

Appendix

Table of Contents

A Cryo-EM image formation model	12
B Architecture and training details	12
C Additional details on Sim2Struct-1000	13
D Volume metrics for heterogeneous reconstruction	13
E Ablations	14
F. Additional Quantitative Results	15
F.1. Traditional FSC	15
F.2. Supervised Classification Metrics	15
F.3. Additional baselines	18
G Conformational Heterogeneity	18
H Additional Qualitative Results	18

A. Cryo-EM image formation model

In this section, we expand on Section 2.1. The real space image formation model is not generally used during the training of neural reconstruction methods due to the cost of computing the projection as an integral. Instead, most reconstruction methods exploit the Fourier Slice Theorem (FST) [4] to simulate the image formation model efficiently in Fourier space. The FST states that the Fourier transform of a 2D projection of a density volume V is a 2D slice through the origin of V of the 3D Fourier transform of V . The FST simplifies the image formation model by circumventing the need to do projection by computing an integral. The Fourier image formation model (Equation 1) is derived from the real image formation model using the FST.

B. Architecture and training details

Here, we provide architectural and training details for our experiments.

Transformer encoder architecture For the Tomotwin-100 experiment, our transformer encoder consists of six transformer blocks, with each block consisting of a multi-head self-attention layer followed by an MLP consisting of two linear layers with GeLU activations [21]. Each MHSA layer consisted of 12 heads with each head

having dimension 64, with the total dimension being 768. The hidden dimension of each MLP is 3072. The Transformer encoder used for the Sim2Struct-1000 experiments were identical except eight blocks were used instead of six.

INR decoder architecture For all experiments, CryoHype used an INR consisting of 5 linear layers with hidden dimension 1024, random Fourier feature positional encoding, and residual connections between each layer. For fair comparison, these INR architecture hyperparameters were used for both CryoHype and all baseline models.

Training hyperparameters The Tomotwin-100 experiment was carried out with a batch size of 64, learning rate of 1e-4, a cosine learning rate schedule with a linear warmup of 5 epochs, patch size of 16, and was trained for a total of 50 epochs. For the 10 structure subset of Sim2Struct-1000, we used a learning rate of 5e-4, while for experiments with more structures, we used a learning rate of 2e-4. All other hyperparameters were the same as the Tomotwin-100 experiment. Finally, for the EMPIAR-10076, we used a patch size of 4 with a Gaussian low-pass filter cutoff of 50. A Gaussian low-pass filter is applied to each input token for the Vision Transformer (ViT), with all other hyperparameters being the same as the Tomotwin-100 experiment. Hyperparameters were tuned using grid search.

CryoDRGN We trained cryoDRGN using the official PyTorch implementation¹ (version 3.4.0b). For the synthetic datasets, all results were obtained using the default settings, with the z-dimension set to 8 and the total number of training epochs is 20 as described in [24]. For experiments on Sim2Struct-1000, we used a batch size of 64 and 50 training epochs. For experiments on EMPIAR-10076, we followed the settings outlined in [62], using 50 training epochs with the latent z-dimension set to 10.

Training splits As is standard for cryo-EM reconstruction, for all datasets we did not split any of the datasets and used the entire dataset for training. Datasets from CryoBench [24], including Tomotwin-100, can be downloaded from Zenodo².

GPUs, Memory, and Compute Time Model were trained on either NVIDIA A100, NVIDIA V100, NVIDIA A6000 GPU, or NVIDIA L40S GPUs. Each model was

¹<https://github.com/ml-struct-bio/cryodrgn>
²IgG-1D: <https://zenodo.org/records/12528292>,
Tomotwin-100: <https://zenodo.org/records/12528292>

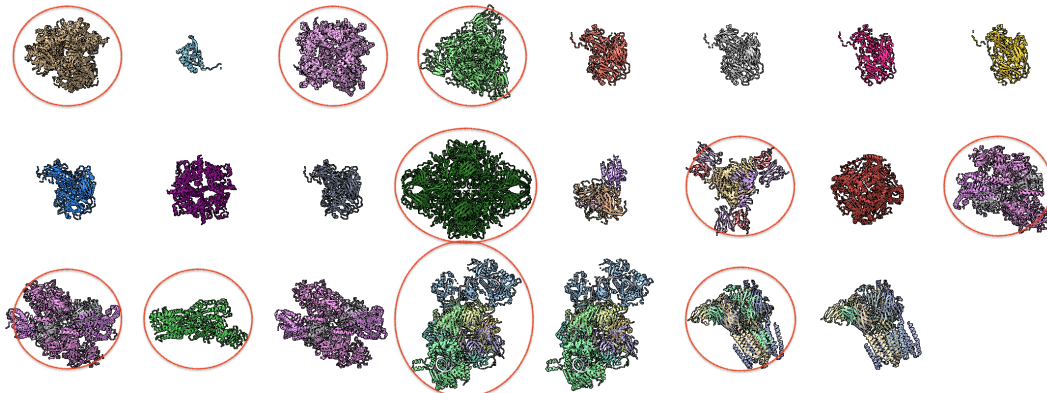


Figure 9. **Qualitative results of Sim2Struct-1000 filtering.** Examples of PDB structures from the Cryo2Struct dataset [16], highlighting those selected (circled) after filtering based on size and structural distinctiveness.

trained using 2 GPUs using PyTorch Lightning [13]. Training a CryoHype model for 50 epochs and performing inference took approximately 8 hours on 2 A100 GPUs for a dataset with 100K total particles (e.g. Tomotwin-100 or the 100 structure subset of Sim2Struct-1000). Sim2Struct-1000 used 77.7GB CPU memory and uses 24.5GB VRAM for a batch size of 32. For larger datasets, CPU memory consumption can be reduced using lazy dataset loading, at the cost of slower computation.

