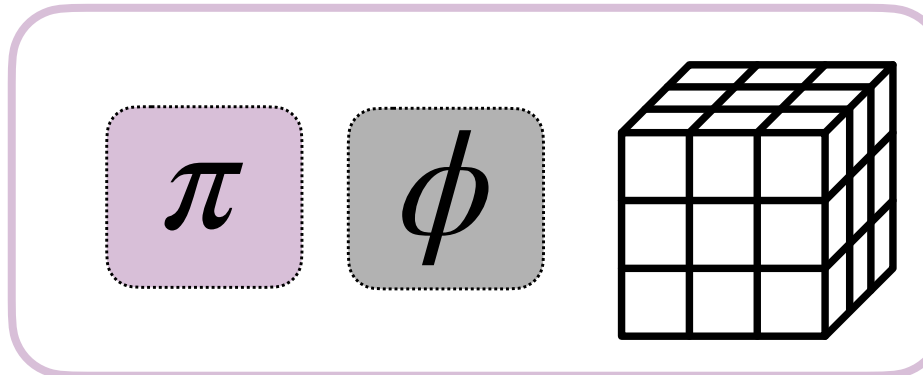
3DVA



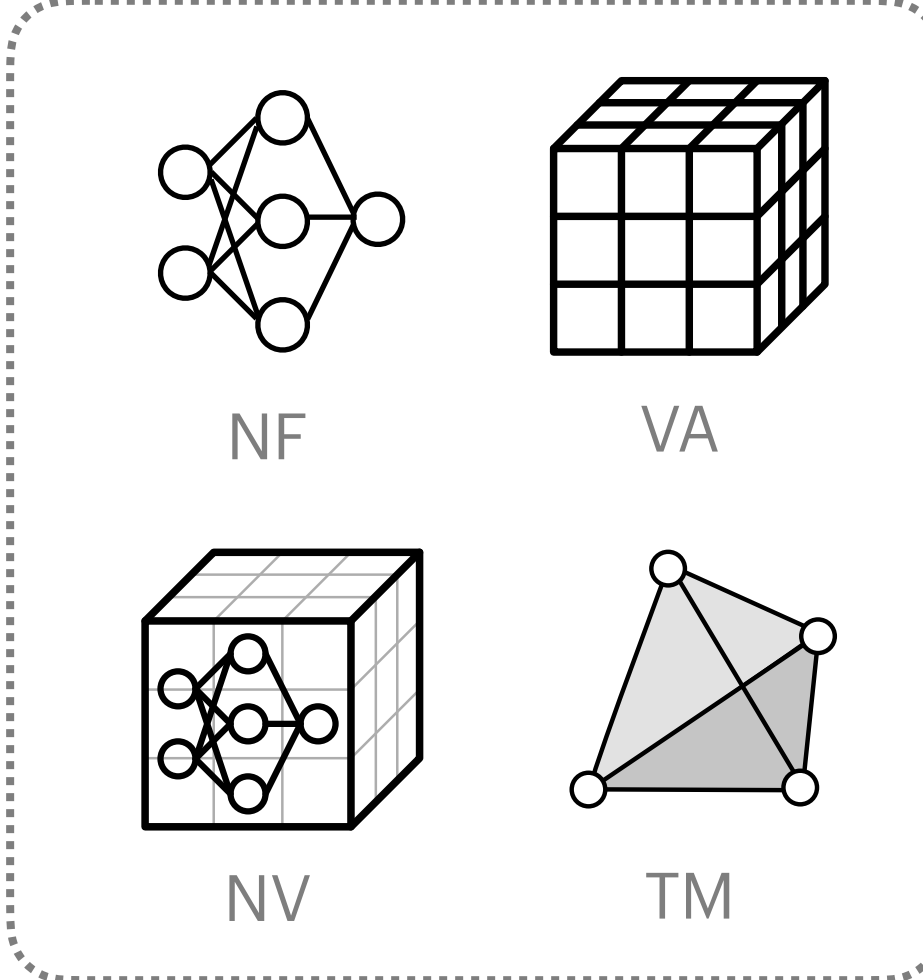
RECOVER



