

Reconstructing Heterogeneous Biomolecules via Hierarchical Gaussian Mixtures and Part Discovery

Supplementary Material

A Webpage

In addition to this pdf file in the supplement package, we share a webpage, [main.html](#), presenting short videos describing CryoSPIRE and showing qualitative comparisons with baseline methods on synthetic and experimental data. We use ChimeraX [29] to create detailed 3D visualizations of reconstructions shown in Figs. 12, 13, and 14. We also demonstrate how navigating the learned latent space leads to various structural states reconstructed by CryoSPIRE. Please open [main.html](#) in a web browser (best viewed in Google Chrome).

B Broader Impact

Cryo-electron microscopy (cryo-EM) has emerged as a revolutionary technique in structural biology, enabling the determination of macromolecular structures with significant societal impact. Computational methods, grounded in machine learning and computer vision have now been used to determine many thousands of biological structures. Notably, cryo-EM played a pivotal role in elucidating the structure of the SARS-CoV-2 spike protein, revealing its pre-fusion conformation and aiding in the assessment of medical countermeasures. Complementing computational methods such as AlphaFold for protein structure prediction, cryo-EM has revolutionized our understanding of cellular processes and accelerated the development of novel therapeutics, including synthetic antibodies. Nevertheless, we strongly condemn any usage of our proposed hierarchical 3D GMM representation for generating malicious data, improperly modifying signals, or spreading misinformation.

C GMM Image Formation, Parameterization and Rendering

In cryo-EM, the image formation process follows integral projection of a 3D density to the 2D image plane. For a 3D Gaussian mixture, the projection is analytically tractable, as described in Sec. 2, Eq. 2. For the purposes of optimization, however, we adopt a slightly different parameterization as we find it to be somewhat better behaved. Assuming each 3D Gaussian component in the mixture is parameterized with center $\mathbf{c} \in \mathbb{R}^3$, isotropic scale $s \in \mathbb{R}$, and an amplitude $m \in \mathbb{R}$, we define the 2D noise-free projection along the canonical z -axis, for location $\tilde{\mathbf{p}} \in \mathbb{R}^2$, as

$$\tilde{I}(\tilde{\mathbf{p}}) = \sum_i m_i \exp\left(-\frac{\|\tilde{\mathbf{p}} - [\mathbf{c}_i]_{xy}\|_2^2}{2s_i^2}\right) \quad (8)$$

Here, we modify the weight of terms such that the peak intensity of each Gaussian term solely depends on the amplitudes m_i , whereas, in Eq. 2, it is proportional to both s_i and m_i . This leads to a direction of ambiguity in the optimization landscape where increasing one parameter (e.g. m_i) and decreasing the other (e.g. s_i) can compensate; the coupling between m_i and s_i makes it challenging to set learning rates and it destabilizes the optimization dynamics. The alternative parameterization in Eq. 8 that we use is mathematically equivalent to the following weighting of the 3D mixture:

$$f(\mathbf{p}) = \sum_i \frac{m_i}{\sqrt{2\pi}s_i} \exp\left(-\frac{\|\mathbf{p} - \mathbf{c}_i\|_2^2}{2s_i^2}\right), \quad (9)$$

for location $\mathbf{p} \in \mathbb{R}^3$. This simply involves a change of variables, i.e., $m_i \rightarrow m_i/\sqrt{2\pi}s_i$.

As discussed in the introduction of the paper, cryo-EM images are extremely noisy, since low electron dosages are used to minimize radiation damage to the particles. To ensure sufficient image contrast, the microscope is defocused, which is modeled as convolution with a image-specific point-spread function (PSF), $g^{(n)}$, or, more commonly as modulation in the Fourier domain with a contrast transfer function (CTF) [41]. Finally, we model all sources of noise with additive, zero-mean Gaussian noise, $\epsilon^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. Taken together, the final image can be expressed as,

$$\hat{I}^{(n)} = g^{(n)} \star \tilde{I}^{(n)} + \epsilon^{(n)} \quad (10)$$

For optimization, we minimize a squared L2 reconstruction loss between model predictions and observed image, which is proportional to the negative log-likelihood,

$$\mathcal{L}(I^{(n)}, \hat{I}^{(n)}) = \|I^{(n)} - \hat{I}^{(n)}\|_2^2. \quad (11)$$