C. Additional details on Sim2Struct-1000

We constructed our benchmark, Sim2Struct-1000, using the Cryo2Struct dataset, initially containing 7,600 cryo-EM density maps paired with corresponding PDB structures [16]. We specifically chose Cryo2Struct over databases such as AlphaFold DB because each PDB structure is paired with an experimentally determined cryo-EM map, thereby avoiding potential issues associated with synthetic proteins, such as disordered regions complicating downstream structure determination [1].

We filtered the initial dataset based on the axis-aligned bounding box dimensions of each PDB structure, retaining only those whose maximum side length fit within the interval [88, 118]. This interval ensures each protein comfortably fits within a 256-pixel reconstruction grid, approximately double the protein’s maximal length, to accommodate translations within a ± 20 -pixel range during image projection. This step was critical, as overly small proteins lacked distinctive structural features at our target resolution and too-large proteins would be truncated during image projection. After filtering, we further refined the dataset by eliminating near-duplicate structures, retaining only the first instance of structures sharing identical initial three-character prefixes from their four-character PDB identifiers (e.g. retaining 6cs3.pdb out of 6cs3.pdb, 6cs4.pdb, 6cs5.pdb, etc.). Figure 9 illustrates the filtering process.

From the resulting structures, we selected the first 1000 for further processing, subsequently creating smaller subsets (10, 100, 200, and 500 structures) to evaluate model performance at varying dataset scales.

Each retained PDB structure was centered by translating the atomic coordinates to place the geometric centroid at the origin. Next, standardized density maps were generated using the `molmap` command in ChimeraX with parameters set at 3 Å resolution, 1.5 Å grid spacing, and a 256-pixel box size. Although the Cryo2Struct dataset included original EMDB maps, variations in their resolution (ranging 1–4 Å) and box dimensions prompted us to generate standardized synthetic volumes to ensure downstream data consistency [16].

From these standardized volumes, we simulated cryo-EM images by generating 1000 projections per structure, applying a contrast transfer function (CTF), introducing noise corresponding to an SNR of 0.01, and downsampling images to 128×128 pixels. The final standardized images have dimensions of 128 pixels, 6 Å resolution, and pixel size of 3 Å.

Future work includes evaluating additional protein structure databases to determine their suitability for simulating cryo-EM datasets, particularly to facilitate the development of more generalizable and large-scale cryo-EM reconstruction methods. We plan to release our dataset upon publication on Zenodo³ with the CC-BY license. A backup copy of the dataset will temporarily be available on Dropbox⁴.

D. Volume metrics for heterogeneous reconstruction

Per-Image FSC We use *per-image FSC* to jointly assess heterogeneity and reconstruction quality following [24]. In

³<https://zenodo.org/uploads/17916258>

⁴<https://www.dropbox.com/home/Jeffrey%20Gu/Sim2Struct>

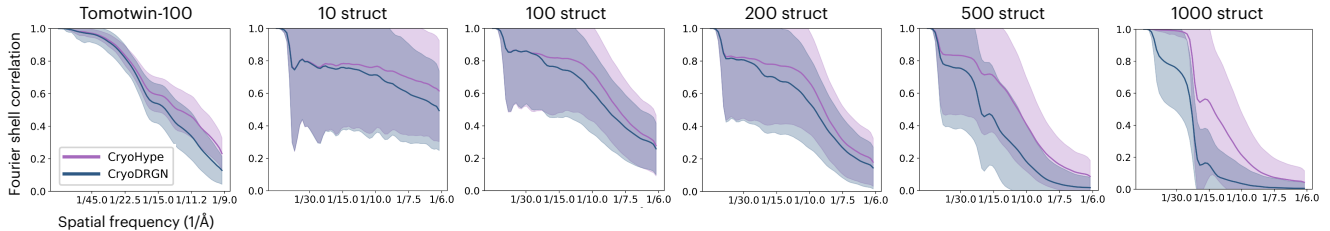


Figure 10. **Per-Image FSC.** Each curve shows the average FSC curve across all conformations with error bars indicating the standard deviation. The full FSC curves are shown in Appendix.

cryo-EM, using the Fourier Shell Correlation (FSC) curve is a standard technique for comparing two volumes. The FSC curve calculates the correlation between two volumes (e.g. a reconstructed volume and a ground truth volume) across spherically averaged radial shells in the Fourier domain. Cryo-EM reconstruction methods often reconstruct a volume from a single input image. *Per-image FSC* evaluates the heterogeneous reconstruction quality of a method by computing the FSC between these per-image reconstructions and ground truth volumes. As in [24], to compute the per-image FSC we sample one image per conformation to assess the distribution of reconstructions and compute the area under the FSC curve as a summary statistic.