3D Class

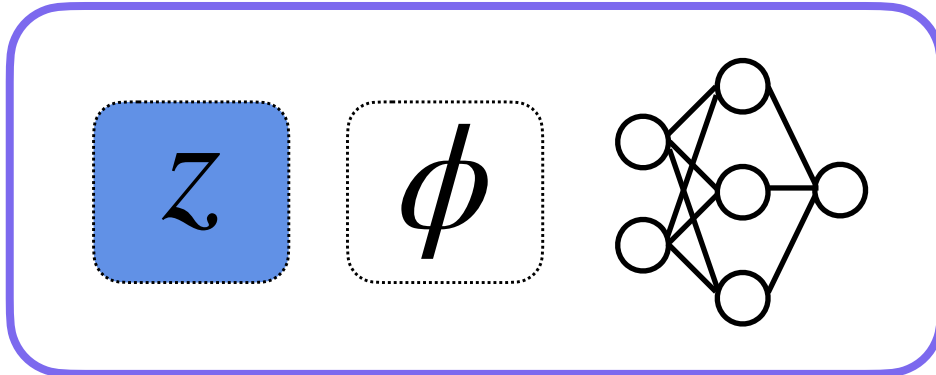


Volume Parameterization

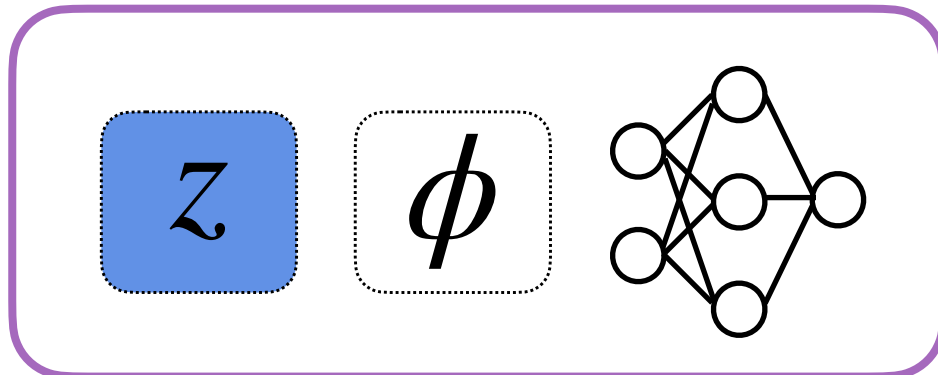