Implementation. Among the Gaussian parameters, the scale is constrained to be positive, and amplitudes are constrained to be non-negative. The centers have no hard constraints. To realize such constraints, we define s_i in the log-scale domain, $s_i = \exp(\tilde{s}_i)$. We also use a ReLU activation function to ensure that amplitudes are non-negative, i.e., $m_i = \text{ReLU}(\tilde{m}_i)$. Thus, the free parameters are $\{(\mathbf{c}_i, \tilde{s}_i, \tilde{m}_i)\}_i$.

The rendering equation (Eq. 8) defines a function I on the 2D plane. We discretize the function within a box $[-0.5, 0.5]^2$ using a regular $L \times L$ grid, denoted $\Lambda = \{(x_u, y_v)\}_{u,v=1}^L$. A naive way to evaluate I on is to compute the contribution of Gaussian terms separately per-point on the grid, which can be stored collectively in a $G \times L \times L$ matrix. Following [6], we can simplify that computation since 2D Gaussians can be expressed as a separable product of 1D Gaussians (on rows and columns):

$$\exp\left(-\frac{\|\mathbf{p}_{xy} - \tilde{\mathbf{c}}_{i,xy}\|_2^2}{2s_i^2}\right) = \exp\left(-\frac{\|\mathbf{p}_x - \tilde{\mathbf{c}}_{i,x}\|_2^2}{2s_i^2}\right) \exp\left(-\frac{\|\mathbf{p}_y - \tilde{\mathbf{c}}_{i,y}\|_2^2}{2s_i^2}\right). \quad (12)$$

All points within a single row or column in the grid share corresponding 1D Gaussian terms, which thus need to be computed only once. To this end, we decompose the 2D grid Λ into two 1D grids $\Lambda_x = \{x_u\}_{u=1}^L$ and $\Lambda_y = \{y_v\}_{v=1}^L$, and accordingly compute two $G \times L$ matrices, M_x and M_y ,

$$M_x(i, x_u) = \exp\left(-\frac{\|x_u - \tilde{c}_{i,x}\|_2^2}{2s_i^2}\right), \quad M_y(i, y_v) = \exp\left(-\frac{\|y_v - \tilde{c}_{i,y}\|_2^2}{2s_i^2}\right). \quad (13)$$

These matrices store the value of 1D Gaussian terms on 1D grids.

We also define all amplitudes within a $G \times 1$ matrix, denoted as W , with $W_{i,1} = m_i$. Next, we use fast matrix operations on W , M_x and M_y , to realize the above rendering equation (Eq. 8). We first compute vectorized outer product between W and M_x , yielding a new $G \times 1 \times L$ matrix. We then compute vectorized outer product between the resulting matrix and M_y , yielding a final matrix of size $G \times 1 \times L \times L$, followed by reducing the first dimension using summation. As a result, we obtain a $1 \times L \times L$ matrix which is our desired discretized projection. Importantly, compared to the naive process, this approach avoids redundant computations using separability of 2D Gaussians.

D CryoSPIRE Initialization

Before describing the initialization process of the coarse-grained GMM for part discovery, we outline how to obtain a preliminary density map as a rigid reconstruction, as well as per-image poses and CTF parameters. For CryoBench [15] datasets, the ground-truth CTFs and poses are provided for each particle image. Given the poses, we use a simple form of backprojection to compute an initial rigid reconstruction. For Ribosome experimental data [7] (EMPIAR-10076), we estimate per-particle poses and an initial rigid reconstruction using CryoSPARC [34]. For Spliceosome [30] (EMPIAR-10180), an initial rigid reconstruction and per-particle poses are available (computed using RELION [37]). For both experimental datasets, CTFs are estimated in a standard preprocessing stage.

