Mixture of neural fields for heterogeneous reconstruction in cryo-EM

Axel Levy[*]
Stanford University
axlevy@stanford.edu

Rishwanth Raghu[*]
Princeton University
rraghu@princeton.edu

David Shustin[*]
Princeton University
dshustin@princeton.edu

Adele Rui-Yang Peng
Princeton University
adelep@princeton.edu

Huan Li
Columbia University
hl3170@columbia.edu

Oliver Biggs Clarke
Columbia University
oc2188@cumc.columbia.edu

Gordon Wetzstein
Stanford University
gordon.wetzstein@stanford.edu

Ellen D. Zhong
Princeton University
zhonge@princeton.edu

Abstract

Cryo-electron microscopy (cryo-EM) is an experimental technique for protein structure determination that images an ensemble of macromolecules in near-physiological contexts. While recent advances enable the reconstruction of dynamic conformations of a single biomolecular complex, current methods do not adequately model samples with mixed conformational and compositional heterogeneity. In particular, datasets containing mixtures of multiple proteins require the joint inference of structure, pose, compositional class, and conformational states for 3D reconstruction. Here, we present Hydra, an approach that models both conformational and compositional heterogeneity fully *ab initio* by parameterizing structures as arising from one of K neural fields. We employ a new likelihood-based loss function and demonstrate the effectiveness of our approach on synthetic datasets composed of mixtures of proteins with large degrees of conformational variability. We additionally demonstrate Hydra on an experimental dataset of a cellular lysate containing a mixture of different protein complexes. Hydra expands the expressivity of heterogeneous reconstruction methods and thus broadens the scope of cryo-EM to increasingly complex samples. Webpage: https://hydra.cs.princeton.edu

1 Introduction

Structural information is key to understanding the function of macromolecular complexes, making protein structure determination a crucial tool in basic structural biology and rational drug design. Among experimental structure determination methods, cryogenic electron microscopy (cryo-EM) is unique in its capability to reveal dynamic information about large macromolecular complexes in near-native states.

[*] Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1816

In single particle cryo-EM, a set of biomolecular complexes (i.e. particles) is flash-frozen and imaged with a transmission electron microscope. Each collected image consists of a noisy, randomly oriented projection of an individual particle of unknown identity or composition, frozen in an unknown state. Reconstruction algorithms processing these data without information from upstream algorithms are called *ab initio heterogeneous* reconstruction methods. Classical reconstruction techniques use a Bayesian approach to optimize a finite number of voxel-based representations [50, 45]. These algorithms enabled biologists to process datasets made of a mixture of different protein compositions and, thus illuminate the molecular details of fundamental biological processes. These methods, however, tend to aggregate all the proteins of the same composition in a *single class*, despite these proteins being trapped in different conformational states, due to thermal fluctuations prior to the freezing step. Another line of work has extensively studied the possibility of reconstructing continuous motion from cryo-EM datasets, using linear combinations of voxel arrays [43], neural-based representations [68, 69, 25, 26], Gaussian mixture models [5, 20], or combining a voxel array with a flow field [44]. These methods sometimes leverage a structural prior about proteins (e.g., proteins are made of a fixed number individual atoms) and enable the reconstruction of molecular motion at high resolution. However, none of these methods can handle datasets containing different types of proteins, thereby strongly limiting their application, especially in the context of *in situ* cryo-EM.

As of today, there exist no approaches to simultaneously reveal compositional (discrete) and conformational (continuous) heterogeneity in cryo-EM datasets, potentially limiting the structures that can be revealed from the data. Unraveling this information poses a nontrivial problem that sequential strategies cannot solve. Due to their low signal-to-noise-ratio, cryo-EM images cannot be clustered depending on the type of protein they contain prior to the reconstruction. Because orientations are unknown, the problem cannot be solved by handling compositional heterogeneity before conformational heterogeneity as consensus poses (i.e., orientations) are inaccurate for large motions.

Here, we introduce Hydra, a neural-based method for *ab initio* heterogeneous reconstruction in cryo-EM. Inspired by the success of implicit neural representations in cryo-EM, we extend the neural field representation of DRGN-AI [26] with a mixture model of K neural fields. Using a new likelihood-based loss function, we simultaneously optimize orientations, conformations and class assignments and circumvent the pitfalls of sequential approaches. We demonstrate that our method allows neural-based methods to handle strong compositional heterogeneity and enables the simultaneous reconstruction of compositional and conformational heterogeneity with state-of-the-art accuracy. In an experimental cryo-EM dataset of a cell lysate mixture, we reveal three compositional states in a single pass, fully *ab initio*. We therefore make the following contributions:

- We develop a mixture of neural fields model for *ab initio* heterogeneous reconstruction in cryo-EM;
- We demonstrate that our method improves the expressiveness of neural-based methods for handling strong compositional heterogeneity;
- We enable simultaneous reconstruction of conformational and compositional heterogeneity beyond the limits of existing methods;
- We demonstrate the reconstruction of multiple protein complexes from an experimental dataset of an unpurified sample.

2 Related Work

2.1 Heterogeneous Reconstruction in Cryo-EM

Discrete Variability. Cryogenic electron microscopy offers the potential to reveal the structure of macromolecules in heterogeneous samples, where multiple types or multiple conformations are mixed together [24]. The first reconstruction algorithms handling 3D variability modeled the set of particles as a finite set of static structures. RELION [50] popularized the Bayesian approach to tackle heterogeneous reconstruction and the later introduced *multi-body refinement* tool [34] offered the possibility to segment static density maps and model continuous motion as a combination of rigid transformations. The software suite was recently improved with the Blush regularization tool [22, 4], leveraging a data-driven prior to enable the reconstruction of small protein-nucleic acid complexes (≤ 40 kDa). CryoSPARC [45] accelerated the inference with stochastic gradient descent and, in the

https://doi.org/10.52202/079017-1816

3DVA [43] extension, modeled continuous motion with a linear combination of density maps. These methods represent the state of the art for cryo-EM reconstruction but do not have the capability to represent complex non-linear motion and tackle the discrete heterogeneity (e.g., with "multi-class *ab initio*") before the continuous variability (e.g., with 3DVA or [21]), leading to inaccurate pose estimation for states exhibiting large motion.

Non-Linear Methods for Heterogeneous Reconstruction. In the past years, significant progress has been made in revealing non-linear dynamics from cryo-EM data. Manifold embedding [11] and Laplacian methods [33] are among the first attempts to model non-linear continuous motion but have only been applied to a small number of macromolecular complexes [8, 7]. HEMNMA [20] used a decomposition over the low-energy normal modes of an atomic model, thereby leveraging a prior over the structure and the dynamics of macromolecules. Herreros et al. [17] proposed the use of 3D Zernike polynomials and eliminated the need for pseudo-atomic models. CryoDRGN [69] used neural networks to continuously represent deformable density maps as well a variational auto-encoding framework to estimate conformational states. E2GMM [5], cryoFold [71], and DynaMight [51] followed this encoder-decoder framework and used Gaussian mixture models to represent density maps, thereby reducing the memory footprint of previous non-linear methods. 3DFlex [44] introduced the use of a parametric flow field to smoothly deform canonical 3D density maps, leveraging the knowledge that energetically favorable deformations tend to preserve the local geometry of proteins. Although these methods demonstrated the ability to reconstruct molecular motions and heterogeneity, they all need a coarse initialization of the density map, or the poses to be provided by an upstream reconstruction algorithm. In practice, this *ab initio* step is error prone in the presence of large conformational motions.

***Ab Initio* Heterogeneous Reconstruction.** Recent works investigated the problem of reconstructing an ensemble of density maps where poses are unknown. The preliminary cryoDRGN method [68] tackled the *ab initio* reconstruction problem by combining traditional pose search algorithms with the encoder-decoder neural-based architecture, which was refined in cryoDRGN2 [70]. Multi-CryoGAN [15] sidestepped pose estimation by casting the reconstruction problem as a distribution matching problem and successfully revealed 3D variability in synthetic cryo-EM datasets. Rosenbaum et al. [49] showed that the auto-encoding framework could be applied to jointly estimate poses and conformations of atomic models, and Levy et al. [25] demonstrated a fully autoencoding framework for *ab initio* heterogeneous reconstruction on real benchmark datasets. DRGN-AI [26] recently introduced a hybrid pose search strategy combined with an autodecoding architecture to handle low-signal datasets and demonstrated *ab initio* heterogeneous reconstruction on challenging datasets containing a significant number of junk images. These methods extend neural-based reconstruction to scenarios where poses cannot be reliably estimated by any upstream algorithms, but have a limited capability to represent mixtures of biomolecular complexes, due to the limited capacity of a single neural representation.