Ab initio

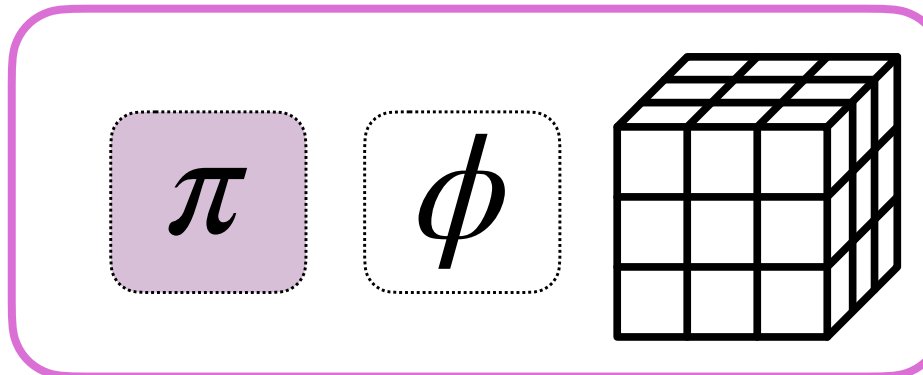
CryoDRGN2



DRGN-AI



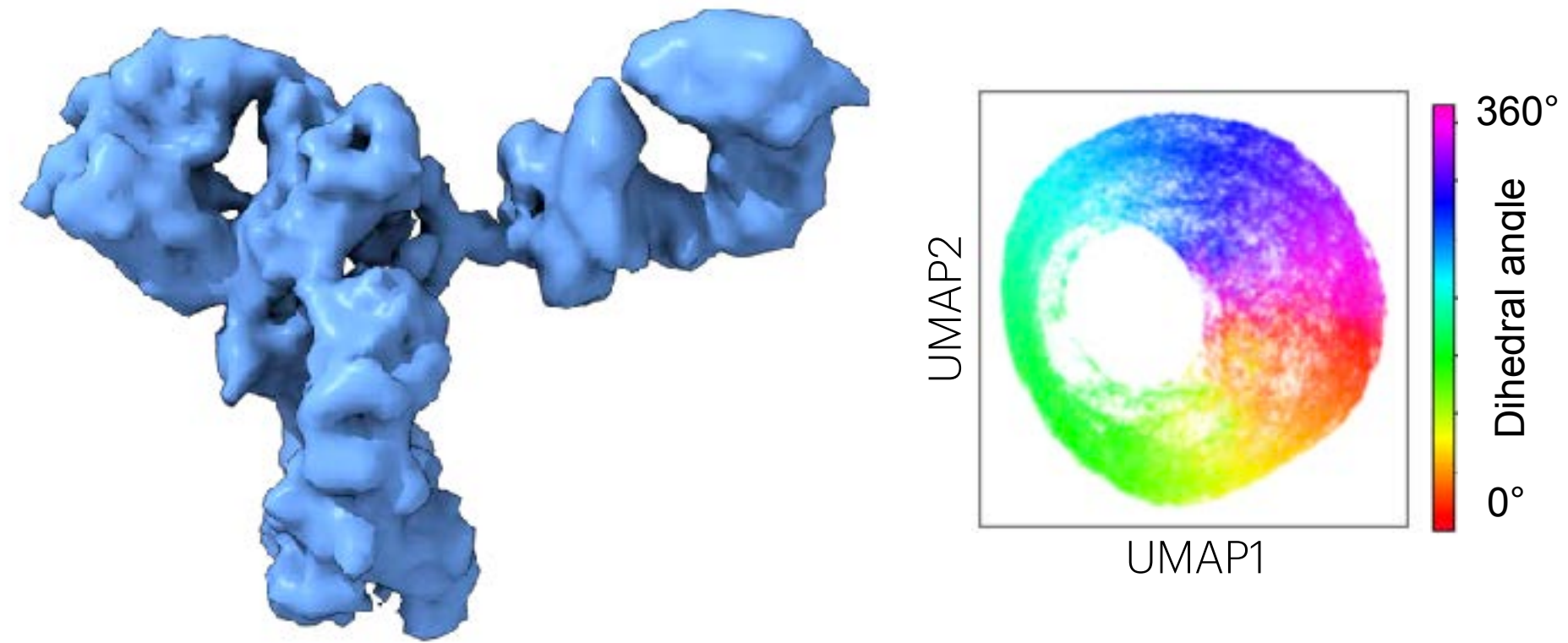
3D Class abinit



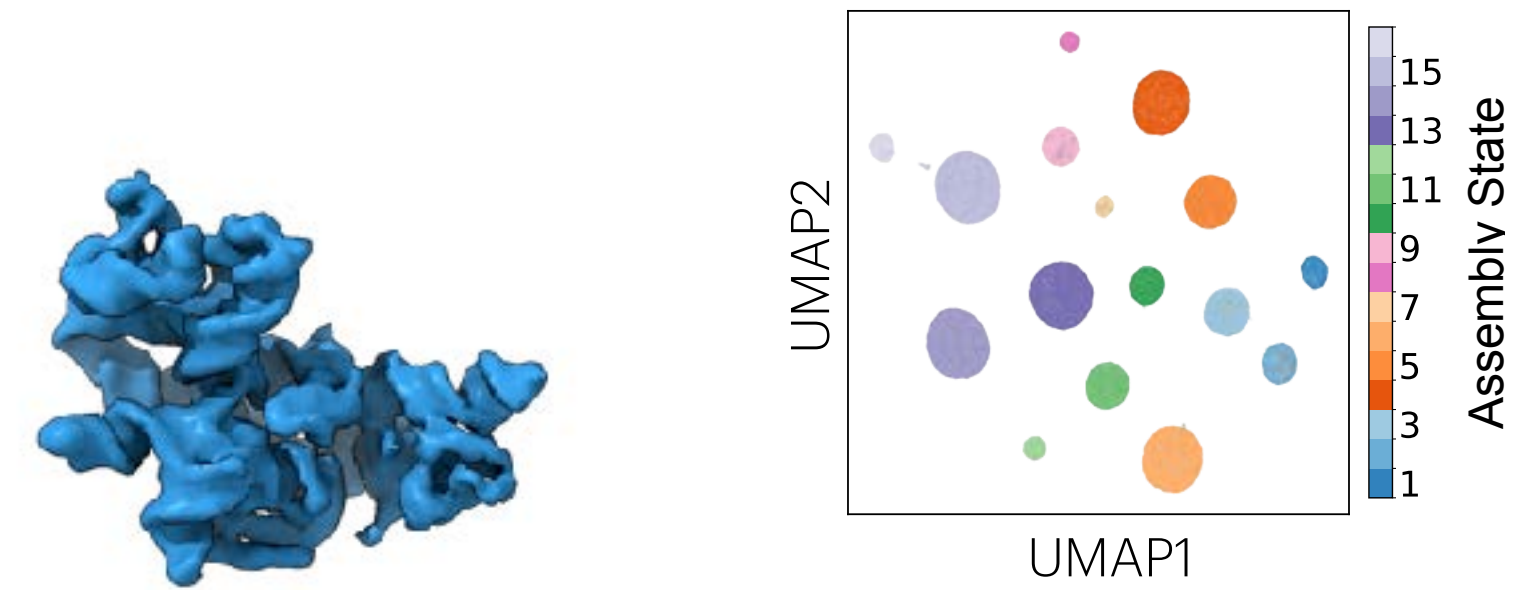
Discrete

Qualitative Results