As demonstrated in the Gaussian Splatting literature [16], the optimization dynamics are highly sensitive to the initial values of Gaussian parameters. For the coarse-grained GMM used for part discovery, we use the input rigid reconstruction to seed $G = 2048$ Gaussians. Given the input density map, we discard voxels with density below a user-defined threshold, and then sample G of the remaining voxels with probability proportional to density. Gaussians are then seeded at those positions. Note that, since flexible regions often exhibit lower density in the rigid reconstruction, setting the threshold too high may exclude these regions, resulting in poor initialization. Finally, we initialize the amplitude and scales to user-defined values, $m_i = 0.15$, $s_i = 0.02$.

To initialize the hierarchical model, we need to determine the anchors. We run k-means++ [1] clustering on the learned Gaussian features from the part discovery model. UMAP visualizations (Figs. 4, 5, 6, 7, 8) indicate that the feature space contains well-separated clusters of Gaussian components. This clear distinction between clusters provide the potential to automatically set the

number of anchors, yet for now we choose number of anchors manually (eg with UMAP visualization). Each anchor receives a feature vector that is set to the centroid of its features in the cluster, and its position is set to that of the closest Gaussian component in the feature space. Feature-based clustering is followed by spatial clustering for IgG data to further divide clusters into local regions, improving coverage of the density map. Furthermore, the part discovery model provides an improved density map, where amplitude modulation helps to identify unused components that can be discarded. Based on this density map, we follow a similar procedure to seed a denser set of Gaussian components. Each new Gaussian is connected to the anchor that the closest Gaussian in the part discovery model is assigned to. Feature offsets are initialized to zero, so Gaussians initially inherit the features of their anchors. We initialize amplitudes as $m_i = 0.15$. But we use lower initial scale values $s_i = 0.01$, since this initialization uses a high quality, more reliable density map.

E FSC-based Performance Metrics

Evaluation of cryo-EM methods is very challenging, in part because ground truth 3D density maps do not exist for experimental data. To date, the most widely used metric, namely, Fourier Shell Correlation (or FSC), is a measure of consistency, defined as the normalized cross-correlation of two 3D density maps computed as a function of frequency [12, 43]. Given two 3D density maps A and B with Fourier coefficients A_j and B_j , the FSC at wavelength λ is defined as

$$\mathcal{F}_\lambda = \frac{\sum_{j \in S_\lambda} A_j B_j^*}{\sqrt{\sum_{j \in S_\lambda} |A_j|^2 \sum_{j \in S_\lambda} |B_j|^2}}, \quad (14)$$

where S_λ is the set of Fourier indices for frequencies within a spherical shell with wavelength λ centered at the origin. Here, $|z|$ is the modulus of the complex scalar z , and z^* is the complex conjugate of z .

FSC can measure consistency between an estimated density map and a ground truth map. Signal-to-noise ratios in particle images, and hence in 3D reconstructions will decrease with frequency. So FSC is usually close to 1 at low frequencies, and then decrease toward zero at higher frequencies where observations are missing or dominated by noise. As a consequence, there are two common ways to characterize the quality of 3D reconstructions. One is the area-under-the-curve (AUC) of the FSC curve. The other is a resolution of the map, defined as the wavelength at which FSC drops below a threshold. A threshold of 0.5 is used for FSC computed between an estimated map and a ground truth map [36], which corresponds to a SNR of 1 (an estimator of spectral SNR from FSC is simply $|FSC_\lambda| / (1 - |FSC_\lambda|)$ [9, 28, 42]). Since we do not generally have ground truth density maps for real experimental datasets, FSC between ground truth and estimated maps is typically only used with synthetic data, e.g. CryoBench [15].