Here, we propose to represent the landscape of accessible structures via an ensemble of K neural networks, each *specialized* in representing the variability within a single compositional state. By jointly estimating structures and orientations, our method handles datasets that exhibit strong compositional heterogeneity and large motions.

2.2 Neural Fields for Large and Dynamic Scenes

Dynamic Scenes. In graphics, neural networks have also been used as light-weight, differentiable representations of continuously defined signals (e.g., occupancy fields [31], signed distance functions [37, 6], surface light fields [66], latent representation of appearance [52, 53], radiance fields [32, 1], and light fields [54]). In order to handle time-dependency and represent dynamic scenes, two approaches have been explored, as described in [39]. Similar to the cryo-EM method 3DFlex [44], *deformation-based* approaches apply a spatially-varying deformation to some canonical radiance field [38, 42, 59]. Methods following this strategy recover detailed dynamic scenes but are unable to model any topological variations. Similar to cryoDRGN [69], *modulation-based* approaches directly condition the radiance field of the scene on some latent vectors encoding temporal or dynamic information [13, 27, 65, 12]. These techniques are capable of modeling arbitrary deformations, topological changes, and other complex phenomena, but require additional regularization strategies

https://doi.org/10.52202/079017-1816

to avoid trivial, non-plausible solutions. HyperNeRF [39] was introduced as a combination of both approaches and enabled photo-realistic reconstruction of dynamic scenes with topological changes.

Large-Scale Scenes. Due to their limited capacity, neural networks have essentially been used to represent single objects or small-scale scenes. Several methods [46, 47] have looked into the possibility of using an ensemble of neural fields to represent large scenes such as a neighborhood in the city of San Francisco [57]. Another approach is to provide extra capacity with a coarse 3D grid of latent codes [27, 55, 40] or a block-coordinate multi-scale decomposition [28]. These works focus on static reconstruction, and if dynamic objects are present in the scene, these are simply masked out [57]. Ost et al. [36] enabled the representation of complex, dynamic multi-object scenes by decomposing them into their static and dynamic parts and learning one neural radiance field per dynamic object.

Here, we use the *modulation-based* approach to handle the continuous motion of proteins and an ensemble of neural fields to increase the representational capacity of our model, allowing it to reconstruct compositional mixtures of proteins. In contrast with the works cited above, we need to estimate which compositional state each image belongs to – which is done using a variational approach – while simultaneously optimizing the ensemble of neural networks and individual image poses.

2.3 Adaptive Mixtures of Experts

A mixture of experts (MoE) model uses an ensemble of neural networks to process a given input [19]. Doing so, the input space can be partitioned into subspaces on which only one of the neural networks needs to become “expert”. The mechanism with which the input space gets partitioned is usually referred to as the *gating* mechanism. Masoudnia and Ebrahimpour [29] partitions MoE approaches into two groups, depending on the gating mechanism and both how and when this mechanism is involved: (1) mixtures of explicitly localized experts (MELE) cluster the input data before the experts' training phase starts while (2) mixtures of implicitly localized experts (MILE) jointly optimize the expert networks and the gating mechanism. In cryo-EM, clustering methods operating directly on images are usually ineffective because of the high level of noise and because of the presence of other nuisance variables (orientation and conformation), making the MELE approach irrelevant.

Jacobs et al. [19] examined the use of different error functions (or loss functions) to optimize the expert networks and the gating mechanism, and the best performance was obtained using the negative log-probability of a Gaussian mixture model [35]. Several architectures have been explored, and one of the most applied is the mixture of MLP-experts (MME) [63, 9], where multi-layer perceptrons (MLPs) are used for both the experts and the gating networks. Our approach can be viewed as an application of the MILE method where, as in Jacobs et al. [19], the error function corresponds to the negative log-probability of a Gaussian mixture model. Due to the high level of noise, the gating mechanism ignores the content of input images and relies on an autodecoding framework.

3 Methods

In this section, we first describe the image formation model in single particle cryo-EM (3.1). We then define a 3D variability model that models both compositional and conformational heterogeneity (3.2). Based on these two first sections, we describe our system with a latent variable model (3.3) and explain our optimization method (3.4). The method is schematically described in Figure 1.

3.1 Image Formation Model

In cryo-EM, a purified solution of macromolecules is flash-frozen inside a thin layer of vitreous ice. Each molecule gets trapped in a random orientation with respect to the microscope and in a random conformational state, approximately following the Boltzmann distribution at ambient temperature [2]. The sample is exposed to an electron beam and individual 2D projections are extracted from micrographs via a step known as “particle picking”. The reconstruction task therefore starts with a given set of N images (or particles). Each image I_i can be modeled as [60, 50]

$$I_i = C_i * \mathcal{P}_{\phi_i} V_i + \eta_i, \tag{1}$$

https://doi.org/10.52202/079017-1816

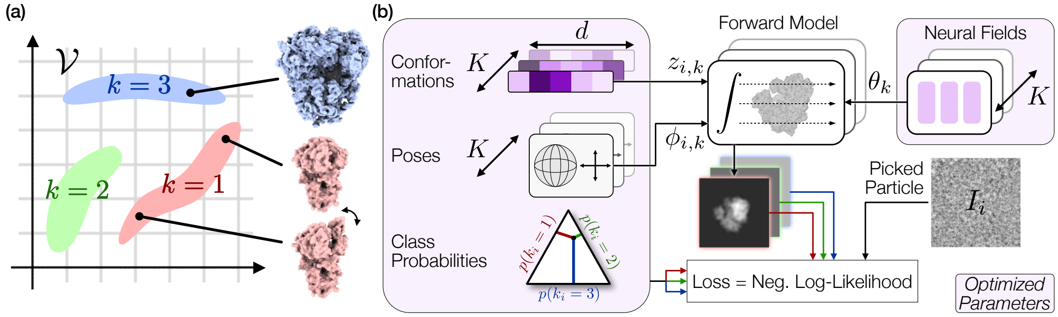


Figure 1: Overview of Hydra. (a) Schematic representation of the space of energetically plausible density maps in a heterogeneous cryo-EM dataset. We approximate this space with a finite union of low-dimensional manifolds. The compositional states (or classes) are labeled by k. The "conformation" within class k refers to intrinsic coordinates within the k-th manifold. **(b)** Optimization pipeline. The conformations, poses, class probabilities and neural fields are optimized such as to maximize the likelihood of the observed images ("picked particles") under the model described in Section 3.3.

where C_i models the the Contrast Transfer Function (CTF), V_i is a scalar 3D field ($\mathbb{R}^3 \to \mathbb{R}$) known as the "electron scattering potential" or the "density map", $\phi_i = (\mathbf{R}_i, \mathbf{t}_i)$ is a "pose" ($\mathbf{R}_i \in \text{SO}(3)$, $\mathbf{t}_i \in \mathbb{R}^2$) and $\mathcal{P}$ projects V_i on a 2D grid:

$$\mathcal{P}_{(\mathbf{R},\mathbf{t})}V_i = \left\{ \int_t V\left(\mathbf{R} \cdot [x_{m,n} - t_x, y_{m,n} - t_y, t]^T\right), (m,n) \in \{1, ..., D\}^2 \right\}. \tag{2}$$

η_i models isotropic Gaussian noise ($\eta_i \sim \mathcal{N}(0, \sigma^2)$). In a typical experiment, N can vary between 10^5 and 10^7; The signal-to-noise ratio can vary between 10^{-1} and 10^{-2}.

3.2 Heterogeneity Model

Structural heterogeneity among the macromolecules can originate (1) from continuous motion along a small number of *degrees of freedom*, or (2) from discrete compositional changes. We refer to the first kind of heterogeneity as "conformational heterogeneity" and to the second one as "compositional heterogeneity".