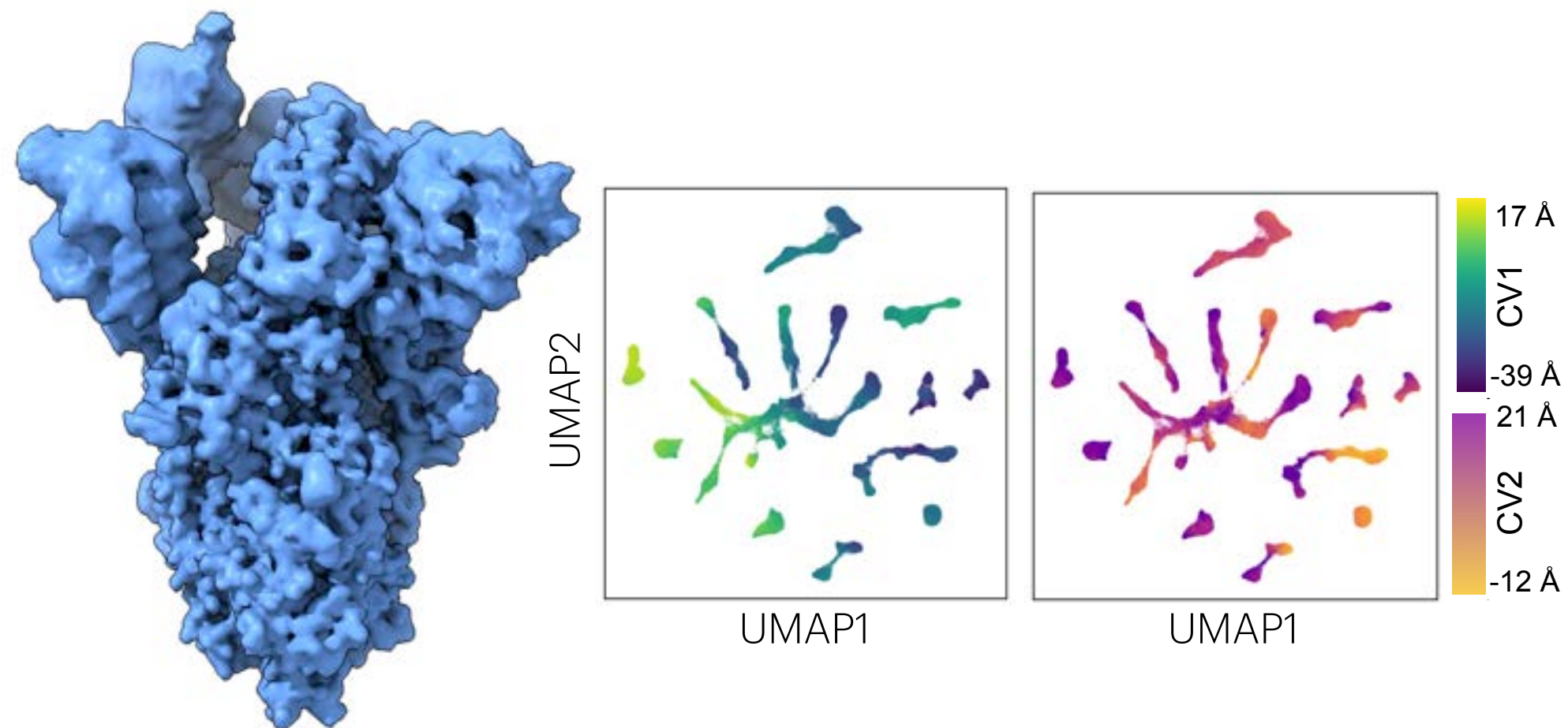
IgG-1D



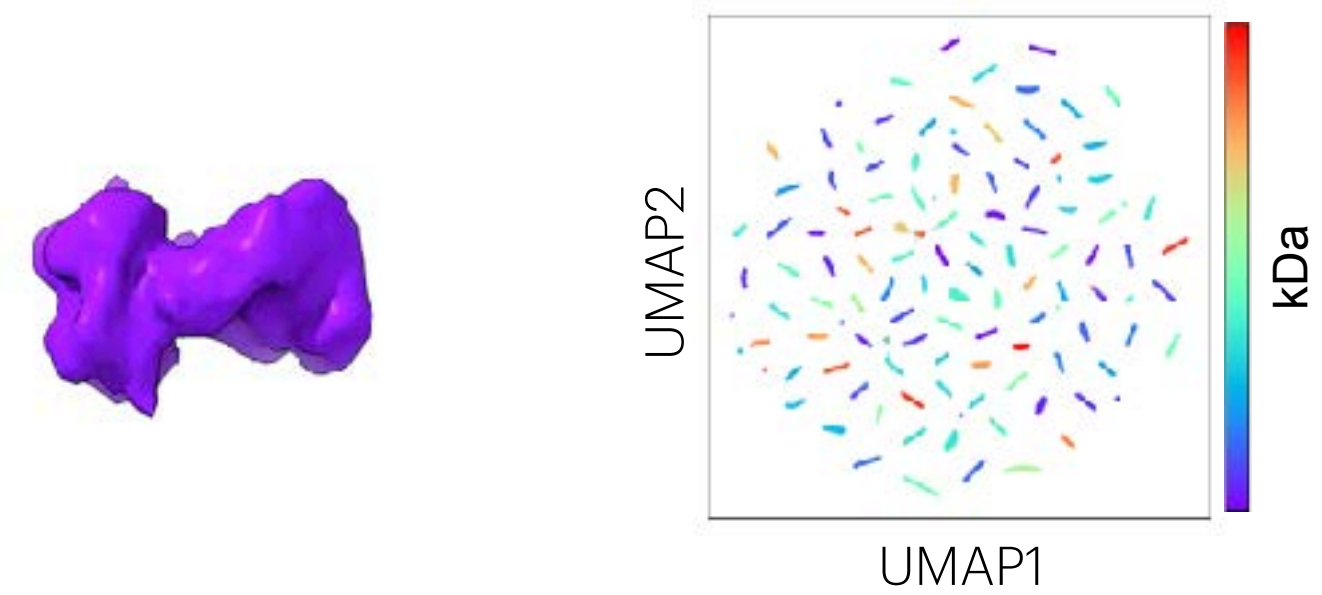