For homogeneous reconstruction with real experimental data, where ground truth is unavailable, FSC is used as a measure of consistency between two 'independent' 3D reconstructions. In the "gold standard" protocol, one randomly divides the set of particle images into halves, from which one estimates two 3D density maps that we assume are conditionally independent given the true density. FSC then provides a measure of consistency as a function of frequency between the two half-maps. When using the gold standard FSC protocol (comparing two half-maps estimated from independent halves of the data), a threshold of 0.143 is used to measure the map resolution, which approximately corresponds to when the spectral SNR of a map computed from the full set of images is 1 [36].

Despite the widespread use of FSC with the gold standard protocol, there is no accepted extension to heterogeneous reconstruction. The best existing benchmark, namely CryoBench [15], leverages synthetic data, in which case one has ground truth density maps available. To handle heterogeneity, they define the Per-Conformation FSC metric as the average AUC-FSC or FSC curve taken over all ground truth states generated in the synthetic dataset, whether compositional or conformational in its heterogeneity. This is the performance metric reported in Table 2 and Figure 3. A closely related measure is the Per-Image FSC, which is the mean AUC-FSC between the ground truth density maps and those estimated from the latent state by taking one sample particle per ground-truth state. For completeness we report this below in Table 4 and visualize FSC curves in Fig. 10. Due to the high memory requirements, we were unable to reproduce FSC curves for RECOVER by the time of submission. For more details on RECOVER performance, see the CryoBench paper [15].

Method	IgG-1D		IgG-RL		Ribosembly	
	Mean (std)	Med	Mean (std)	Med	Mean (std)	Med
3D Classification [39]	0.297 (0.019)	0.291	0.309 (0.01)	0.307	0.289 (0.081)	0.288
CryoDRGN [47]	0.351 (0.028)	0.356	0.331 (0.016)	0.333	0.412 (0.023)	0.415
CryoDRGN-AI-fixed [19]	0.364 (0.002)	0.364	0.348 (0.012)	0.350	0.372 (0.032)	0.375
3DFlex [33]	0.335 (0.003)	0.335	0.337 (0.007)	0.337	-	-
3DVA [32]	0.349 (0.004)	0.350	0.333 (0.014)	0.335	0.375 (0.038)	0.375
RECOVAR [10]	0.386 (0.001)	0.388	0.363 (0.011)	0.363	0.429 (0.018)	0.432
CryoSPIRE (Ours)	0.396 (0.013)	0.400	0.375 (0.020)	0.391	0.422 (0.015)	0.421

Table 4: Mean AUC of **Per-Image FSC** is reported for various methods on CryoBench datasets [15]. In parentheses, we report the standard deviation indicating the spread of AUC among different structural states (100 states for IgG-1D and IgG-RL and 16 states for Ribosembly). FSC curves are computed after masking out background noise. We AUC numbers from CryoBench [15] for RECOVAR. (Best method in bold, second best underlined.)

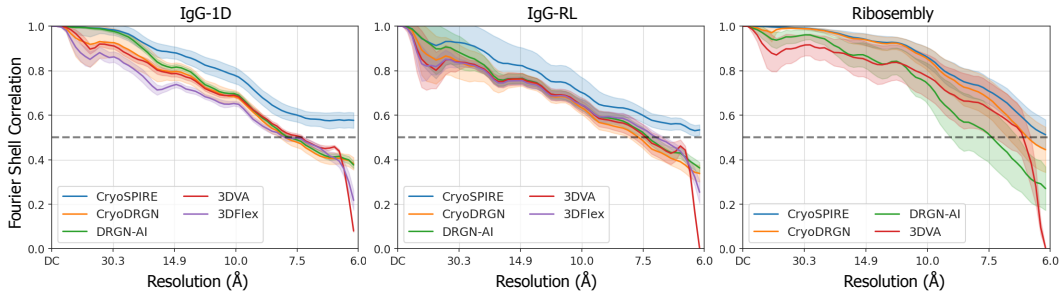


Figure 10: Per-Image FSCs on CryoBench datasets. Error bars indicate standard deviation across different states. The highest possible resolution is 6 Å on these synthetic datasets.