To model this mathematically, we make the assumption that all the density maps V_i belong to a finite union of low-dimensional manifolds:

$$\forall i \in \{1, \dots, N\}, V_i \in \{\mathcal{V}(z; \theta_k), z \in \mathbb{R}^d, k \in \{1, \dots, K\}\}, \tag{3}$$

where $K \in \mathbb{N}$, $d \in \mathbb{N}$ and, for all k, $\theta_k \in \Theta$ is an unknown parameter. In other words, we assume that there exist K compositional states and that the conformational motion of state k can essentially be described with d degrees of freedom.

3.3 Latent Variable Model

We statistically describe the set of observed images with a latent variable model. Here, the *latent variables* are the poses ϕ_i, the conformations z_i and the class identities k_i, while $\{\theta_k\}_{k=1}^K$ is a set of *shared parameters*.

Given the image formation model described by Eq. 1 and the heterogeneity model described by Eq. 3, each observed image can be seen as a sample from a multivariate mixture model:

$$p(I_i) = \sum_{k=1}^{K} p(k_i = k) \iint p(\phi | k_i = k) p(z | k_i = k) \mathcal{N}(C_i * \mathcal{P}_\phi \mathcal{V}(z; \theta_k), \sigma^2) \mathrm{d}\phi \mathrm{d}z. \tag{4}$$

We parameterize a probability distribution over the class identity, such that:

$$\forall i \in \{1, \dots, N\}, \forall k \in \{1, \dots, K\}, \quad p(k_i = k) = \text{softmax}(\mathbf{s}_i)_k \doteq \frac{\exp(s_{i,k})}{\sum_j \exp(s_{i,j})}, \tag{5}$$

where $\mathbf{s}_i \in \mathbb{R}^K$ is a free parameter called the "score". For each k, we parameterize point estimates of the continuous latent variables:

$$\begin{aligned} p(\phi|k_i = k) &= \delta(\phi - \phi_{i,k}), \quad \phi_{i,k} \in \text{SO}(3) \times \mathbb{R}^2 \\ p(z|k_i = k) &= \delta(z - z_{i,k}), \qquad\qquad z_{i,k} \in \mathbb{R}^d. \end{aligned} \tag{6}$$

Under the model in Equation (4) and the variational parameterization described above, the negative log-likelihood of an image is given by

$$\begin{aligned} &\ell_i(\{\phi_{i,k}\}_{k=1}^K, \{z_{i,k}\}_{k=1}^K, \mathbf{s}_i; \{\theta_k\}_{k=1}^K) \\ &\qquad = -\log \sum_{k=1}^K \exp\left(-\frac{1}{2\sigma^2}\|I_i - C_i * \mathcal{P}_{\phi_{i,k}}\mathcal{V}(z_{i,k};\theta_k)\|_2^2 + \log(\text{softmax}(\mathbf{s}_i)_k)\right) \end{aligned} \tag{7}$$

In our implementation, each low-dimensional manifold $\mathcal{V}(.;\theta_k)$ is implemented with a residual MLP. We provide further details on the architecture of the neural networks in the supplementary materials.

3.4 Hybrid Optimization Strategy

We aim at minimizing the negative log-likelihood of the set of observed images,

$$\mathcal{L} = \sum_{i=1}^N \ell_i(\{\phi_{i,k}\}_{k=1}^K, \{z_{i,k}\}_{k=1}^K, \mathbf{s}_i; \{\theta_k\}_{k=1}^K), \tag{8}$$

over $(\{\phi_{i,k}\}, \{z_{i,k}\}, \{\mathbf{s}_i\}, \{\theta_k\})$. All the parameters are optimized using a combination of gradient-based optimization and exhaustive search. Here, we use an autodecoding framework and do not *amortize* the inference of latent variables with an encoder, following the observation in Levy et al. [26] that encoders tend to memorize images in highly-noisy setups.

We handle the minimization over $\{z_{i,k}\}$, $\{\mathbf{s}_i\}$ and $\{\theta_k\}$ with stochastic gradient descent. The minimization over the poses $\{\phi_{i,k}\}$ is more challenging, due to the existence of local minima. We take inspiration from the two-step pose estimation strategy introduced in DRGN-AI [26] and adapt it to cope with the simultaneous representation of multiple maps. First, optimization over poses is done with hierarchical pose search (HPS) in alternation with stochastic gradient descent on the other variables. On a given random *minibatch* of indices $\mathcal{I} \subset \{1, \ldots, N\}$,

$$\forall i \in \mathcal{I}, \forall k \in \{1, \ldots, K\}, \quad \phi_{i,k} \leftarrow \underset{\phi \in \Omega}{\text{argmin}}\, \|I_i - C_i * \mathcal{P}_\phi \mathcal{V}(z_{i,k};\theta_k)\|_2^2, \tag{9}$$

where Ω is a predefined grid in the space of poses ($\text{SO}(3) \times \mathbb{R}^2$). See [26] and the supplementary for more details on how the minimization can be done efficiently with a hierarchical strategy. This robust but computationally-expensive strategy is used for a predefined number of steps, usually between 5×10^5 and 2×10^6. After that, most poses are located in the basin of attraction of their global optimum and switching to stochastic gradient descent (SGD) makes computation more efficient while improving pose accuracy (poses are not constrained to belong to a predefined grid of fixed resolution during SGD).

At the end of optimization, the estimated class of image I_i is simply given by the index k_i of the largest entry in $\mathbf{s}_i \in \mathbb{R}^K$.

4 Experiments

We run three experiments to evaluate Hydra. In Section 4.1, we show that our method improves the expressiveness of neural-based methods and can reveal strong compositional heterogeneity fully *ab initio*. In Section 4.2, we use Hydra to reveal compositional heterogeneity in an experimental dataset containing protein complexes of diverse sizes, in a single run. In Section 4.3, we demonstrate the simultaneous reconstruction of compositional and conformational heterogeneity.

https://doi.org/10.52202/079017-1816

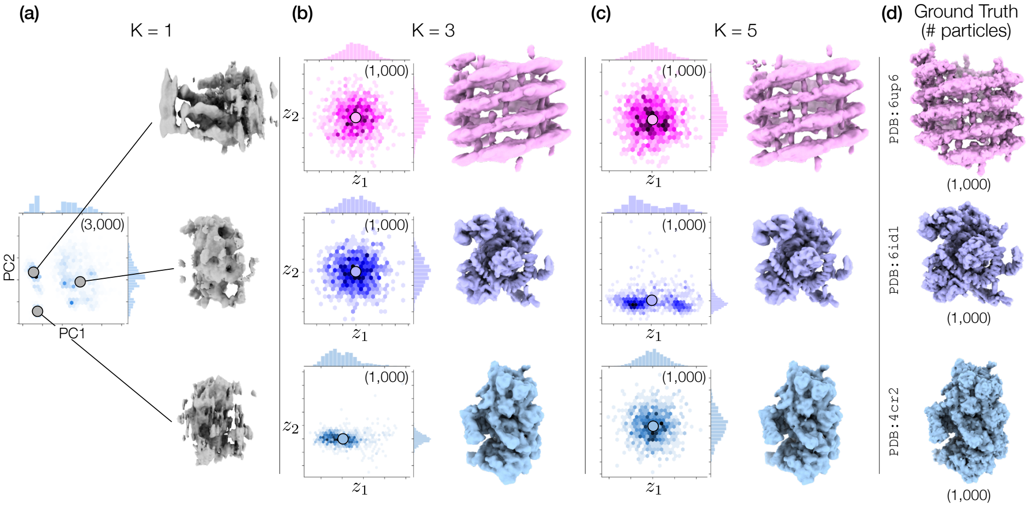


Figure 2: Hydra captures strong compositional heterogeneity in the `tomotwin3` dataset. (a-c) Reconstructed densities and estimated conformations with $K \in \{1, 3, 5\}$. We report the number of particles in each class between parenthesis. We represent density maps using isosurfaces. **(a)** With $K = 1$ (DRGN-AI), the model fails to reconstruct the three density maps, in spite of using $d = 8$ dimensions to represent conformations. **(b)** With $K = 3$ ($d = 2$), Hydra recovers the three density maps with perfect classification accuracy. **(c)** With $K = 5$ ($d = 2$), the model is over-parameterized and 2 classes out of 5 end up empty at the end of optimization. **(d)** Ground truth density maps for the `tomotwin3` dataset.