The averaged FSC curves for cryoDRGN and CryoHype on the Tomotwin-100 and all subsets of the Sim2Struct-1000 dataset are presented in Figure 10 and summarized in Table 1.

Additional Volume Metrics To assess the quality of reconstructed volumes against ground truth volumes, we propose the usage of two new metrics, volumetric IoU and Chamfer distance. *Volumetric IoU* (vIoU) is defined as the intersection of the ground truth and predicted volume divided by their union. Higher vIoU values indicate better alignment between the predicted and ground truth volumes. We evaluate shape accuracy using *Chamfer Distance* (CD), a metric commonly used for assessing the quality of point cloud reconstructions. CD measures the average bidirectional distance between points in the ground truth and predicted point clouds. Lower CD values indicate more accurate reconstructions. Together, these two metrics provide complementary insights into reconstruction quality by evaluating volumetric overlap and shape precision.

The results in Table 4 demonstrate the impact of CryoHype’s main components on these metrics, demonstrating their ability to capture insights on performance. Figure 11 validates these two metrics on the IgG-1D dataset, a dataset modeling a 1D circular motion of one of the fragment antibody (Fab) domains of the human immunoglobulin G (IgG) protein [24]. Figure 12 highlights metric sensitivity to dataset-specific structural characteristics and the increased difficulty posed by more diverse

datasets like Sim2Struct-1000. Figure 13 shows the correlation between per-image FSC and our proposed metrics.

Density Threshold Selection To determine which voxels constitute the foreground (object) versus the background in volumetric data, we tuned near-optimal density thresholds to 6.0 for predictions on the Tomotwin-100 dataset, 220 for Sim2Struct-1000 predictions, and 5×10^{-5} for all ground truth structures, ensuring consistent voxel segmentation for evaluation. To aid in selecting these thresholds, we opened the figures in ChimeraX [39] to visually assess and refine the isosurface settings. Final levels were chosen to maximize the completeness of the visualized structures and minimize noisy artifacts.

Thresholding Limitations and Implications A current limitation of our new metrics is that each predicted structure may require individual density level tuning, which limits the ability to draw dataset-wide performance conclusions using these density-dependent metrics. A potential solution to this is to optimize the density level for each predicted structure on a given metric with a greedy search algorithm. Moreover, Chamfer distance is highly sensitive to outliers, resulting in occasional extreme values, particularly at lower density thresholds where noisy or disconnected points are included. Higher density thresholds reduce the number of high-CD outliers but must be carefully tuned to avoid excessive loss of volume and structural details. This sensitivity highlights the importance of selecting density thresholds that balance object completeness with noise reduction.

E. Ablations

In this section, we provide further details on the ablation experiments in Section 4.4. Additional quantitative results using our proposed Chamfer distance and volumetric IoU metrics can be found in Table 4. We also provide ablations of different sized cryoDRGN [62] variants vs CryoHype.

Concatenation ablation We modify our architecture to only condition the INR decoder by concatenation, turning

Table 4. Ablation study on CryoHype examining the two main components of the model, evaluated by Chamfer distance and volumetric IoU.

Method	Tomotwin-100			
	↓ CD (std)	Med	↑ vIoU (std)	Med
Concatenation	7.157 (11.300)	2.725	0.487 (0.198)	0.556
U-Net encoder	5.154 (2.360)	4.549	0.451 (0.081)	0.455
CryoHype	2.185 (0.462)	2.111	0.615 (0.061)	0.621

Table 5. Comparison of CryoHype vs different-sized variants of CryoDRGN [62] on the Tomotwin-100 dataset, evaluated using FSC_{AUC} .

Method	Params	Tomotwin-100	
		Mean FSC_{AUC} (std)	Med FSC_{AUC}
CryoDRGN (base)	20M	0.316 (0.046)	0.321
CryoDRGN (6 layer encoder)	21M	0.316 (0.040)	0.322
CryoDRGN (8 layer encoder)	23M	0.319 (0.038)	0.321
CryoDRGN (6 layer decoder)	21M	0.324 (0.040)	0.328
CryoDRGN (8 layer decoder)	23M	0.324 (0.041)	0.330
CryoDRGN (encoder dim 4096, decoder dim 4096)	155M	0.338 (0.027)	0.340
CryoHype	50M	0.346 (0.033)	0.353

our network into an autoencoder. We do this by removing all linear heads $Head_i$ except one (see Sec 2.3) and modifying the forward pass of the network to apply the single linear head to just the last output token to produce a latent vector \mathbf{z} , which is used to condition the decoder by concatenation.

U-Net hypernetwork The U-Net [5, 47] encoder for the U-Net ablation takes as input images of shape $[1, D, D]$ and has output shape $[L, D, D]$, where L is the number of layers of the INR and $D = 256$ is a multiple of one of the dimensions of the INR’s weight matrices and initial features size 144. Since our input originally has shape $[1, 129, 129]$, we reshape the input to size $[1, D, D]$ using interpolation with nearest upsampling. We then apply each linear head $Head_i, 1 \leq i \leq L$ to a channel of the output, transforming D to the correct size for the given layer. We then repeat the other dimension as necessary to get the full correct shape. Training is unchanged.