136 F Complete Qualitative Result on Ribosembly

137 While Figure 6 in the main body of the paper shows example compositional states reconstructed by CryoSPIRE, Figure 11 shows all 16 states.

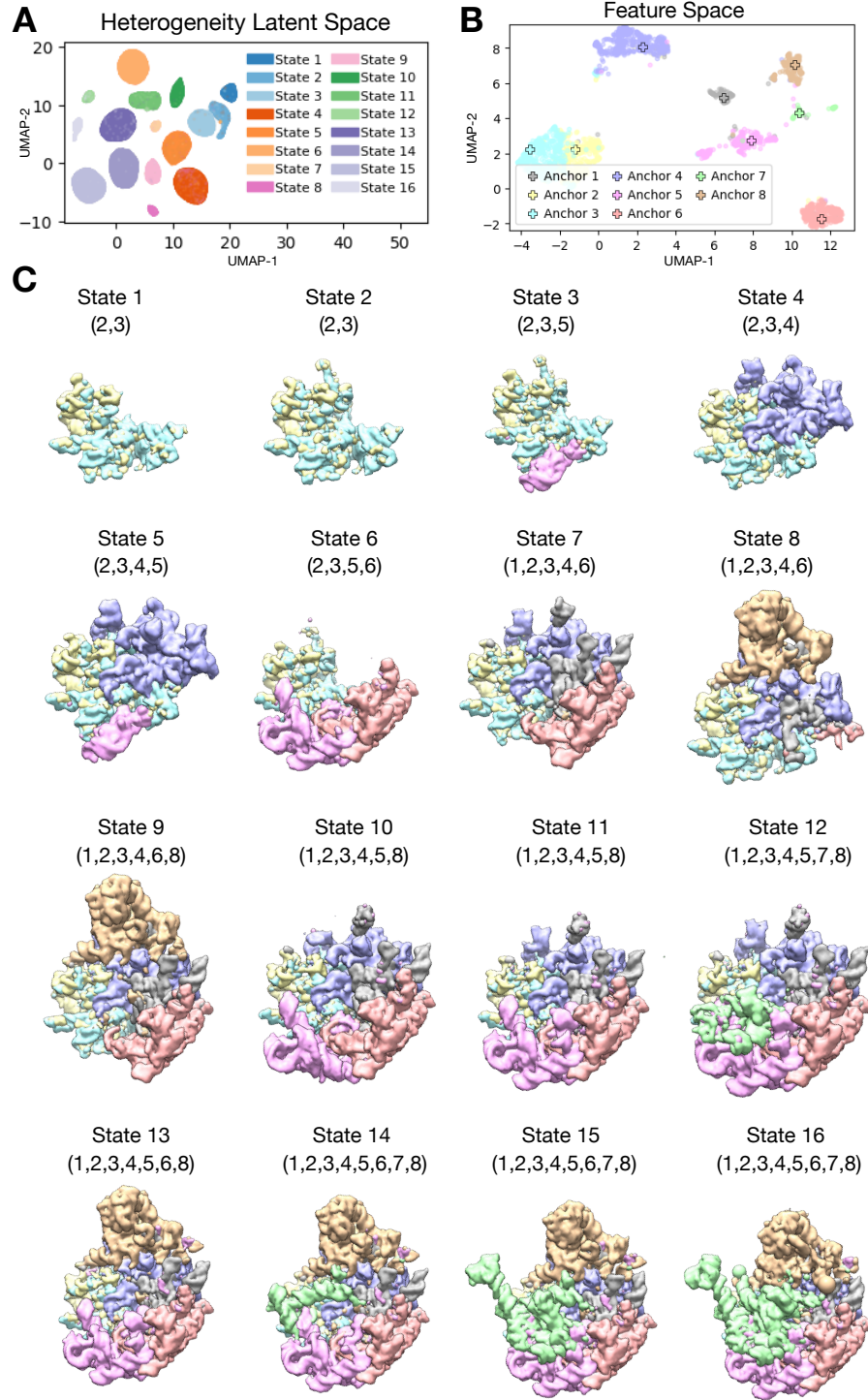


Figure 11: Complete qualitative results on Ribosembly [15] (A) Gaussian feature space, \mathcal{F} , showing eight major parts identified through clustering. (B) Heterogeneity latent space, \mathcal{Z} , colored coded with the ground-truth compositional state. (C) Visualizations of all 3D density maps corresponding to 16 compositional states, with colors depicting parts (given in parentheses).