Ribosembly



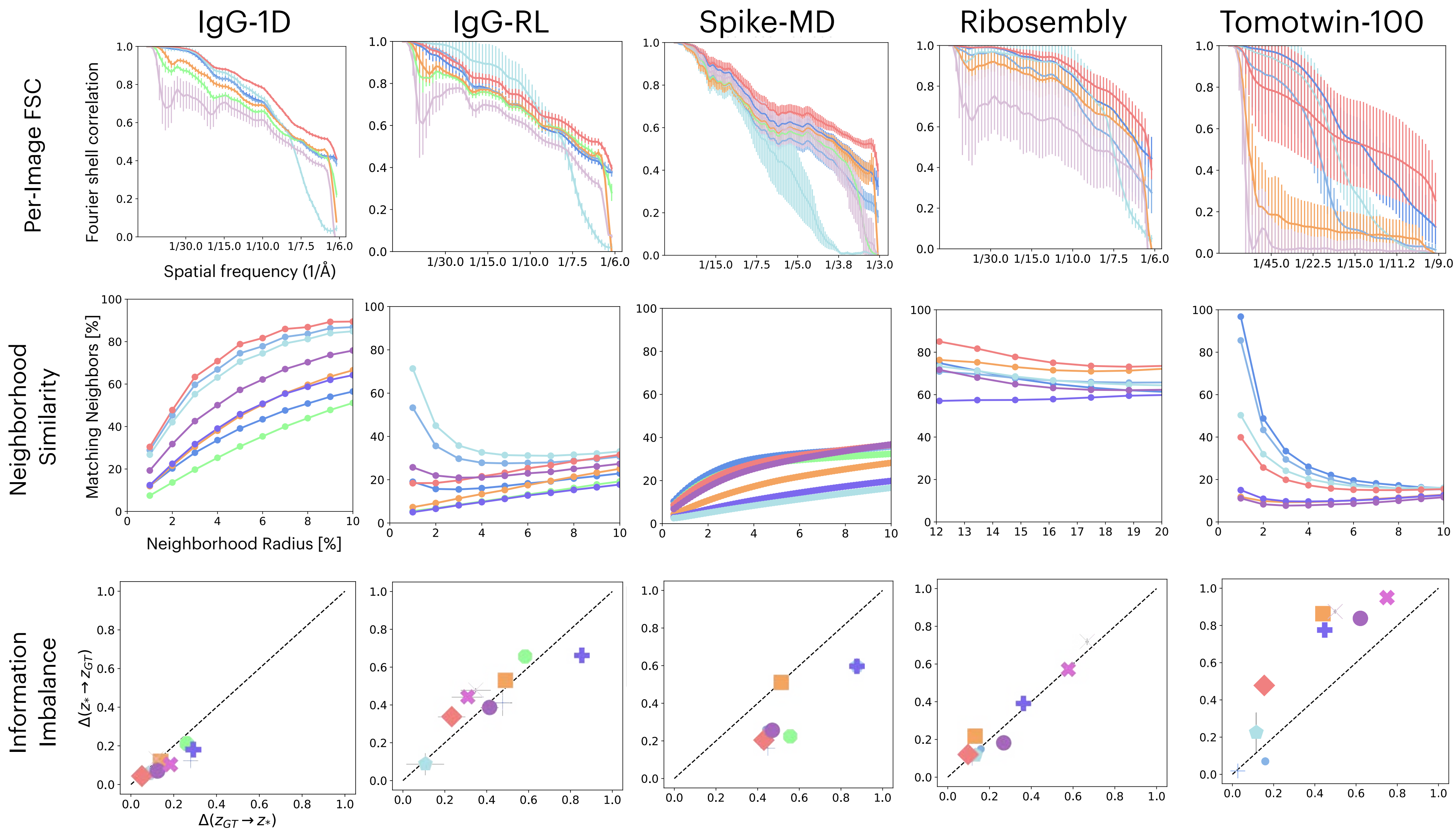
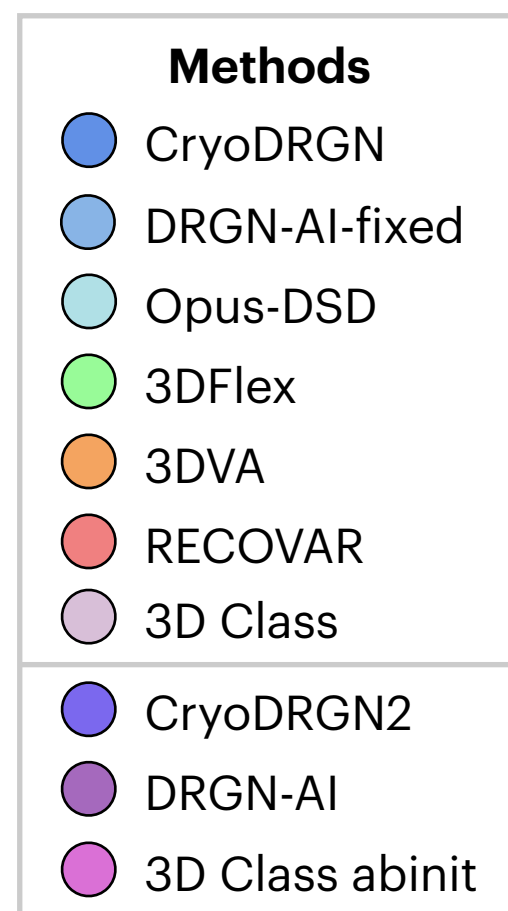
Spike-MD