	ARI ↑	**Per-image FSC** ↑		
Model	All	PDB `6up6`	PDB `6id1`	PDB `4cr2` ↑
Hydra ($K = 3$)	**1.00**	0.25 ± 0.03	0.394 ± 0.001	$\mathbf{0.367 \pm 0.001}$
Hydra ($K = 5$)	**1.00**	0.25 ± 0.03	$\mathbf{0.396 \pm 0.001}$	$\mathbf{0.367 \pm 0.001}$
DRGN-AI ($K = 1$)	0.59	0.0242 ± 0.003	0.040 ± 0.005	0.042 ± 0.004
CryoDRGN2	0.36	0.08 ± 0.01	0.070 ± 0.006	0.3 ± 0.1
CryoSPARC	**1.00**	**0.284**	0.367	0.338

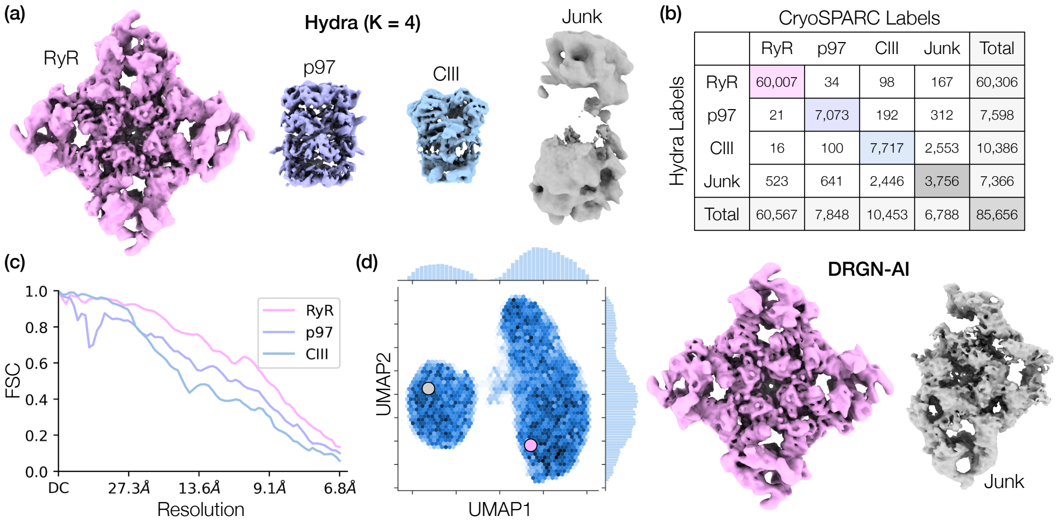
Table 1: Hydra captures compositional heterogeneity in a challenging synthetic dataset and outperforms other neural-based methods. The classification accuracy is evaluated for each method using the adjusted Rand index (ARI) [18]. To evaluate each method's reconstruction quality, we use the mean area under the Fourier shell correlation (FSC) curve for 20 images per class (we report ±1 standard deviation). We **bold** the best result, and underline the second best result.

4.1 *Ab initio* reconstruction of compositional heterogeneity

We evaluate Hydra on `tomotwin3`, a synthetic dataset of 3,000 images emulating a protein sample containing multiple species with static structures. We selected the 6th, 7th, and 8th largest proteins by atomic weight from the TomoTwin training dataset [48], which correspond to entries `6up6`, `6id1`, and `4cr2` in the RCSB PDB (Protein Data Bank) [3]. We rendered 1,000 synthetic images of each specimen following the cryo-EM image formation model. For each protein structure, we simulated density maps with the ChimeraX `molmap` command [30, 58]. We padded the density maps to a box size of $D = 384$ pixels and centered them. We sampled 1,000 orientations per ground truth class uniformly from $\mathrm{SO}(3)$. We projected the density maps using Eq. 1 and applied a translation vector uniformly chosen from a 30-pixel-wide box centered at the origin. We applied a CTF sampled from EMPIAR-11247 [10] and added Gaussian noise (SNR = 0.01). We downsampled each image to a box size of $D = 128$, yielding 3,000 images in total (samples shown in Figure S1).

We compare the performance of Hydra to several state-of-the-art *ab initio* methods for resolving compositional heterogeneity in cryo-EM datasets, including DRGN-AI [26], cryoDRGN2 [70], and cryoSPARC [45]. Hydra only uses 2 dimensions for the conformational space, while DRGN-AI and CryoDRGN2 are evaluated with a more expressive 8-dimensional space. We test Hydra using the correct number of classes ($K = 3$, the true number of specimens present in the sample) and

https://doi.org/10.52202/079017-1816



CryoSPARC Labels

Hydra Labels	RyR	p97	CIII	Junk	Total
RyR	60,007	34	98	167	60,306
p97	21	7,073	192	312	7,598
CIII	16	100	7,717	2,553	10,386
Junk	523	641	2,446	3,756	7,366
Total	60,567	7,848	10,453	6,788	85,656

Figure 3: Hydra captures compositional heterogeneity in a real dataset containing a mixture of membrane and soluble protein complexes. (a) Density maps obtained with Hydra ($K = 4$) on the Ryanodine receptor dataset. **(b)** Confusion matrix between Hydra and cryoSPARC $K = 6$ heterogeneous refinement (three classes representing RyR were combined for analysis). **(c)** Fourier shell correlation (FSC) between the Hydra density maps and refined cryoSPARC density maps. **(d)** *Left:* latent space plot and *right:* representative density maps from each of the latent space clusters from DRGN-AI.

using an over-parameterized setup ($K = 5$ in Figure 2 and $K = 7$ in Figure S2). For DRGN-AI and CryoDRGN2, we classify each image using k-means clustering on the conformational space with 3 clusters. We evaluate per-image FSC for each class using the dataset's ground truth class labels.

In Table 1, we report results for each configuration of Hydra and other methods, (best of three replicas with distinct seeds for Hydra and DRGN-AI, full results in Table 3). We observe that other methods with implicit neural volume representations, including DRGN-AI and CryoDRGN2, fail to capture the compositional heterogeneity of the dataset. Hydra matches the classification quality of cryoSPARC, which models independent, discrete mixtures. Hydra also outperforms all methods on per-image FSC, a distributional volume reconstruction quality metric based on the Fourier Shell Correlation (FSC) [68]. We provide additional results with a larger version of this dataset in Figure S3 and additional metrics (including roto-translation accuracy) in Table 4.

4.2 *Ab initio* reconstruction of an experimental cryo-EM mixture dataset

We evaluate our method on an experimental dataset of a protein mixture from red blood cell lysate subjected to density-gradient centrifugation followed by chemical cross-linking. Through an exhaustive, expert-driven data processing pipeline in cryoSPARC [45], we determined that the dataset consisted of substantial compositional heterogeneity from three main component protein complexes: the membrane proteins ryanodine receptor (RyR) and mitochondrial respiratory chain complex III (CIII), and a dimeric complex of the soluble valosin-containing protein (p97). A sweep of K-values from cryoSPARC ab-initio reveals that $K \geq 5$ is necessary in order for cryoSPARC to reconstitute distinct RyR, p97, CIII, and junk classes (Figure S4).