138

139 G Qualitative Comparisons on CryoBench Data

140 In Figures 4, 5 and 6, we showed qualitative result of CryoSPIRE on CryoBench synthetic datasets of
 141 IgG-1D, IgG-RL and Ribosembly, respectively. Here, in Figure 12, we compare CryoSPIRE with
 142 the state-of-the-art methods 3DFlex [33], 3DVA [32], CryoDRGN [47], DRGN-AI [19], and with
 143 ground-truth structural states. Please find detailed 3D visualizations of the reconstruction in the
 144 webpage.

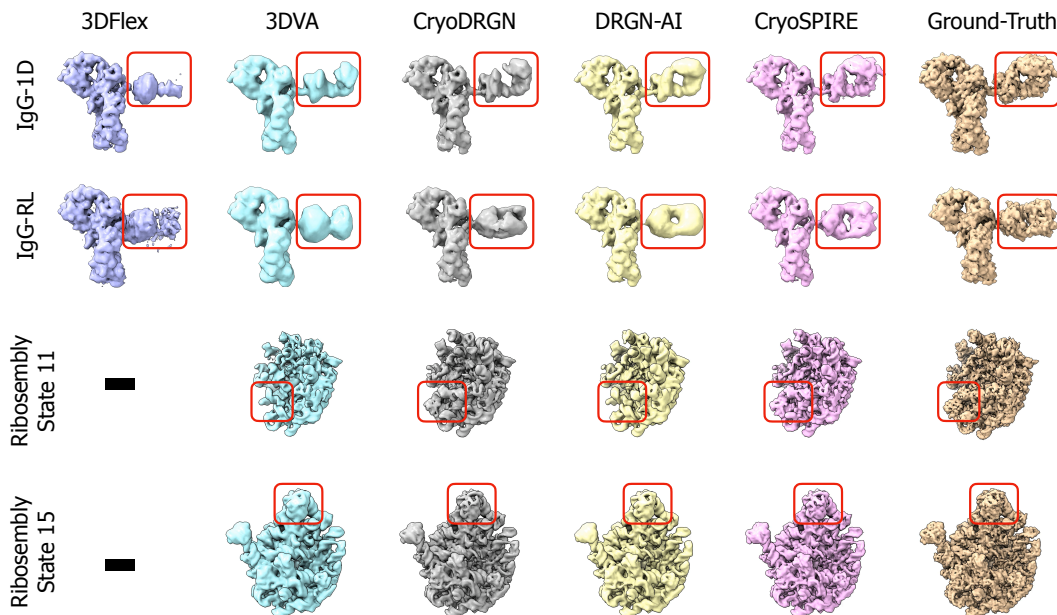


Figure 12: Qualitative comparison of CryoSPIRE with four state-of-the-art methods on CryoBench synthetic datasets [15]. Last column corresponds to the ground-truth state. In the first two rows, reconstruction of all methods for a sample conformational state is provided for IgG-1D and IgG-RL, demonstrating that our method outperforms others in recovering higher frequency details in the Fab domain (highlighted in red). For Ribosembly, we provide reconstructions of two example compositional states (labeled as 11 and 15). Since 3DFlex is limited to conformational heterogeneity, it is not evaluated on this dataset. Reconstructions of the state 11 by DRGN-AI and 3DVA clearly miss a subunit, while CryoSPIRE is able to capture it. Moreover, for state 15, DRGN-AI and 3DVA are overall less detailed while CryoSPIRE appears slightly better than CryoDRGN. 3D visualizations of the above reconstructions are presented in the webpage.