MLP hypernetwork We base our MLP hypernetwork architecture on that of Light Field Networks [55], which uses a separate MLP encoder to generate the weights of each layer. We adapt the architecture of [55] to produce weight tokens instead of directly producing the weights, leaving the rest of the architecture unchanged from the ViT hypernetwork (see Section 2).

Larger cryoDRGN variants The base CryoHype model requires more parameters than the base cryoDRGN model

(50M vs 20M). In this ablation study, we show that scaling cryoDRGN does not allow it to match the performance of CryoHype (Table 5). While the performance of cryoDRGN increases with increasing encoder and decoder capacity, the performance of cryoDRGN increases only very gradually, showing that CryoHype scales much better than cryoDRGN. The largest cryoDRGN variant still shows lower performance than the base CryoHype model on Tomotwin-100 (c.f. Table 1), despite having 3 times the number of parameters (155M to 50M).

F. Additional Quantitative Results

In this section, we provide additional quantitative evaluations of our method.

F.1. Traditional FSC

In Table 6, we evaluate CryoHype vs cryoDRGN [62] using the traditional FSC at 0.143 and FSC at 0.5 metrics on both Tomotwin-100 and all Sim2Struct-1000 subsets. We find that, in line with our other quantitative metrics, the performance of CryoHype matches or exceeds the performance of cryoDRGN on all datasets, with CryoHype performing better as the degree of compositional heterogeneity, as measured by the number of distinct structures, gets more and more extreme.

F.2. Supervised Classification Metrics

Since CryoHype was evaluated quantitatively on synthetic datasets in the fixed-pose setting where the particle poses are known, it also makes sense to evaluate the perfor-

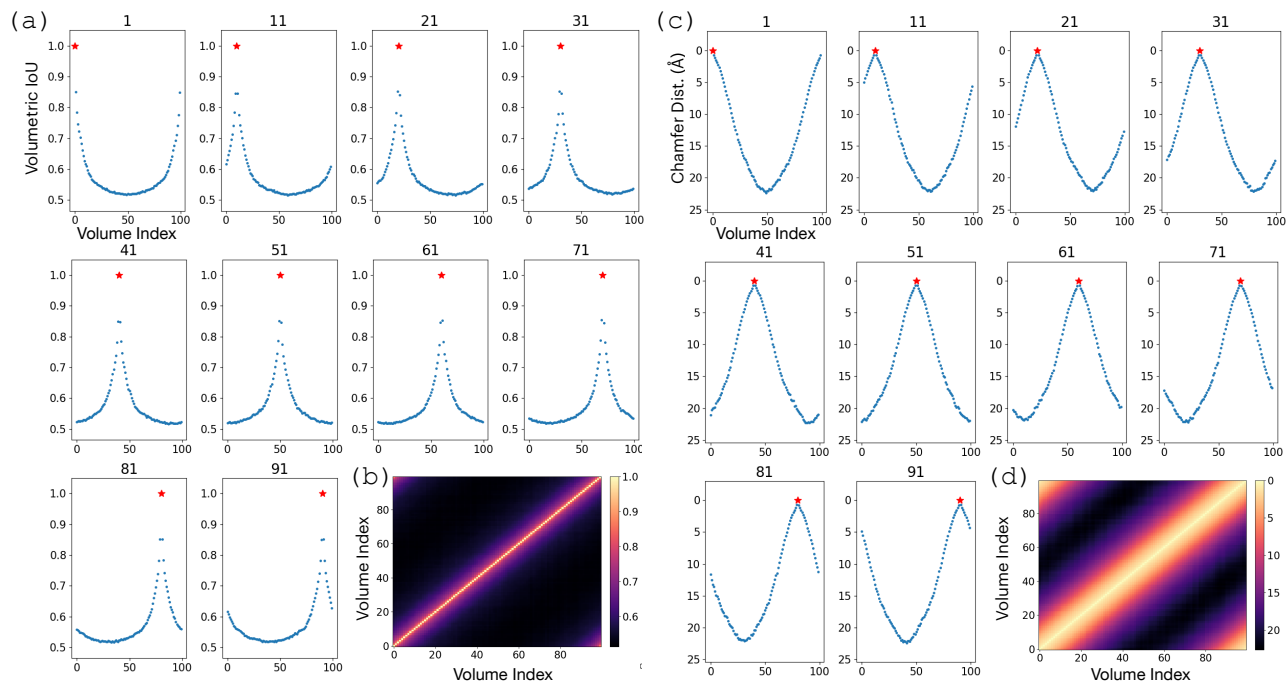


Figure 11. **Dynamic range of CD and vIoU metrics between ground truth volumes for IgG-1D.** (a) Volumetric IoU computed between one reference G.T. structure and all 100 G.T. structures of the IgG-1D dataset. Each plot corresponds to one reference G.T. volume, indicated by the number above the plot and the red star in each plot. Points higher on the y-axis indicate greater structure similarity. (b) Heatmap of vIoU showing the pairwise volumetric IoU for all pairs of G.T. structures. Lighter colors indicate greater structure similarity. Parts (c) and (d) display the corresponding plots for Chamfer distance.

