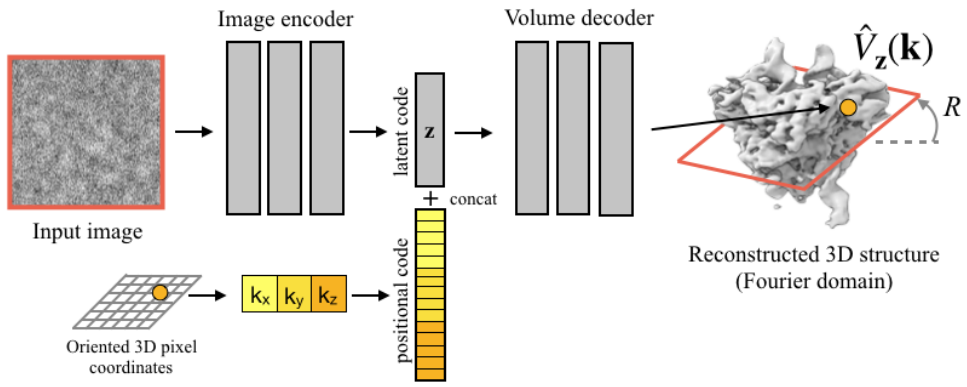

Figure S1: The cryoDRGN architecture for heterogeneous cryo-EM reconstruction. Adapted from Zhong et al [55].

A. Additional Methodological Details

A.1. cryoDRGN background and architecture

Figure 1(c) shows the architecture used in cryoDRGN's neural representation for 3D volume and Figure S1 shows an overview of the cryoDRGN method for heterogeneous cryo-EM reconstruction.

CryoDRGN directly approximates the volumetric density function, $V : \mathbb{R}^3 \rightarrow \mathbb{R}$, with a coordinate-based MLP where each input Cartesian coordinate is featurized with a fixed positional encoding function consisting of D sinuosids of varying frequency:

$$pe^{(2d)}(k) = cos(\gamma(d)k); d = 0, ...D/2 - 1 \tag{3}$$

$$pe^{(2d+1)}(k) = sin(\gamma(d)k); d = 0, ...D/2 - 1 \tag{4}$$

$$\gamma(d) = D\pi \left(\frac{2}{D}\right)^{d/(D/2-1)} \tag{5}$$

Without loss of generality, the volume is represented on a sphere centered at the origin with radius 0.5. Wavelengths of the positional encoding follow a geometric series from $2/D$ to 1, i.e. the Nyquist limit of the imaging dataset to the width of the volume. For noisy (e.g. real) datasets, we follow prior work [55, 54] and scale all wavelengths of the featurization by 2π. Volume and image data are represented in Hartley space [16], which is closely related to Fourier space as the real minus imaginary component for real-valued signals. We use a white noise error model; the loss is computed as the mean square error between the input images and central slices from the 3D volume.

For heterogeneous reconstruction, we use the coordinate-based neural network to parameterize a generative model over volumes (Figure S1). While the latent pose variables are constrained as their geometric operations on the input coordinates, the unconstrained variables are directly input as additional dimensions to the MLP. We learn the generative model by amortized variational inference based on the standard variational autoencoder (VAE): An encoder predicts the approximate conditional posterior, $q(z|X)$, whose form is an isotropic Gaussian given a cryo-EM image. A sample from this distribution is broadcast to all pixel coordinates of an (oriented) image. Then the decoder is then evaluated pixel-by-pixel to reconstruct the slice. The prior $p(z)$ is a standard normal. The image orientation is unknown for ab initio reconstruction; we therefore propose an efficient search algorithm over the 5-D space of poses performed within the training loop.

A.2. 5-D pose search

Given a cryo-EM image and the current weights of the volume decoder, $\hat{V}$, we perform a grid search over the complete space of poses, $SO(3) \times \mathbb{R}^2$ (Algorithm 1). Our method exhaustively evaluates a discretization over the 5-D space at a base grid resolution and iteratively refines the top K poses ($K = 8$ by default), doubling the resolution of the grid for M iterations ($M = 4$ by default). When computing the error for a given pose at the base resolution, we band-limit the image and model slice to $|k| < k_{min}$ for computational efficiency and to prevent overfitting to high-frequency noise. During the refinement iterations, the frequency band-limit is linearly increased up to k_{max}. As an additional computational speedup, we initialize

k_{max} to a low value and let it increase over multiple epochs of training as the volume representation in early epochs will be lower resolution.