H Qualitative Comparisons on Experimental datasets

Figures 7 and 8 in the main body of the paper showed qualitative result of CryoSPIRE on Large Ribosomal Subunit (EMPIAR-10076 [7]) and Pre-Catalytic Spliceosome (EMPIAR-10180 [30]), respectively. Here, in Figs 13 and 14, we show comparisons with state-of-the-art methods 3DVA [32] and CryoDRGN [47] and 3DFlex [33]. As these Ribosome data mainly exhibit compositional heterogeneity, 3DFlex is not evaluated on this dataset. For both 3DVA and 3DFlex, we use CryoSPARC v4.4.0 [34] with default setting. We use the default 3 variability components for 3DVA, while for 3DFlex, we run *3DFlex Training Job*, followed by *3DFlex Reconstruction* to obtain a high-resolution canonical structure. For CryoDRGN, we use the final result provided by the authors. Please find detailed 3D visualizations of the reconstruction in the webpage.

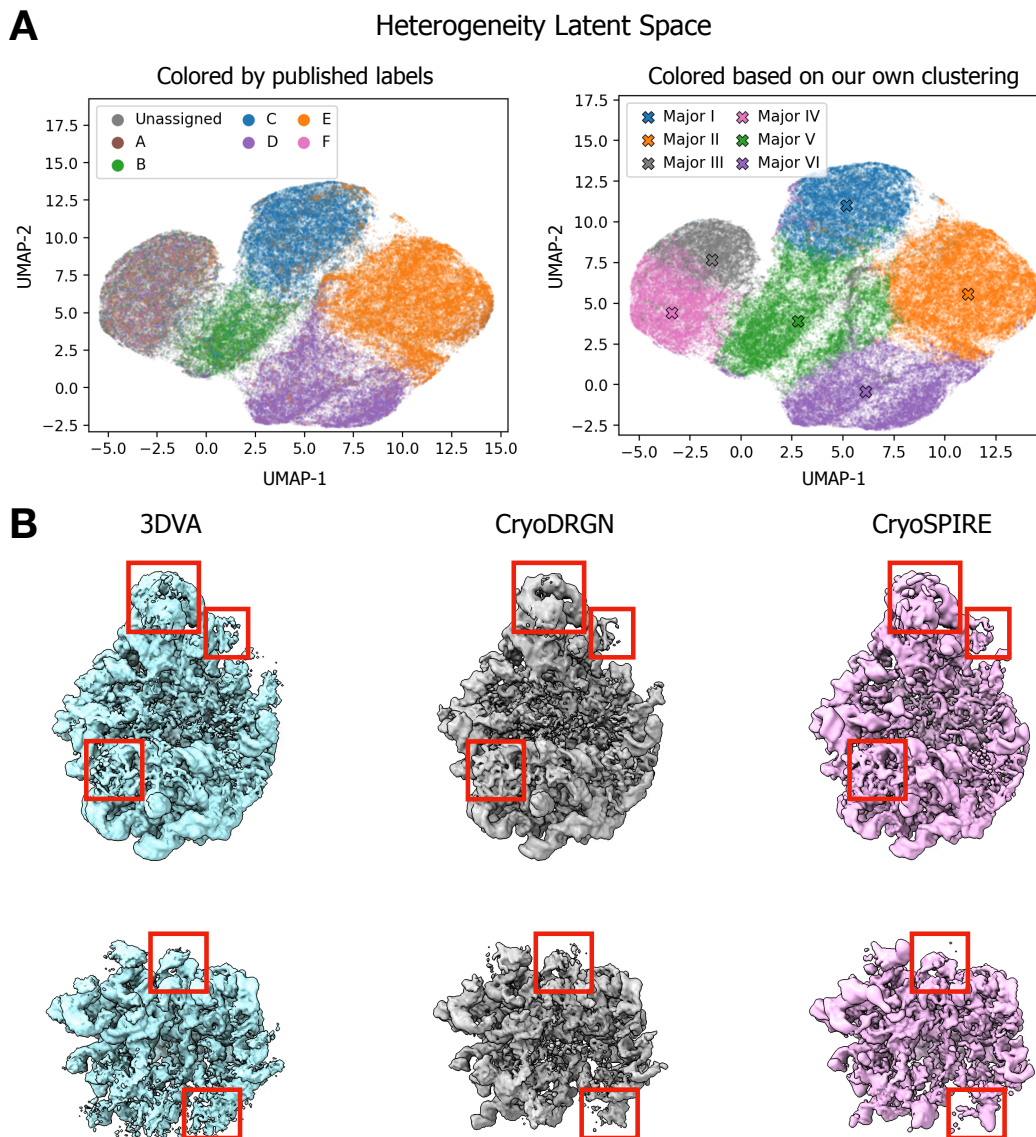