Table 6. CryoHype and cryoDRGN [62] evaluated using traditional FSC metrics on both the Tomotwin-100 dataset and all subsets of the Sim2Struct-1000. Lower is better, with 2 the best possible FSC.

Method	Dataset	Structures	FSC at 0.143		FSC at 0.5	
			Mean	Median	Mean	Median
CryoDRGN CryoHype	Tomotwin-100	100	2.17 2.03	2.06 2.00	3.14 2.81	3.08 2.61
CryoDRGN CryoHype	Sim2Struct-10	10	2.00 2.00	2.00 2.00	2.00 2.00	2.00 2.00
CryoDRGN CryoHype	Sim2Struct-100	100	2.01 2.00	2.00 2.00	2.73 2.30	2.75 2.29
CryoDRGN CryoHype	Sim2Struct-200	200	2.18 2.01	2.01 2.00	3.09 2.63	2.91 2.72
CryoDRGN CryoHype	Sim2Struct-500	500	3.96 2.68	3.37 2.29	5.93 3.49	4.74 3.20
CryoDRGN CryoHype	Sim2Struct-1000	1000	6.45 3.87	6.10 3.76	9.31 4.53	6.74 4.57

formance of CryoHype using supervised classification evaluation metrics. We find that overall, CryoHype outperforms cryoDRGN in supervised classification metrics (Table 7). CryoHype maintains close to perfect classification metrics regardless of the number of structures, whereas cry-

oDRGN’s performance is close to perfect at 10 and 100 structures and begins to drop at 200 structures, degrading rapidly as the number of structures increases. This mirrors the trends observed in the other metrics (FSC_{AUC}, CD, vIoU). The large drops at 500 and 1000 structures can

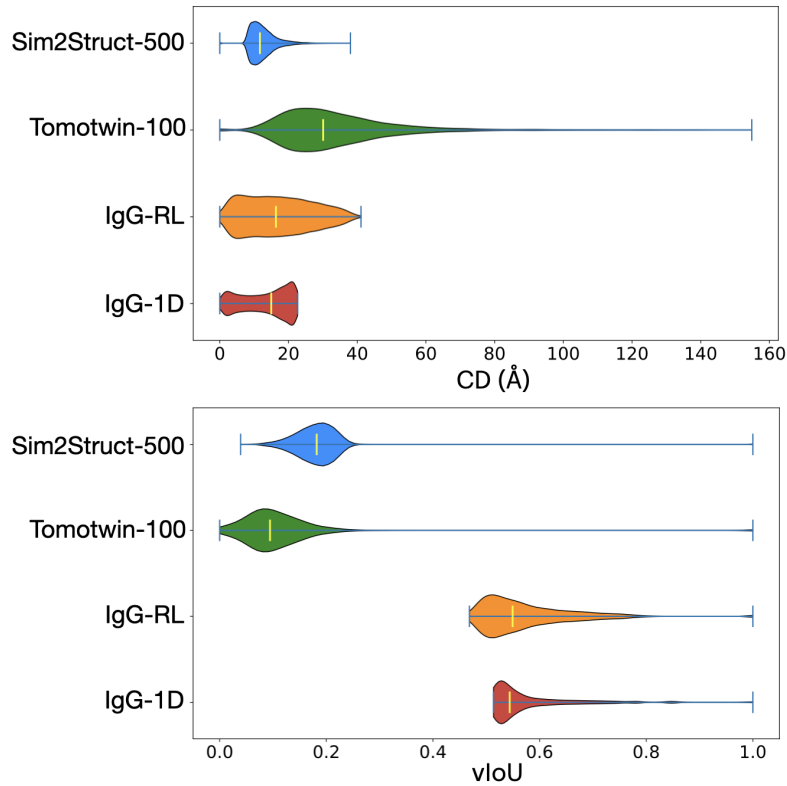


Figure 12. **Dynamic range of CD and vIoU metrics between ground truth volumes for each dataset.** Violin plots show the distribution of Chamfer distance (CD) and volumetric IoU (vIoU) metrics on all G.T. structures for four datasets: Sim2Struct-1000, Tomotwin-100, IgG-RL, and IgG-1D. Lower CD and higher vIoU indicate better shape and volume similarity. Sim2Struct-1000 and Tomotwin-100 exhibit wider variation due to greater structural diversity, while IgG-1D shows the most compact distributions, reflecting its homogeneity.