Tomotwin-100



Quantitative Results

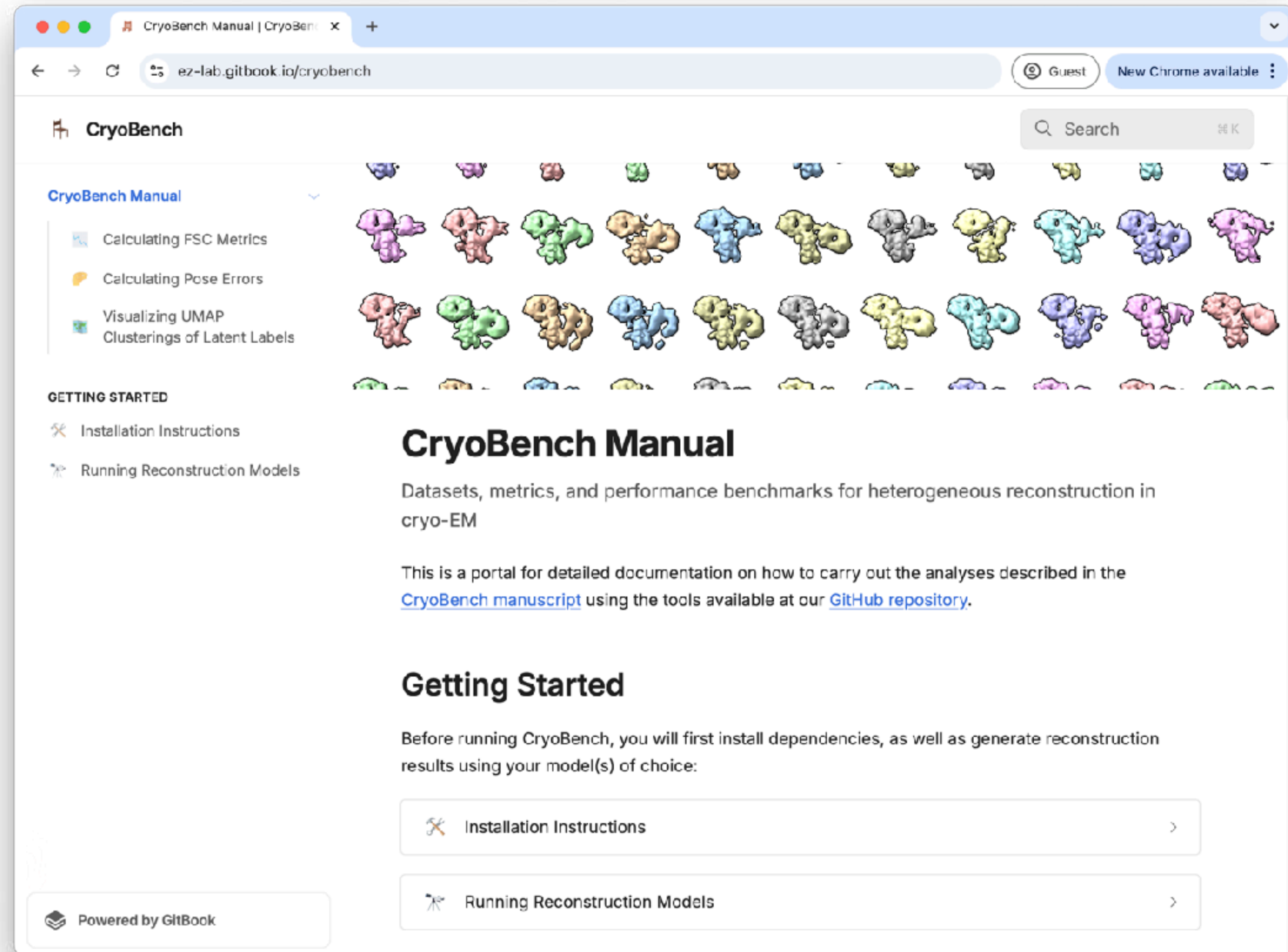


Quantitative Results

Method	IgG-1D		IgG-RL		Ribosemby		Tomotwin-100		Spike-MD	
	Mean (std)	Median	Mean (std)	Med	Mean (std)	Med	Mean (std)	Med	Mean (std)	Med
CryoDRGN	0.351 (0.028)	0.356	0.331 (0.016)	0.333	<u>0.412 (0.023)</u>	<u>0.415</u>	0.316 (0.046)	0.321	<u>0.340 (0.009)</u>	<u>0.340</u>
DRGN-AI-fixed	<u>0.364 (0.002)</u>	<u>0.364</u>	<u>0.348 (0.012)</u>	<u>0.350</u>	0.372 (0.032)	0.375	0.202 (0.044)	0.207	0.301 (0.012)	0.303
Opus-DSD	0.335 (0.026)	0.339	0.343 (0.016)	0.346	0.362 (0.083)	0.382	0.237 (0.049)	0.251	0.229 (0.027)	0.242
3DFlex	0.335 (0.003)	0.335	0.337 (0.007)	0.337	-	-	-	-	0.304 (0.011)	0.306
3DVA	0.349 (0.004)	0.350	0.333 (0.014)	0.335	0.375 (0.038)	0.375	0.088 (0.04)	0.077	0.324 (0.010)	0.323
RECOVAR	0.386 (0.005)	0.388	0.363 (0.011)	0.363	0.429 (0.018)	0.432	<u>0.258 (0.109)</u>	<u>0.254</u>	0.362 (0.011)	0.365
3D Class	0.297 (0.019)	0.291	0.309 (0.01)	0.307	0.289 (0.081)	0.288	0.046 (0.026)	0.037	0.307 (0.023)	0.308
CryoDRGN2	<u>0.32 (0.062)</u>	<u>0.342</u>	<u>0.301 (0.03)</u>	<u>0.306</u>	0.341 (0.059)	<u>0.356</u>	0.076 (0.016)	0.072	<u>0.245 (0.042)</u>	<u>0.260</u>