In Figure 3, we compare the performance of DRGN-AI [26] with Hydra. Figures 3d and S5 show that DRGN-AI successfully partitions the dataset into two clusters corresponding to the RyR and non-RyR particles; however, DRGN-AI is unable to learn distinct shapes for all three discrete structures and instead appears to learn structural artifacts from the non-RyR particles that match the overall RyR shape. By contrast, Hydra with $K = 4$ successfully separates the particles into the three component protein complexes and a fourth junk class (Figures 3 and S6). As a baseline for comparison, we also carried out the typical workflow of first generating poses from a consensus reconstruction in cryoSPARC, followed by fixed-pose DRGN-AI to separate the heterogeneity. As can be seen in Figure S7, DRGN-AI recovers a high resolution RyR density but fails to learn non-RyR protein structures when training from poses from homogeneous reconstruction. A multi-step reconstruction with DRGN-AI does not either reveal the CIII class (Figure S8).

https://doi.org/10.52202/079017-1816

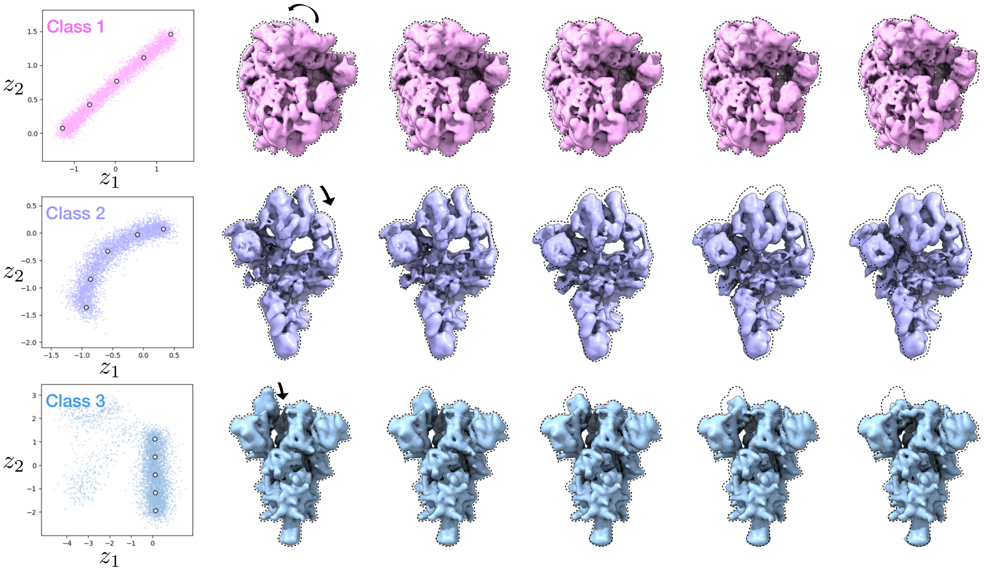


Figure 4: Hydra effectively recovers both compositional and conformational heterogeneity in the `ribosplike` dataset. Particles within each latent space are colored by class. Representative density maps are generated from the latent points denoted in white dots.

4.3 *Ab initio* reconstruction of conformational and compositional heterogeneity.

We prepare a synthetic dataset exhibiting both compositional and conformational heterogeneity (`ribosplike`). For each of the pre-catalytic spliceosome, 80S ribosome, and SARS-CoV-2 spike protein, 5,000 images are generated from different conformational states of the macromolecule along a 1-dimensional trajectory of motion [41, 64, 61] yielding a dataset of 15,000 images (See SI Section F for additional details and Figure S1 for sample images).

As seen in Figure 4, our method with $K = 3$ nearly perfectly separates the particles into their three classes. Within each neural field, the conformational change of the dominant particle type is well captured both by the conformational space. Due to the trimeric structure of the spike protein, we observe three continuous manifolds in the conformational space, though the majority of spike particles are aligned to one of the symmetries. Tables 2 and 5 compare the performance of Hydra with other *ab initio* baselines, demonstrating superior performance in classification accuracy as measured by the Adjusted Rand Index (ARI) and volume reconstruction quality (per-image FSC). It additionally achieves the lowest pose error amongst all methods. Representative reconstructions for baseline methods are shown in the supplementary material (Figure S9).

Model	**Img-FSC ↑**	**ARI ↑**	**Pose Err. ↓**
Hydra (K=3)	**0.414**	**0.997**	**1.070**
DRGN-AI	0.207	0.994	45.379
CryoDRGN2	0.399	0.986	1.2120
CryoSPARC	0.344	0.972	1.576

Table 2: Hydra outperforms state-of-the-art methods on the `ribosplike` dataset. Quantitative results include reconstruction quality (per-image FSC), particle classification (ARI), and median pose accuracy (geodesic distance between rotations, in degrees).

5 Discussion

This work introduces Hydra, a method that expands the expressiveness of neural-based models for heterogeneous reconstruction in cryo-EM. Hydra models structures as arising from 1 of K neural fields and is designed to capture heterogeneity in datasets containing a mixture of multiple species. Notably, our method runs fully *ab initio*, i.e. does not rely on pre-estimated poses, which are not straightforward to obtain in cases of strong compositional heterogeneity. We evaluated Hydra on both synthetic and experimental datasets, showed improved performance over previous neural-based methods on compositionally heterogeneous datasets, and demonstrated that it could simultaneously handle compositional and conformational heterogeneity.

Our method can be seen as an instance of a mixture of experts model, where the gating mechanism is handled by an autodecoding framework, due to the low signal-to-noise ratio in the input data. However, predicting classes with a neural network can be viewed a *classification* task and may be easier than pose or conformation estimation. We view the possibility of using a (potentially pretrained) neural network for classification, thereby making use of amortized inference, as an exciting avenue for future work.

One important limitation of our approach is the need to specify the number of classes K prior to reconstruction. Other methods that cope with compositional heterogeneity, like RELION [50] and cryoSPARC [45], possess the same requirement, and this is solved in practice by running the algorithm several times with a sweep on the hyperparameter K. As this process is time and energy-consuming (a single reconstruction can run for up to 4 GPU-days on high-end GPUs), more efficient methods for hyperparameter selection, for example, with an adaptive and/or hierarchical strategy, is an impactful direction for future work. We would also like to note that, in its current implementation, Hydra must give the same latent dimension (d) to all the classes, but this constraint could be relaxed with, for-example, a dictionary-based representation for the conformations.

With Hydra, we broaden the scope of datasets that cryo-EM reconstruction algorithms can process. Our method would be especially suitable for data collected *in situ* (i.e., inside the cell), where mixtures of dynamic biomolecular complexes coexist. This data is usually acquired by cryogenic electron tomography (cryo-ET), where the sample is progressively tilted throughout the data collection process. Extending our method to enable subtomogram averaging of cryo-ET data is another potential future direction that we hope can enable future discoveries in structural biology and facilitate the design of new therapeutic compounds.

6 Acknowledgments

The author(s) are pleased to acknowledge that the work reported on in this paper was substantially performed using the Princeton Research Computing resources at Princeton University which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing. The Zhong lab gratefully acknowledges support from the Princeton Catalysis Initiative, Princeton School of Engineering and Applied Sciences, and Generate Biomedicines. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[2] Lars V Bock and Helmut Grubmuller. Effects of cryo-em cooling on structural ensembles. *Biophysical Journal*, 121(3):148a, 2022.

[3] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.

[4] Alister Burt, Bogdan Toader, Rangana Warshamanage, Andriko von Kugelgen, Euan Pyle, Jasenko Zivanov, Dari Kimanius, Tanmay AM Bharat, and Sjors Scheres. An image processing pipeline for electron cryo-tomography in relion-5. *bioRxiv*, pages 2024–04, 2024.

[5] Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional gaussian mixture model for characterizing variability in cryo-em. *Nature methods*, 18(8):930–936, 2021.

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5939–5948, 2019.

[7] Ali Dashti, Peter Schwander, Robert Langlois, Russell Fung, Wen Li, Ahmad Hosseinizadeh, Hstau Y Liao, Jesper Pallesen, Gyanesh Sharma, Vera A Stupina, et al. Trajectories of the ribosome as a brownian nanomachine. *Proceedings of the National Academy of Sciences*, 111 (49):17492–17497, 2014.

[8] Ali Dashti, Ghoncheh Mashayekhi, Mrinal Shekhar, Danya Ben Hail, Salah Salah, Peter Schwander, Amedee des Georges, Abhishek Singharoy, Joachim Frank, and Abbas Ourmazd. Retrieving functional pathways of biomolecules from single-particle snapshots. *Nature communications*, 11(1):4734, 2020.

[9] Reza Ebrahimpour, Ehsanollah Kabir, Hossein Esteky, and Mohammad Reza Yousefi. View-independent face recognition with mixture of experts. *Neurocomputing*, 71(4-6):1103–1107, 2008.