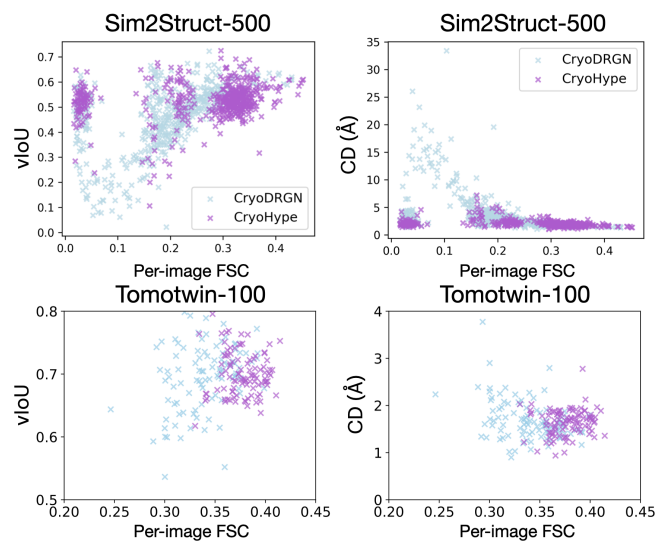


Figure 13. **Quantitative Metric Comparison for Tomotwin-100 and Sim2Struct-1000.** Higher FSC values positively correlate with vIoU and negatively correlate with CD, reflecting improved reconstruction accuracy. CryoHype (purple) outperforms cryoDRGN (blue) across both datasets.

Table 7. CryoHype and cryoDRGN [62] evaluated using supervised classification metrics on all subsets of the Sim2Struct dataset.

Method	Dataset	Structures	Accuracy	Precision	Recall	F1
CryoDRGN CryoHype	Sim2Struct-10	10	0.9998 0.9999	0.9998 0.9999	0.9998 0.9999	0.9998 0.9999
CryoDRGN CryoHype	Sim2Struct-100	100	0.9621 0.9506	0.9523 0.9421	0.9621 0.9506	0.9555 0.9451
CryoDRGN CryoHype	Sim2Struct-200	200	0.9397 0.9632	0.9298 0.9557	0.9397 0.9632	0.9311 0.9581
CryoDRGN CryoHype	Sim2Struct-500	500	0.8375 0.9719	0.8202 0.9635	0.8375 0.9719	0.8166 0.9663
CryoDRGN CryoHype	Sim2Struct-1000	1000	0.6792 0.9753	0.6508 0.9713	0.6792 0.9753	0.6468 0.9724

Table 8. Performance of cryoSPARC’s fixed pose 3D classification method on the Sim2Struct dataset, evaluated using FSC_{AUC} .

Method	Dataset	FSC_{AUC}	
		Mean (std)	Med
cryoSPARC 3D class (fixed pose)	Sim2Struct-10	0.204 (0.036)	0.224
cryoSPARC 3D class (fixed pose)	Sim2Struct-100	0.071 (0.055)	0.037

be seen visually in latent space, where the poor quantitative classification metrics for cryoDRGN are reflected as increasingly poor organization at the center of latent space (see Figure 5).

F.3. Additional baselines

In this section, we report the results of traditional maximum likelihood classification methods with fixed pose (e.g. cryoSPARC 3D classification [42]) on our new Sim2Struct-1000 dataset. As with the cryoSPARC results reported in our paper (originally from CryoBench [24], see Table 1), we confirm that cryoSPARC’s 3D classification method struggles with compositional heterogeneity and especially the extreme compositional heterogeneity considered in our paper (Table 8). In particular, the results on the 100 structure subset of Sim2Struct-1000 mirror that of Tomotwin-100. At even more extreme levels of heterogeneity (>100 classes), we find that cryoSPARC 3D classification throws an error.

G. Conformational Heterogeneity

Although CryoHype was not designed to target conformational heterogeneity, it is still able to recover conformational heterogeneity. In Figure 14, CryoHype is able to successfully recover the circular motion of IgG-1D [24] when augmented with an additional smoothness loss.

H. Additional Qualitative Results

In this section, we provide additional examples of reconstructed volumes for both CryoHype and cryoDRGN on

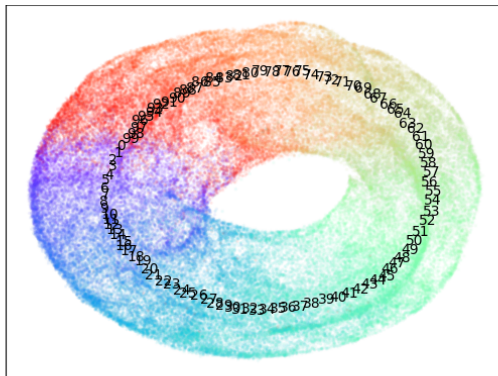


Figure 14. **IgG-1D latent space** We find that CryoHype successfully captures the 1D circular motion of the IgG-1D dataset. The numbers $0 \leq i \leq 99$ label the mean latent for each of the 100 conformations of IgG-1D, where $(3.6i)^\circ$ is the angle of the Fab domain relative to the fixed domains.

each dataset.

Figures 15 and 16 show the groundtruth, CryoHype, and cryoDRGN reconstructions for Tomotwin-100 and Sim2Struct-1000, respectively. We see that overall, CryoHype is both able to reconstruct some shapes that cryoDRGN cannot. In particular, cryoDRGN struggles with proteins with loops, while CryoHype does a better job of reconstructing these proteins. Additionally, CryoHype generally has higher resolution and preserves fine details better than cryoDRGN.

Figure 17 shows reconstructed volumes for each class of EMPIAR-10076. Each CryoHype structure is generated by randomly sampling from latent encoding of particles with the corresponding class assignments from [10].