The uniform, incremental grids on $SO(3)$ are parameterized using the Hopf fibration [52], which is composed of the product of the Healpix [12] grid on the sphere and an ordinary grid on the circle S^1. The cryoDRGN2 base grid on $SO(3)$ has 4,608 rotations with spacing of 15° by default. For in-plane translations, we use a grid centered at the origin with extent, $[-t_{ext}, t_{ext}]^2$ and a spacing of $2 * t_{ext}/T = 2 * (20 \text{ pixels})/14 \approx 2.86$ pixels for D=128 images by default. As detailed in Section 3.1.2, after the base resolution poses are evaluated, we select the top K unique rotations, and subdivide the grid to get $8K$ new rotations at the next incremental grid. We select the best rotation t^*, then generate a new grid centered at t^* with half the extent and the same number of grid points for the next incremental grid.

Algorithm 1 CryoDRGN2 pose search

procedure OPTPHI($\hat{X}, \hat{V}$) ▷ Find the optimal pose for $\hat{X}$ given the current decoder $\hat{V}$
  $k_{min} \leftarrow 12,\ k_{max} \leftarrow min(48, D/2),\ M \leftarrow 4,\ K \leftarrow 8$
  $\gamma \leftarrow 15^\circ$
  $\Phi \leftarrow (\phi, (0,0))$ for ϕ in $SO(3)$ rotation grid at resolution γ.
  $\Psi \leftarrow$ Translation grid in $\mathbb{R}^2$ at base resolution centered at the origin.
  $k \leftarrow k_{min}$ ▷ frequency cutoff for computing err
  for $iter = 1 \ \ldots \ M$ **do**
    $errors \leftarrow \{\}$
    for $(\phi, t) \in \Phi$ **do** ▷ Compute error for all rotations
      $err, dt^* = \arg\min_{dt \in \Psi} \left|\left| \text{SLICE}(\hat{V}_z, \phi) - S_l(t + dt)\hat{X} \right|\right|^2$
      $t^* = t + dt^*$
      $errors \leftarrow errors \cup (err, \phi, t^*)$
    $\Phi^* \leftarrow$ the K pairs (ϕ, t^*) with lowest err from $errors$.
    $\gamma \leftarrow \gamma/2$ ▷ Halve rotation grid spacing
    $\Psi \leftarrow \Psi/2$ ▷ Halve translation grid spacing
    $\Phi_{new} \leftarrow \{\}$
    for $(\phi, t) \in \Phi^*$ **do**
      $\Phi_{new} \leftarrow \Phi_{new} \cup \text{SUBDIVIDE}(\phi_i, \gamma, t)$ ▷ Subdivide rotations to 8 new rotations at 2x resolution
    $\Phi \leftarrow \Phi_{new}$
    $k \leftarrow k + (k_{max} - k_{min})/(N_{iter} - 1)$ ▷ Increase frequency band limit
  return the min-err element of Φ^*

A.3. Hyperparameter sweep

We perform a grid search over possible hyperparameter values, including the base resolution for rotations (γ), the base resolution for translations (T), the frequency band-limit bounds (k_{min}, k_{max}), the number of subdivisions of the grid (M), and the number of kept poses at each subdivision (K) (Table S1). Each of the settings are evaluated by computing the alignment error for 1000 held out images of the synthetic spike dataset, the real 80S, and the real RAG dataset, aligned on a cryoDRGN model pre-trained using either ground truth (synthetic) or reference (real) poses.

We find that for the tested datasets, using a base resolution beyond 15° does not lead to improved accuracy. Increasing the maximum band-limit frequency k_{max} or the number of poses to refine K leads to increased accuracy, though increased training time. Our chosen defaults ("base" in Table S1) are $\gamma = 15^\circ$, $K = 8$, $M = 4$, $k_{min} = 12$, $k_max = 48$, $T = 14$, and $t_{ext} = 20$. With these settings the final grid resolution for poses is 0.94° for rotations and 0.18 pixels for translations.

B. Additional dataset details

Example images and reference volumes for the datasets used in our study are shown in Fig. 1a, Fig. 2, and Fig. 5. Dataset statistics are provided in Table S2.