[10] J. Ryan Feathers, Erica K. Richael, Kayla A. Simanek, J. Christopher Fromme, and Jon E. Paczkowski. Structure of the rhlr-pqse complex from pseudomonas aeruginosa reveals mechanistic insights into quorum-sensing gene regulation. *Structure*, 30(12):1626–1636.e4, 2022. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2022.10.008. URL `https://www.sciencedirect.com/science/article/pii/S0969212622003975`.

[11] Joachim Frank and Abbas Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-em. *Methods*, 100:61–67, 2016.

[12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[13] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.

[14] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622 (2):759, 2005.

https://doi.org/10.52202/079017-1816

[15] Harshit Gupta, Thong H Phan, Jaejun Yoo, and Michael Unser. Multi-cryogan: Reconstruction of continuous conformations in cryo-em using generative adversarial networks. In *European Conference on Computer Vision*, pages 429–444. Springer, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] David Herreros, Roy R Lederman, James Krieger, Amaya Jiménez-Moreno, Marta Martínez, David Myška, David Strelak, Jiri Filipovic, Ivet Bahar, Jose Maria Carazo, et al. Approximating deformation fields for the analysis of continuous heterogeneity of biological macromolecules by 3d zernike polynomials. *IUCrJ*, 8(6):992–1005, 2021.

[18] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[20] Qiyu Jin, Carlos Oscar S Sorzano, José Miguel de La Rosa-Trevín, José Román Bilbao-Castro, Rafael Núñez-Ramírez, Oscar Llorca, Florence Tama, and Slavica Jonić. Iterative elastic 3d-to-2d alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure*, 22(3):496–506, 2014.

[21] Dari Kimanius, Kiarash Jamali, and Sjors Scheres. Sparse fourier backpropagation in cryo-em reconstruction. *Advances in Neural Information Processing Systems*, 35:12395–12408, 2022.

[22] Dari Kimanius, Kiarash Jamali, Max E Wilkinson, Sofia Lövestam, Vaithish Velazhahan, Takanori Nakane, and Sjors HW Scheres. Data-driven regularisation lowers the size barrier of cryo-em structure determination. *bioRxiv*, pages 2023–10, 2023.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Roy R Lederman and Amit Singer. Continuously heterogeneous hyper-objects in cryo-em and 3-d movies of many temporal dimensions. *arXiv preprint arXiv:1704.02899*, 2017.

[25] Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022.

[26] Axel Levy, Frederic Poitevin, Jake D Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D Zhong. Revealing biomolecular structure and motion with neural ab initio cryo-em reconstruction. *biorxiv:preprint*, 2024.

[27] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022.

[28] Julien NP Martel, David B Lindell, Connor Z Lin, Eric R Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021.

[29] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293, 2014.

[30] Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[33] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-em reconstruction of continuous heterogeneity by laplacian spectral volumes. *Inverse Problems*, 36(2):024003, 2020.

[34] Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors HW Scheres. Characterisation of molecular motions in cryo-em single-particle data by multi-body refinement in relion. *elife*, 7: e36861, 2018.

[35] Steven Nowlan and Geoffrey E Hinton. Evaluation of adaptive mixtures of competing experts. *Advances in neural information processing systems*, 3, 1990.

[36] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021.

[37] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[39] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.

[41] Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546:617–621, 2017.

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[43] Ali Punjani and David J Fleet. 3d variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-em. *Journal of structural biology*, 213(2): 107702, 2021.

[44] Ali Punjani and David J Fleet. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870, 2023.

[45] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

[46] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14161, 2021.

[47] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14335–14345, 2021.

[48] Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. Tomotwin: generalized 3d localization of macromolecules in cryo-electron tomograms with structural data mining. *Nature Methods*, 20(6):871–880, Jun 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01878-z. URL `https://doi.org/10.1038/s41592-023-01878-z`.

[49] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W Senior, John Jumper, Carl Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. *arXiv preprint arXiv:2106.14108*, 2021.

[50] Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

[51] Johannes Schwab, Dari Kimanius, Alister Burt, Tom Dendooven, and Sjors Scheres. Dynamight: estimating molecular motions with improved reconstruction from cryo-em images. *bioRxiv*, pages 2023–10, 2023.

[52] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[53] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[54] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

[55] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021.

[56] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[57] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[58] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007.

[59] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021.

[60] Miloš Vulović, Raimond BG Ravelli, Lucas J van Vliet, Abraham J Koster, Ivan Lazić, Uwe Lücken, Hans Rullgård, Ozan Öktem, and Bernd Rieger. Image formation modeling in cryo-electron microscopy. *Journal of structural biology*, 183(1):19–32, 2013.

[61] Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.02.058. URL `https://www.sciencedirect.com/science/article/pii/S0092867420302622`.

[62] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281–292, 2020.

[63] Steven Richard Waterhouse. *Classification and regression using mixtures of experts*. PhD thesis, Citeseer, 1998.

[64] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-em structure of the *Plasmodium falciparum* 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, jun 2014. ISSN 2050-084X. doi: 10.7554/eLife.03080. URL `https://doi.org/10.7554/eLife.03080`.

[65] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021.

[66] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020.

[67] Anna Yershova, Swati Jain, Steven M Lavalle, and Julie C Mitchell. Generating uniform incremental grids on so (3) using the hopf fibration. *The International journal of robotics research*, 29(7):801–812, 2010.

[68] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-EM images. In *International Conference on Learning Representations (ICLR)*, May 2020.

[69] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

[70] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021.

[71] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Exploring generative atomic models in cryo-em reconstruction. *arXiv preprint arXiv:2107.01331*, 2021.

Appendix

A Architectural Details

Each manifold of density maps $\mathcal{V}(, ; \theta_k)$ is implemented as a neural network. Conditioned on a conformation $z \in \mathbb{R}^d$, $\mathcal{V}(, ; \theta_k) : \mathbb{R}^3 \to \mathbb{R}$ represents the Hartley transform of the 3D electron scattering potential of a single particle. The frequency coordinate $\mathbf{k} \in [-0.5, 0.5]^3$ is expanded in a sinusoidal basis using Fourier features [56] (64 base frequencies are randomly sampled from a 3D Gaussian distribution of standard deviation 0.5). The neural network contains 3 hidden residual layers [16] of size 128 with ReLU nonlinearities, without any normalization schemes.

B Pose Estimation

"Hierarchical pose search" (HPS) is done on a predefined grid over $\mathrm{SO}(3) \times \mathbb{R}^2$ (4,608 rotations and 49 translations in $[-10 \text{ pix.}, 10 \text{ pix.}]$). The first 8 poses minimizing the reprojection error are kept and refined with a local search over 8 neighbors during 4 additional steps. The images are band-limited during pose search and the cutoff frequency increases linearly from k_{min} to k_{max} ($k_{\text{min}} = 6$, $k_{\text{max}} = 16$ in (image length)$^{-1}$). Grids are parameterized using the Hopf fibration [67], product of the Healpix [14] grid on the 2-sphere and a regular grid on the circle.

Once 2,000,000 images have been fed to the model (with a minimum of 2 epochs of pose search), the pose estimation strategy switches to stochastic gradient descent (SGD). The poses $\phi_{i,k}$ are initialized with the last poses estimated by hierarchical pose search.

C Conformation Estimation

The conformations $\phi_{i,k}$ are independently optimized by SGD. They are initialized randomly from a d-dimensional Gaussian distribution of standard deviation 0.1. Unless stated otherwise, the dimension d of the conformations is 2.

D Score Estimation

The scores $\mathbf{s}_i \in \mathbb{R}^K$ are optimized by SGD and converted into probability vectors of dimension K with a softmax operator.

E Optimization Parameters

We use the Adam optimizer [23] without weight decay and with the following learning rates: 0.1 for the scores, 0.01 for the conformations, 0.001 for the poses, and 0.0001 for the weights of the neural networks.