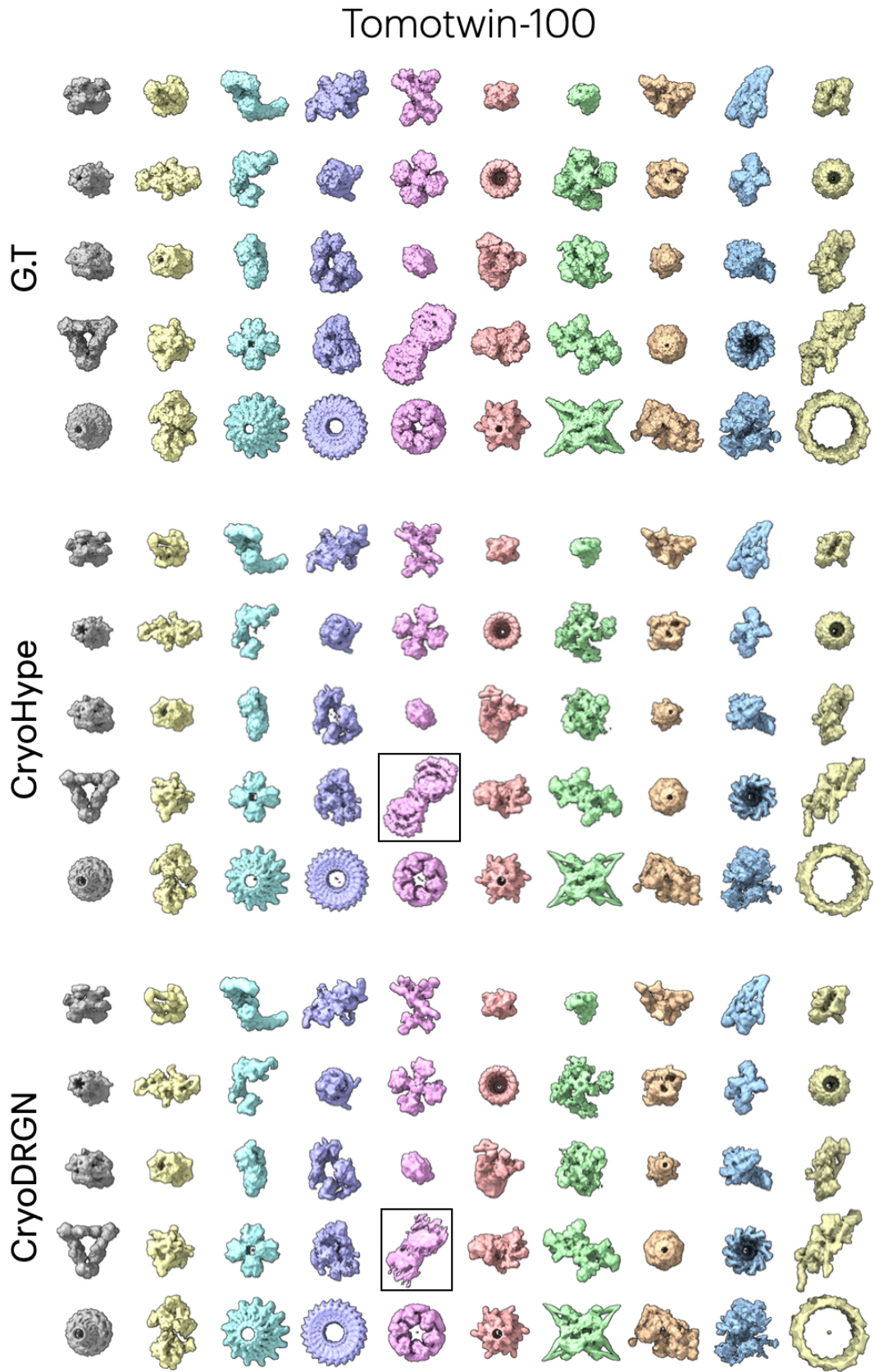


Figure 15. **Tomotwin-100 (SNR 0.01) qualitative results** Last 50 volumes reconstructed by CryoHype and CryoDRGN for the Tomotwin-100 dataset.