	Spike			80S			RAG		
	MSE	*MedSE*	*Time*	*MSE*	*MedSE*	*Time*	*MSE*	*MedSE*	*Time*
base	0.208	0.002	0:00:56	0.004	0.0006	0:00:56	3.436	0.164	0:00:56
legacy	0.799	0.007	0:00:26	0.648	0.0023	0:00:26	4.171	4.993	0:00:26
k=[12,20]	0.259	0.004	0:00:31	0.004	0.0011	0:00:30	3.512	0.363	0:00:31
k=[12,48]	0.165	0.001	0:01:50	0.003	0.0004	0:01:49	3.325	0.107	0:01:49
k=[12,64]	0.159	0.001	0:03:01	0.003	0.0003	0:03:00	3.353	0.097	0:03:00
k=[10,20]	0.271	0.005	0:00:28	0.019	0.0011	0:00:28	3.636	1.645	0:00:28
k=[10,32]	0.208	0.002	0:00:53	0.015	0.0006	0:00:53	3.508	0.242	0:00:53
k=[14,20]	0.260	0.004	0:00:33	0.004	0.0011	0:00:32	3.514	0.405	0:00:32
k=[14,32]	0.196	0.002	0:00:59	0.004	0.0006	0:00:58	3.437	0.164	0:00:58
$\gamma_0 = 30°$	0.872	0.004	0:00:51	0.706	0.0017	0:00:50	4.127	4.901	0:00:50
$\gamma_0 = 15°$	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.385	0.142	0:00:56
$\gamma_0 = 7.5°$	0.219	0.002	0:01:28	0.001	0.0005	0:01:32	3.469	0.200	0:01:27
M=1	0.213	0.011	0:00:23	0.019	0.0095	0:00:23	3.601	1.256	0:00:23
M=2	0.210	0.004	0:00:34	0.006	0.0030	0:00:33	3.414	0.166	0:00:34
M=3	0.208	0.003	0:00:44	0.004	0.0011	0:00:44	3.451	0.160	0:00:44
M=4	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.412	0.144	0:00:56
M=5	0.219	0.002	0:01:08	0.004	0.0005	0:01:07	3.385	0.129	0:01:07
M=6	0.225	0.002	0:01:19	0.003	0.0005	0:01:23	3.436	0.177	0:01:19
M=7	0.232	0.002	0:01:29	0.003	0.0005	0:01:29	3.496	0.275	0:01:29
K=2	0.373	0.003	0:00:17	0.037	0.0008	0:00:17	3.860	3.271	0:00:17
K=4	0.256	0.002	0:00:31	0.017	0.0006	0:00:30	3.477	0.214	0:00:30
K=8	0.208	0.002	0:00:57	0.004	0.0006	0:00:56	3.400	0.141	0:00:56
K=16	0.196	0.002	0:01:50	0.001	0.0006	0:01:48	3.428	0.114	0:01:49
K=24	0.196	0.002	0:02:25	0.001	0.0006	0:02:42	3.420	0.124	0:02:43
$T = 7$	0.208	0.002	0:00:53	0.004	0.0006	0:00:52	3.425	0.155	0:00:52
$T = 28$	0.208	0.002	0:01:13	0.004	0.0006	0:01:12	3.407	0.133	0:01:12

Table S1: Hyperparameter sweep for pose search. The default cryoDRGN2 settings ("base") are $\gamma = 15°$, $K = 8$, $M = 4$, $k_{min} = 12$, $k_{max} = 48$, $T = 14$, and $t_{ext} = 20$, and other rows describe modifications to "base". The "legacy" settings approximate the default hyperparameters used in the cryoDRGN-BNB grid, which are $K = 8$, $M = 4$, $k_{min} = 12$, $k_{max} = 20$, $T = 7$, and $t_{ext} = 20$. Mean, median square error, and timing are computed for aligning 1000 images from each dataset aligned on a pretrained model. Rotation error is defined as the squared Frobenius norm of the rotation matrices after global rigid body alignment.

Dataset	**D**	N_{images}	**Å/pixel**	**Noise?**	**CTF?**
Hand ideal	64	50,000	6	N	N
Hand noisy	64	50,000	6	SNR 0.1	N
Spike ideal	128	50,000	3	N	N
Spike noisy	128	50,000	3	SNR 0.01	Y
80S [50]	128	93,852	3.76875	Y	Y
RAG [1]	128	108,544	1.845	Y	Y
Spliceosome [17]	128	139,722	4.25	Y	Y
Linear1d	128	50,000	6	SNR 0.1	N

Table S2: Dataset statistics

B.1. Synthetic datasets

Synthetic datasets were created by following the standard image formation model: Given a ground truth voxel array, images were generated by rotating the volume by a rotation matrix R, where R is uniformly sampled from $SO(3)$, projecting the

volume along the z-axis, then shifting the resulting image by t, where t is uniformly sampled from $[-t_{ext}, t_{ext}]^2$ pixels. For the Hand dataset, the volume depicting a hand was manually generated on a 64^3 voxel array. For the Spike dataset, the volume was generated by simulating the electron scattering of PDB 6VYB [2] at 6 Åresolution using a grid spacing of 3 Å using the molmap command in UCSF Chimerax [**?**]. The volume was then padded to a final dimension of 128^3. For the Linear1d dataset, 50 volumes (D=128) sampled along a 1-D reaction coordinate were used as the ground truth. 1000 images were generated for each volume at a resolution of 12 Åand grid spacing of 6 Å/pixel. Image CTF parameters were sampled without replacement from EMPIAR-10028 [50] and applied to each image. Noise was added to each dataset to a signal to noise ratio (SNR) level as specified in Table S2, where we define the signal as the whole DxD image for Hand and Spike and a circle of radius 40 pixels for Linear1d. We note that the signal variance changes drastically depending on how much of the background is included in the SNR computation, thus we show example images for each of our synthetic datasets.