F Synthetic Datasets

Simulation of compositional heterogeneity. We first describe the construction of `tomotwin3`, the synthetic dataset with compositional heterogeneity only. We selected the 6th, 7th, and 8th largest proteins by atomic weight from the training dataset used in TomoTwin [48], which correspond to entries `6up6`, `6id1`, and `4cr2` in the RCSB PDB (Protein Data Bank) [3]. We simulate density maps for each entry using the ChimeraX `molmap` command [30, 58] using a grid spacing of 1.5 Å/px and a resolution of 3.0 Å/px. We pad each density map to a box size of $D = 384$ pixels and center it. We sample 3,000 orientations uniformly from $\mathrm{SO}(3)$, and 3,000 translation vectors in a 30-pixel-wide box around the origin (1,000 poses for each PDB entry). We project each density map by applying the corresponding orientation, projecting using Eq. 1, and applying the corresponding translation vector. We apply a contrast transfer function (CTF) uniformly sampled from EMPIAR-11247 [10], a representative cryo-EM dataset. We add Gaussian noise to each image to reach a signal-to-noise ratio (SNR) of 0.01. We downsample each image to an image size of $D = 128$ pixels.

https://doi.org/10.52202/079017-1816

We evaluated Hydra on `tomotwin3` with a 2-dimensional conformational space. We perform HPS on 100,000 images (33 epochs), followed by 100 epochs of SGD pose optimization. We set the batch size to 64 during SGD pose optimization. The score table learning rate is set to 0.01, and σ is set to 0.1. We evaluate Hydra using both the correct number of classes ($K = 3$) and an overparametrized configuration ($K = 5$). All other parameters are set to default values.

We train CryoDRGN2 v3.3.0 for 30 epochs using an 8-dimensional latent space, an encoder width of 1024, 3 encoder layers, and a decoder width of 1024. We evaluate DRGN-AI using an 8-dimensional conformational space, 100k images (33 epochs) of HPS followed by 100 epochs of SGD pose estimation. cryoSPARC *ab initio* was run with $K = 3$ classes. All other parameters are set to default values.

When calculating volume metrics, we compare the reconstructed density map to a downsampled ground truth density map ($D = 128$ pixels). All experiments on `tomotwin3` are run on one NVIDIA A100 GPU. Our experiments required 4h00min for $K = 3$ and 6h20min for $K = 5$. DRGN-AI ran in 1h20min and cryoDRGN2 in 1h50min.

Simulation of conformational and compositional heterogeneity. For the pre-catalytic spliceosome, we obtain a trained cryoDRGN model from Zenodo[2] and generate 500 density maps of box size D=256 at equally spaced points along the first principal component of the latent space [41] [69]. For the 80S ribosome, we likewise obtain a trained cryoDRGN model from Zenodo and generate 500 density maps of box size D=256 at equally spaced points along a linear trajectory connecting the embeddings of particles 34570 and 60629 in the latent space [64, 69]. For the SARS-CoV-2 spike protein, a cryoDRGN model was trained on a filtered subset of the dataset of Walls et al. [62], for 25 epochs with a circular image mask of dimension 64, latent dimension of 8, and all other hyperparameters at default. Then, we generate 500 density maps of box size D=256 at equally spaced points along the first principal component of the latent space [61].

We next match the pixel sizes of the density maps by standardizing them to that of the spliceosome. In particular, we downsample the ribosome density maps to D=228 and pad back to D=256, and downsample the spike protein density maps to D=198 and pad back to D=256. Then, we generate and apply a soft mask for each density map, at thresholds of 0.03, 0.05, and 0.1 for the spliceosome, ribosome, and spike protein, respectively, and with 25Å of dilation and a 15Å cosine edge. Next, we normalize density map intensity values such that the signal ratios between different particle types match approximate "true" signal ratios. We calculate the true signal ratios by running the ChimeraX *molmap* command on the PDBs of each particle (spliceosome 5NRL, ribosome 3J79, spike 6VXX and 6VYB), summing the density within each density map, and calculating signal ratios (where the signal of the spike protein is taken to be the average over the two PDBs). Then, the intensities of the 500 ribosome density maps and 500 spike protein density maps are scaled such that the ratios of total intensities across all density maps for ribosome to spliceosome, and spike to spliceosome, match the true ratios.

Finally, we generate images from all the density maps. For each density map, 100 projection images are generated with uniformly sampled rotations from SO(3) and uniformly sampled in-plane translations within [-20, 20] pixels. CTF is applied to each image, with values drawn (with replacement) from the experimental CTF values of the spliceosome dataset [41]. Noise is added to all images at an SNR of 0.1, with the variance of the signal computed over all 15k particles. Lastly, the images are downsampled from D=256 to D=128.

We train Hydra for 100,000 images of HPS before 100 epochs of SGD. We train cryoDRGN2 for 90 epochs. All other hyperparameters for all methods are set to their defaults. For single-class DRGN-AI and cryoDRGN2, the predicted classes for Adjusted Rand Index calculation are obtained by k-means clustering the latent space with k=3. Image-FSC is computed over 50 images equally spaced along the true conformational trajectory, per each of the 3 particle types (150 images total). Experiments for Hydra were run using 2 NVIDIA V100 GPUs, and experiments for baselines were run using 1 NVIDIA V100 GPU.

[2] `https://zenodo.org/records/4355284`

https://doi.org/10.52202/079017-1816

G Real Dataset

Dataset details. 85,656 particles were manually picked from one round of 2D classification on cryoSPARC of a larger dataset of 148,596 particles. The particles were downsampled from a box size of $D = 300$ pixels and 1.66 Å/pixel to $D = 150$ pixels at 3.32 Å/pixel.

DRGN-AI and Hydra. All training for DRGN-AI and Hydra were carried out on 4 A100 NVIDIA GPUs with 80GB memory. For *ab initio* DRGN-AI, we ran DRGN-AI with default parameters (latent dimension $d = 8$), with 500k images for HPS followed by 100 epochs of SGD. We verified that single-class DRGN-AI could separate the RyR and non-RyR particles by selecting the particles belonging to the non-RyR cluster (as visualized in the latent space) for downstream single-class DRGN-AI analysis, and a second run of single-class DRGN-AI with default parameters, 500k images of HPS and 100 epochs SGD on the non-RyR particles recovered a p97 density map and a non-p97 cluster in the latent space (followed by a third round of DRGN-AI with the same parameters on the non-p97 particles). For Hydra, we trained the model with 2 million images for HPS followed by 100 epochs of SGD and latent dimension $d = 2$. After an initial sweep of the hyperparameters for learning rate and high-entropy prior σ using $K = 3$ or $K = 4$, the optimal parameters for this dataset were determined to be a learning rate of 0.1 and $\sigma = 10.0$. In order to reduce the resource demand of Hydra, we lowered the hypervolume dimension to 128 and reduced the number of hidden layers in the hypervolume to 2, as well as decreased the SGD batch size to 64. For fixed-pose DRGN-AI with consensus cryoSPARC poses, poses from a cryoSPARC homogeneous refinement job were used as input for optimization by single-class DRGN-AI with default parameters.

CryoSPARC processing. We performed a sweep of K-values for 3D classification using cryoSPARC *ab initio* from $K = 1$ through $K = 8$ (all other parameters set to default values), and determined that $K < 5$ is insufficient to separate the junk particles from protein. $K = 6$ *ab initio* yielded the best separation of particles into the four different classes, with three different RyR density maps produced. We performed heterogeneous refinement on the best *ab initio* $K = 6$ replicate for comparison to Hydra. In accordance with best practices for data processing, we performed *ab initio* with the downsampled $D = 150$ particles but used undownsampled particles boxed at $D = 300$ to generate the most accurate poses during refinement against the low-resolution density maps generated by *ab initio*.