Sim2Struct-1000

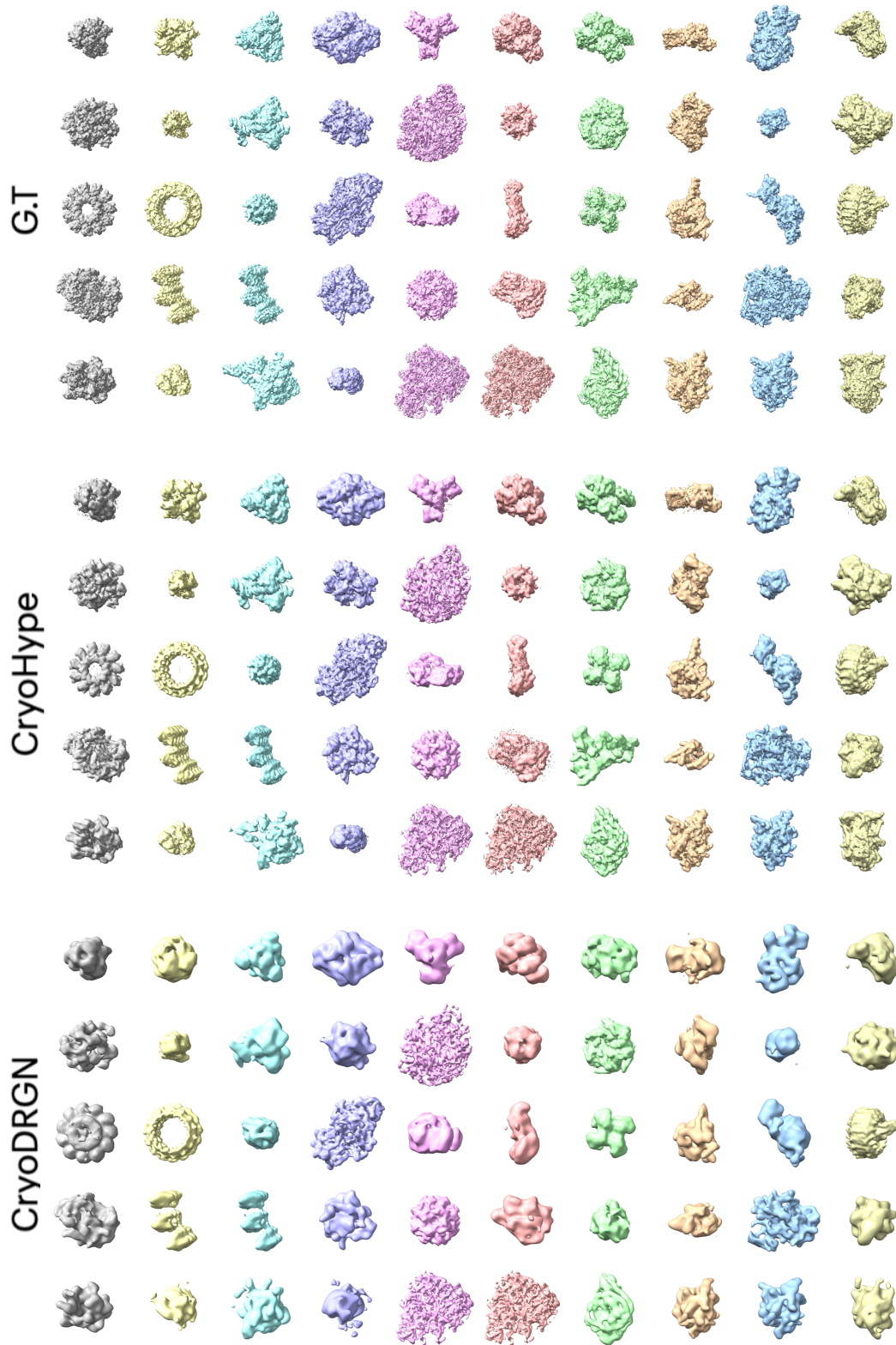


Figure 16. **Sim2Struct-1000 qualitative results with G.T** Last 50 volumes reconstructed by CryoHype and cryoDRGN for the Sim2Struct-1000 dataset.

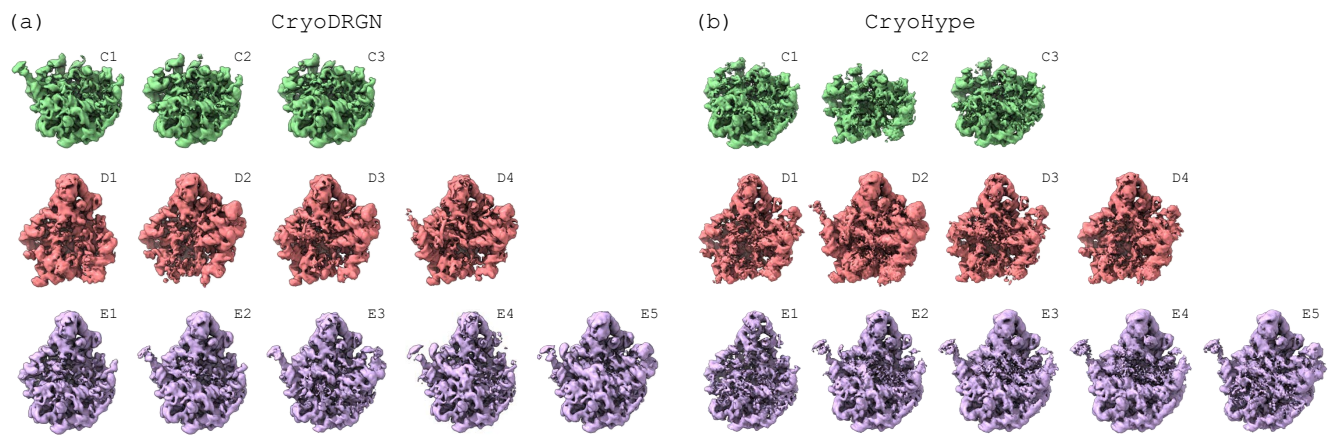


Figure 17. **EMPIAR-10076** Reconstructed volumes of ribosome assembly minor classes for (a) CryoDRGN and (b) CryoHype.