B.2. Real datasets

Real datasets for the *Plasmodium falciparum* 80S ribosome (80S)[50], RAG1-RAG2 complex (RAG)[1], and pre-catalytic spliceosome (Spliceosome)[17] were downloaded from EMPIAR at accession codes 10028, 10049, and 10180, respectively. We used the filtered datasets from Zhong *et al.* [54] available on Zenodo [53]. Images were downsampled to D=128 by clipping in Fourier space. A real space windowing function was applied to the images where the corners of the images are scaled to zero using a linear ramp from a radius of 85% to 95% of the image. We use previously published poses [53] as the "ground truth" poses for comparison. These poses were originally obtained via traditional refinement initialized from previously determined structures and thus include prior 3D information. As real datasets also lack a ground truth volume, to generate a reference volume for visual and quantitative comparison, we use the published poses[54] and train a cryoDRGN MLP with 3 hidden layers of width 256 for 30 epochs. We use the same architecture and amount of training to produce a volume that is more directly comparable to the cryoDRGN2 reconstruction (i.e. controls for model capacity).

C. Experimental setup

C.1. cryoSPARC

We use the cryoSPARC software package [35] as representative of state-of-the-art traditional, voxel-based reconstruction methods. CryoSPARC implements both an *ab initio* reconstruction algorithm using stochastic gradient ascent on the posterior probability distribution of the volume given the data to get an approximately correct low resolution structure, and an iterative refinement (E-M) algorithm which requires a roughly-correct initial model as input. In our baseline experiments, we perform *ab initio* reconstruction followed by homogeneous refinement using all default settings in cryoSPARC v2.15. We experimented with non-uniform refinement [36] in cryoSPARC, however it produced slightly worse pose errors and FSC resolutions for the 80S and RAG datasets and similar performance on the spliceosome dataset, and we report results from standard homogeneous refinement.

C.2. Pose-VAE

As a baseline for homogeneous reconstruction of synthetic datasets, we test a functional approach to pose prediction by training a VAE for SO(3)-valued pose variables. Briefly, we use the $S^2 \times S^2$ parameterization of SO(3) for the homeomorphic mapping of encoder outputs ($\mathbb{R}^6$) to an element of SO(3) and the modified KL-divergence described in [10]. A sample of pose from the approximate posterior is then used to transform a 2D coordinate lattice, which is then fed into the coordinate-based neural decoder to reconstruct the input image. We test on synthetic, centered images and use the VAE for inference of rotations only. As in our cryoDRGN experiments, networks are 256x3 MLPs with residual connections, and are trained for 30 epochs in minibatches of 8 images. We train on centered images.

C.3. Pose-GD

As a baseline for homogeneous reconstruction of synthetic datasets, we test the gradient-based optimization of pose variables. We randomly initialize pose variables (uniformly on SO(3)) and use a pre-trained cryoDRGN MLP on ground truth poses. We jointly update the volume and poses with backpropagation for 200 epochs in mini-batches of 512 images (to achieve more updates of poses). We train on centered images.

C.4. cryoDRGN-BNB

We compare against prior work which performed fully unsupervised reconstruction of cryoDRGN models [55]. We use the default settings for pose search and train the same 256x3 coordinate-based MLP architecture for 30 epochs with pose search

performed every 5 epochs. Unlike in Zhong *et al.* [55], we do not train on tilt-series pairs.

C.5. cryoDRGN2

Unless otherwise specified, all cryoDRGN2 neural networks are instantiated as fully connected neural networks with 3 hidden layers of width 256 and ReLU activations. We use residual connections between layer input and output to whiten gradients, i.e. $y(x) = ReLU(fc(x) + x)$.

In our cryoDRGN2 experiments, we train for 30 epochs total and perform pose search every 5th epoch (where image poses and the model are jointly updated). We additionally test resetting the model with random weights, retraining the model for 30 epochs using the last round of estimated poses, then repeating 30 more epochs of training with pose search every 5 epochs (cryoDRGN2+r). Training is performed in minibatches of 8 images using the Adam optimizer [19] with a learning rate of 1e-4.

For heterogeneous reconstruction, we jointly train an image encoder along with a generated model of 3D volume. Encoder/decoder architectures are 3 layer MLPs of width 256, and we use a 8-D latent variable. We train for 30 epochs with pose search performed every 5 epochs.

Training times varied across datasets. For homogeneous and heterogeneous experiments on the three EMPIAR datasets, training cryoDRGN2 took approximately 7-15 hours on a single V100 Nvidia GPU for 30 epochs of the above training schedule. For comparison, training cryoDRGN with fixed poses for 30 epochs ranged between 3-5 hours, and training cryoDRGN-BNB ranged between 18-38 hours on a single Nvidia V100 GPU.