DRGN-AI	0.351 (0.01)	0.352	0.329 (0.028)	0.333	0.341 (0.083)	0.367	<u>0.072 (0.015)</u>	0.072	0.279 (0.017)	0.281
3D Class abinit	0.13 (0.046)	0.119	0.184 (0.022)	0.188	0.144 (0.036)	0.138	0.032 (0.012)	0.031	0.206 (0.009)	0.208

- Per-image Fourier Shell Correlation (Per-image FSC) is a distributional metric measuring volume reconstruction quality (higher better, best 0.5)
- No method dominates across different forms of heterogeneity

Datasets and code: cryobench.cs.princeton.edu



CryoBench

ez-lab.gitbook.io/cryobench

Search

CryoBench Manual

- Calculating FSC Metrics
- Calculating Pose Errors
- Visualizing UMAP Clusterings of Latent Labels

GETTING STARTED

- Installation Instructions
- Running Reconstruction Models

CryoBench Manual

Datasets, metrics, and performance benchmarks for heterogeneous reconstruction in cryo-EM

This is a portal for detailed documentation on how to carry out the analyses described in the [CryoBench manuscript](#) using the tools available at our [GitHub repository](#).

Getting Started

Before running CryoBench, you will first install dependencies, as well as generate reconstruction results using your model(s) of choice:

- Installation Instructions
- Running Reconstruction Models

Powered by GitBook

Thank you