Figure 13: More qualitative result on Large Ribosomal Subunit (EMPIAR-10076 [7]) (A) Heterogeneity latent space, with latent points colored based on the published labels [7] (left) and colored based on our own clustering of latent space (right). (B) Qualitative comparison of CryoSPIRE with 3DVA [32] and CryoDRGN [47]. Two rows shows two of major assembly states. We identify some areas with red rectangles that shows main discrepancies between different methods. 3D visualizations of the above reconstructions are presented in the webpage.

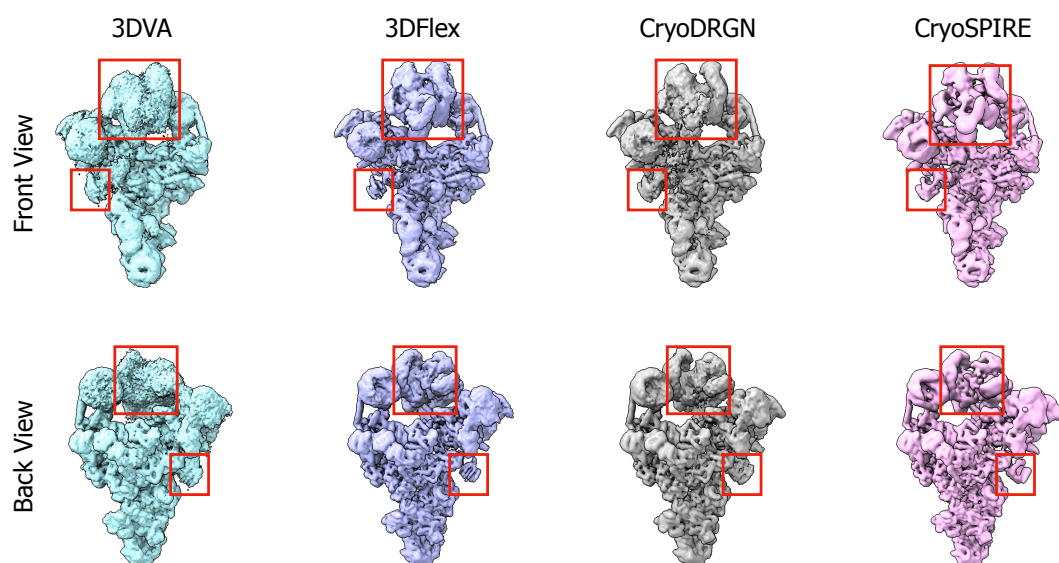


Figure 14: Qualitative comparison of CryoSPIRE with 3DVA [32], 3DFlex [33] and CryoDRGN [47] on Pre-Catalytic Spliceosome (EMPIAR-10180 [30]). Two rows show front and back views of the reconstructions, respectively. In both views, we mark the SF3b and a peripheral subunit of helicase with red rectangles. The reconstructions obtained by 3DVA, 3DFlex and CryoDRGN contain high-frequency noise within the two marked areas, whereas our method is better in resolving corresponding regions. 3D visualizations of the above reconstructions are presented in the webpage.