D. Supplementary results – Ab initio homogeneous reconstruction

In this section, we provide more detailed experimental results for *ab initio* homogeneous reconstruction. To evaluate reconstruction quality, we focus on image pose error, which allows for direct comparisons of reconstruction accuracy between different model classes (voxel vs. neural). To compute the pose error, we first perform a 6-D rigid body alignment of the reconstructed volume into the frame of the reference volume and transform the predicted poses accordingly. Pose errors are reported as the square Frobenius norm between the reference and predicted rotation matrix, $||R_{ref} - R_{pred}||_F^2$, and the square L2 norm between translation vectors, $||t_{ref} - t_{pred}||_2^2$. Pose errors statistics are summarized in Table 2 and Table S3 for synthetic datasets and Table 3 for real datasets.

Reconstructed volumes for the hand dataset for all tested methods are shown in Figure S2, and the reconstructed volumes for the RAG and spliceosome datasets are shown in Figure 3.

We also quantify the correlation between the reconstructed volumes and the reference volume using Fourier Shell Correlation (FSC) curves [2]. In Table S6, we report the map-to-map FSC with a 0.5 criterion (*i.e.* the resolution at which the FSC curve falls below 0.5) between the reference and reconstructed volumes. In Figure S3, we provide the full FSC curves. In Table S5, we report the "gold standard" half-map FSC [39] with a 0.143 criterion, where we performed independent reconstructions on random half subsets of the dataset and compare the resulting volumes, i.e. "half-maps". In Figure S4, we provide gold standard FSC curves.

As described in Section C.5, we train the model in three stages: standard pose search interleaved with model updates (cyoDRGN2), we then reset the model and train on fixed poses from the last iteration (cryoDRGN2+r), followed by standard pose search interleaved with model updates (cryoDRGN2+r+ps). While low pose error is often achieved after the first stage, showing that poses may be accurately aligned on low resolution structures, further iterations are useful for converging the neural model. We note that the cryoDRGN MLP is a relatively small architecture (3 hidden layers of width 256; 296,193 parameters). On the 80S dataset, a large ribosomal complex, we repeated cryoDRGN 80S experiments with a larger MLP architecture (5 hidden layers of width 512; 1,510,913 parameters) (Figure S4,S3).

[2]To remain invariant to differences in absolute scale and normalization imposed by different algorithms and to capture resolution-dependent decay, cryo-EM volumes are typically compared as their correlation as a function of radial shells in Fourier space.

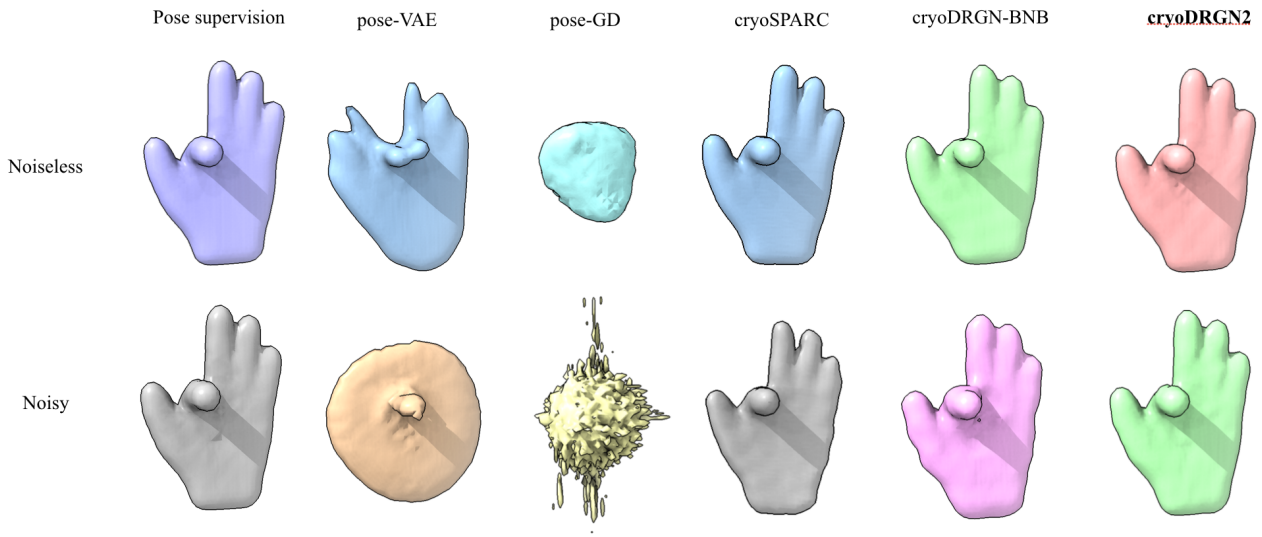

Figure S2: Reconstructed volumes of the synthetic Hand dataset from different homogeneous *ab initio* reconstruction algorithms. Noiselss - top row; Noisy - bottom row.

Method	**Hand**		**Spike**	
	SNR = inf	*SNR = 0.1*	*SNR = inf*	*SNR = 0.1*
Pose VAE	5.99/6.66	5.97/6.64	5.98/6.67	5.98/6.65
Pose GD	5.97/6.61	5.97/6.65	5.96/6.63	5.97/6.66
cryoSPARC	0.012/0.002	0.692/0.071	0.0007/0.0003	0.065/**0.002**
cryoDRGN-BNB	0.39/0.007	**0.06**/0.25	0.10/0.0006	0.066/0.012
cryoDRGN2	**0.002/0.0003**	0.086/**0.027**	**0.0004/0.0001**	**0.057**/0.011

Table S3: Homogeneous reconstruction pose accuracy quantified by mean/median rotation error between the predicted and the ground truth image poses.

Method	**80S** [EMPIAR-10028]		**RAG** [EMPIAR-10049]		**Spliceosome** [EMPIAR-10180]	
	R	t	R	t	R	t
cryoSPARC	**0.0186/0.0001**	**0.0008/0.0006**	3.7806/0.3084	**0.0096/0.0033**	0.0853/0.0015	0.1674/0.1663
cDRGN-BNB	0.6151/0.0020	0.0071/0.0030	4.1621/4.6371	0.0745/0.0775	2.2187/0.1854	0.0418/0.0172
cryoDRGN2	0.0578/0.0008	0.0022/0.0015	3.4254/0.0386	0.0324/0.0265	0.1958/0.0046	0.0094/0.0036
cryoDRGN2+r	0.0590/0.0008	0.0022/0.0015	**3.3730/0.0226**	0.0338/0.0311	**0.1947/0.0044**	**0.0093/0.0035**

Table S4: Homogeneous reconstruction pose accuracy on real cryo-EM datasets quantified by mean/median rotation (R) and translation (t) error between the predicted pose and the reference pose. ”cryoDRGN2+r” refers to model reset followed by additional pose refinement.

Method	Hand	Spike	80S	RAG12	Spliceosome
cryoSPARC	4.00	2.17	2.00	2.98	2.00*
cryoDRGN-BNB	4.27	2.21	2.06	4.57	3.88
cryoDRGN2	2.91	2.03	2.00	2.42	2.21
cryoDRGN2+r	N/A	N/A	2.00	2.21	2.06

Table S5: Quantitative comparison of the reconstructed volumes to the ground truth (synthetic) or reference volume (real) by an FSC=0.5 criterion. Lower values are better; the best possible is 2 pixels. "cryoDRGN2+r" refers to model reset followed by additional pose refinement. *CryoSPARC originally failed to produce the correct structure (resolution of 32 pix) before re-centering the images.

Method	80S	RAG12	Spliceosome
cryoSPARC	2.00	2.42	2.00*
cryoDRGN-BNB	2.00	4.26	2.51
cryoDRGN2	2.00	2.97	2.09
cryoDRGN2+r	2.00	2.13	2.00

Table S6: Resolution of the reconstructed volumes from *ab initio* homogeneous reconstruction assessed with the gold standard FSC=0.143 criterion. Lower values are better; the best possible is 2 pixels. "cryoDRGN2+r" refers to model reset followed by additional pose refinement. *The cryoSPARC reconstruction originally failed to produce the correct structure before re-centering the images.

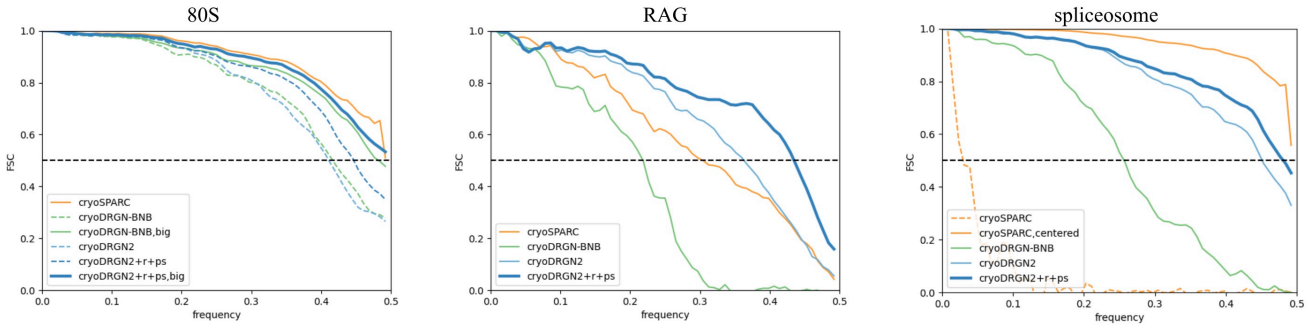

Figure S3: Map-to-map Fourier Shell Correlation (FSC) curves computed between the reference volume and the reconstructed volumes. "cryoDRGN2+r" refers to model reset and training on fixed poses from the last iteration. "cryoDRGN+r+ps" refers to model reset followed by additional iterations of pose search. On the 80S dataset, we show FSC curves after repeating *ab initio* reconstruction with a larger MLP architecture. We also show original cryoSPARC results on the spliceosome before image recentering.

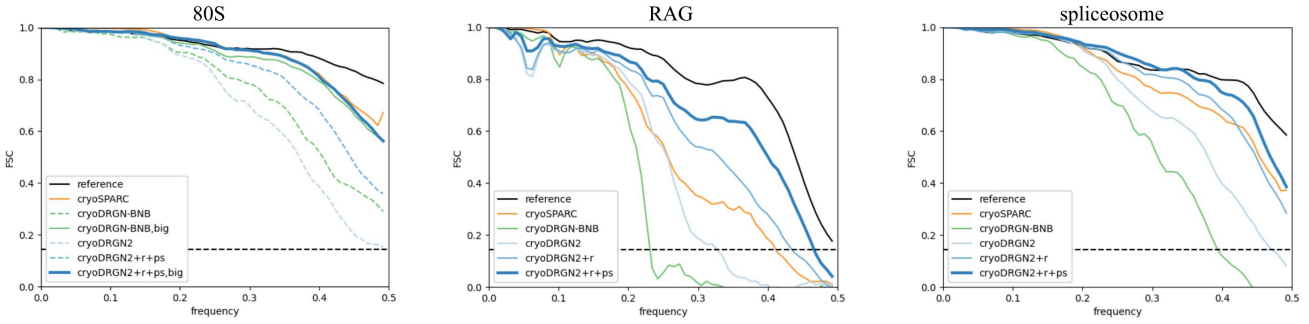

Figure S4: Gold standard Fourier Shell Correlation (FSC) curves from *ab initio* homogeneous reconstruction. "cryoDRGN2+r" refers to model reset and training on fixed poses from the last iteration. "cryoDRGN+r+ps" refers to model reset followed by additional iterations of pose search. On the 80S dataset, we show FSC curves after repeating *ab initio* reconstruction with a larger MLP architecture.

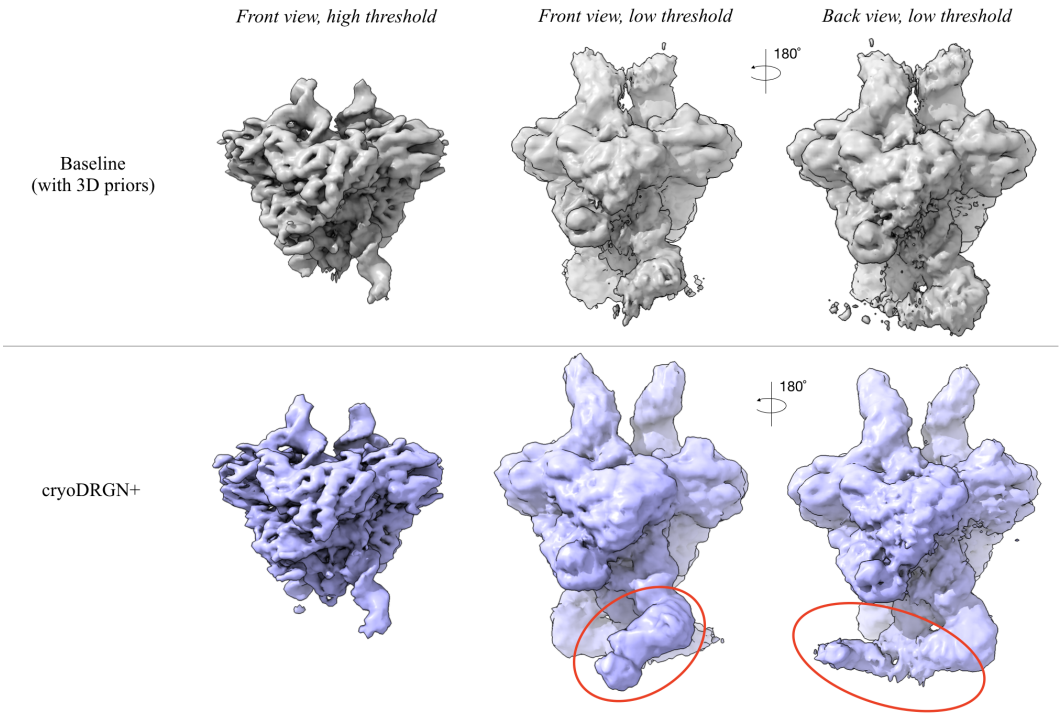

Figure S5: Additional views of the cryoDRGN2 reconstructed volume of the RAG dataset [1] and the baseline volume from traditional homogeneous refinement of the published volume. At a low threshold visualization of the volume, additional density of the DNA extensions is visible in the cryoDRGN2 volume (red circles), which are not resolved in the baseline structure.

E. Supplementary results - *Ab initio* heterogeneous reconstruction

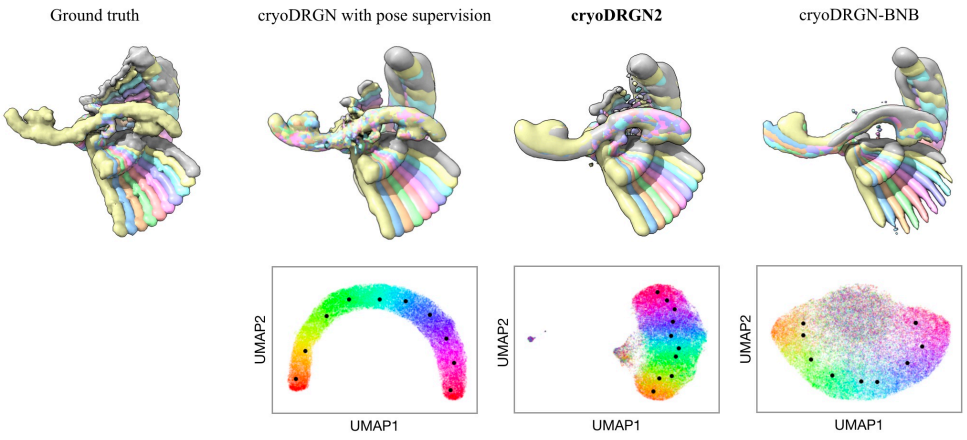

Figure S6: **Comparison of reconstruction algorithms on the Linear1d dataset.** *Top row:* Ten representative ground truth and reconstructed volumes along the 1D motion. *Bottom row:* UMAP visualization of the latent space embeddings of images from the dataset, colored by the ground truth reaction coordinate describing the motion.

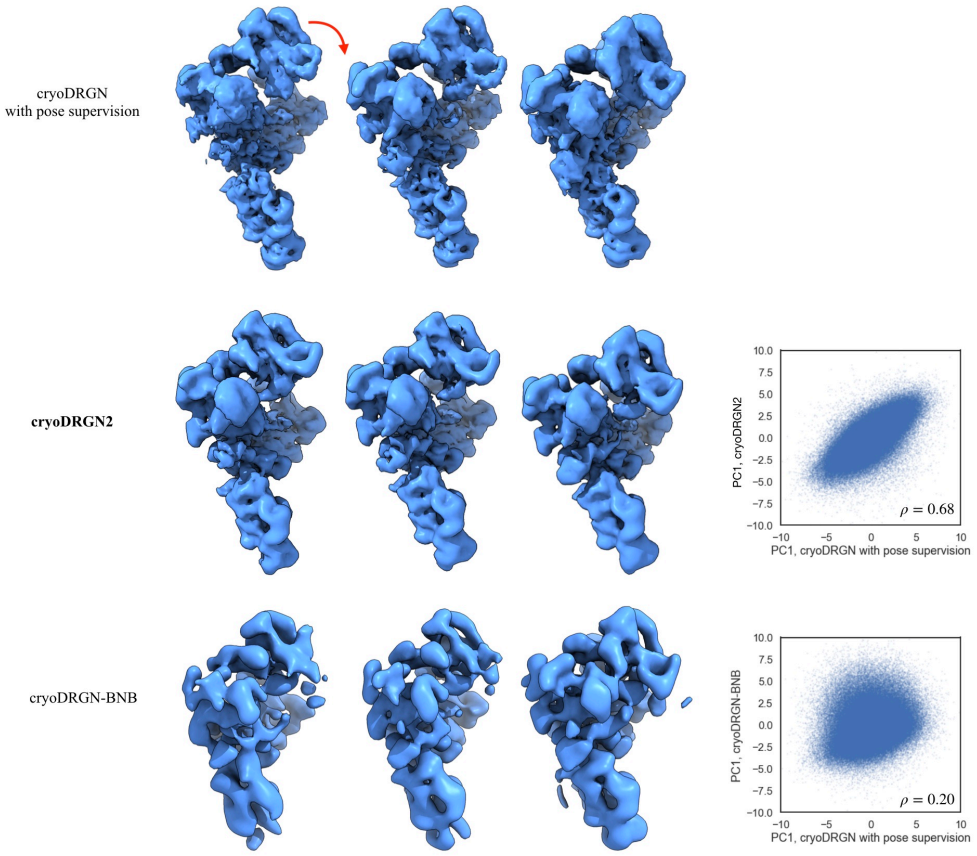

Figure S7: **Comparison of reconstruction algorithms on the pre-catalytic spliceosome dataset [EMPIAR-10180] [17].** Reconstructed volumes are generated along the first PC of the latent space embeddings and show a hinging motion of the complex (red arrow). Comparison of the latent embeddings from *ab initio* reconstruction to the previously published cryoDRGN reconstruction with pose supervision [53]. Spearman correlation noted.