H Additional Files

We provide the following supplementary files:

- Three movies illustrating the continuous motion of the ribosome, spliceosome and spike, as shown in Figure 4. The movies were obtained by sampling 9 points on a linear trajectory of the latent space, for each compositional state.
- The 10 volumes shown in Figure 4 for the ribosome and the spliceosome, in `mrc` format.

https://doi.org/10.52202/079017-1816

I Additional Figures

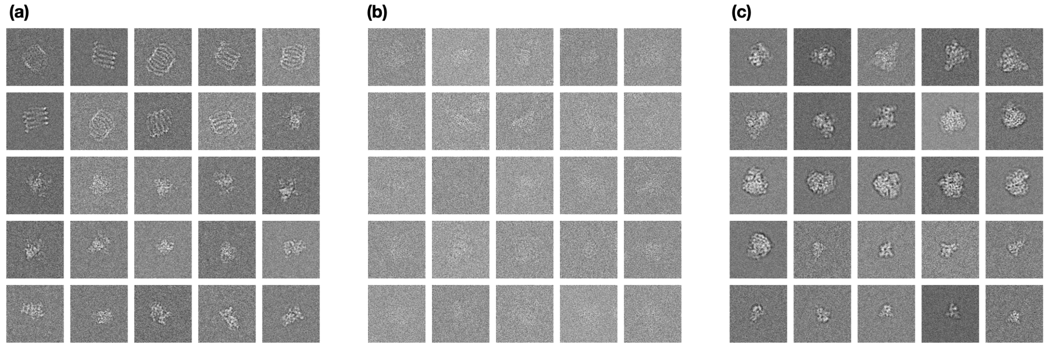
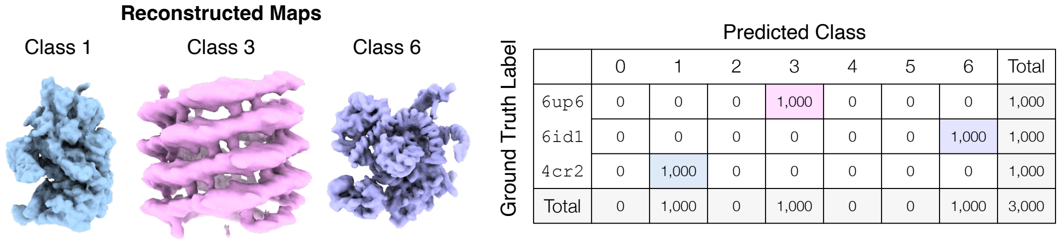
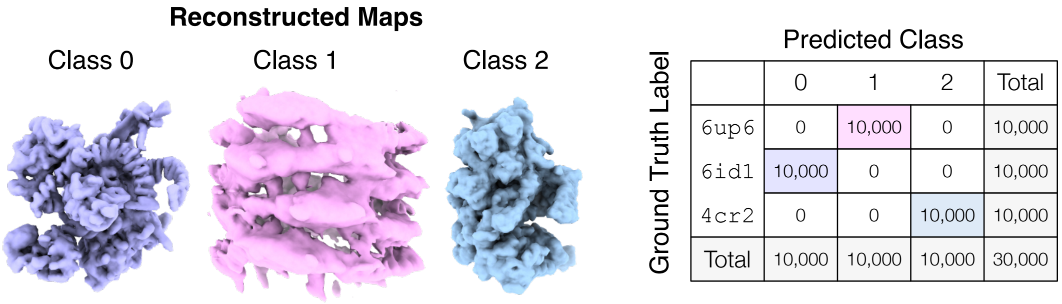


Fig. S1: 25 representative sample images from each of the referenced three datasets. **(a)** Sample images for the `tomotwin3` synthetic dataset, $D = 128$, 4.5 Å/pix. **(b)** Sample images for the experimental ryanodine receptor dataset, $D = 150$, 3.32 Å/pix. **(c)** Sample images for the `ribosplike` synthetic dataset, $D = 128$, 4.24 Å/pix.



	0	1	2	3	4	5	6	Total
6up6	0	0	0	1,000	0	0	0	1,000
6id1	0	0	0	0	0	0	1,000	1,000
4cr2	0	1,000	0	0	0	0	0	1,000
Total	0	1,000	0	1,000	0	0	1,000	3,000

Fig. S2: Results on the `tomotwin` dataset (3k particles) with a larger-than-optimal value for K ($K = 7$). Reconstructed maps on the left and confusion matrix on the right. Hydra is able to accurately reconstruct the three states in the dataset but four classes end up empty.



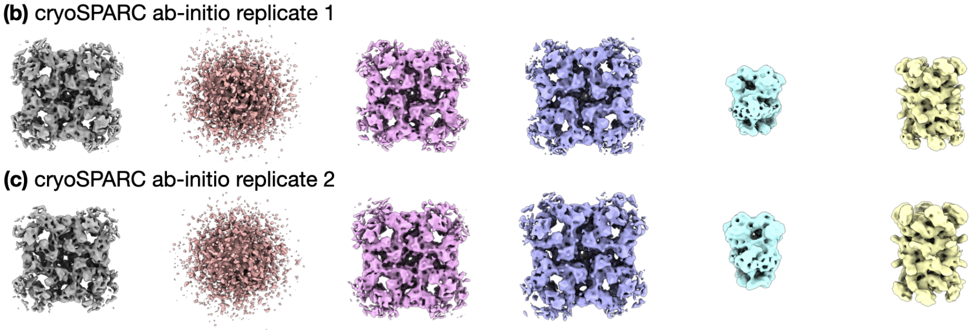
	0	1	2	Total
6up6	0	10,000	0	10,000
6id1	10,000	0	0	10,000
4cr2	0	0	10,000	10,000
Total	10,000	10,000	10,000	30,000

Fig. S3: Results on the `tomotwin` dataset (30k particles, $K = 3$).

https://doi.org/10.52202/079017-1816

(a) CryoSPARC replicate 1 labels

CryoSPARC replicate 2 labels	RyR	p97	CIII	Junk	Total
RyR	60,464	8	46	73	60,591
p97	3	7,714	74	105	7,896
CIII	19	80	9,492	1,501	11,092
Junk	52	98	1,305	4,622	6,077
Total	60,538	7,900	10,917	6,301	85656



(b) cryoSPARC ab-initio replicate 1

(c) cryoSPARC ab-initio replicate 2

Fig. S4: cryoSPARC *ab initio* $K = 6$ shows comparable levels of classification uncertainty as Hydra $K = 4$, with cryoSPARC *ab initio* showing significant classification uncertainty between particles from the CIII and junk classes, similar to the comparison between Hydra class assignments and cryoSPARC *ab initio* reference labels. (a) Confusion matrix between two $K = 6$ replicates of cryoSPARC *ab initio*. **(b)** and **(c)**: reconstructed density maps from the two cryoSPARC *ab initio* replicates.

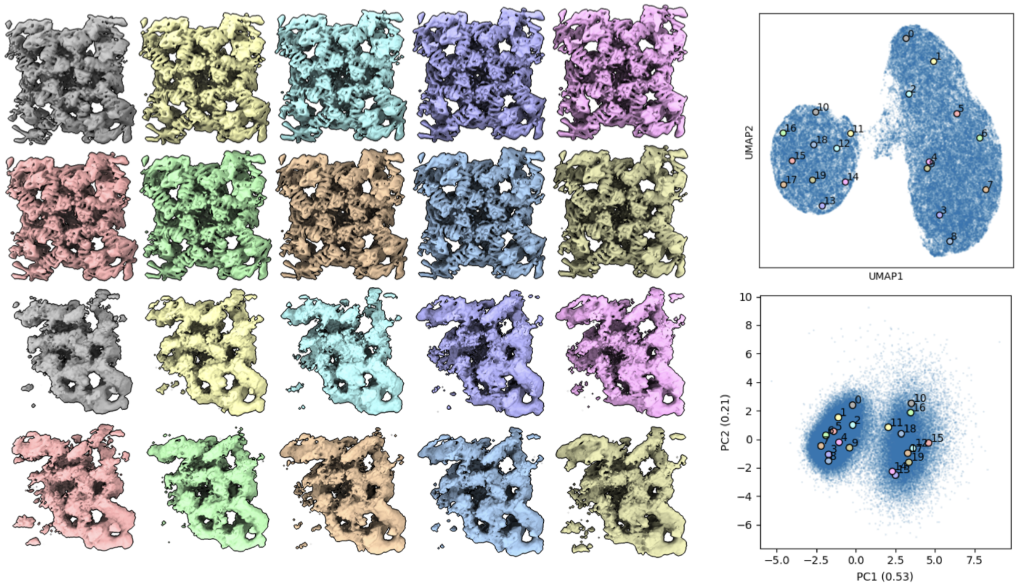


Fig. S5: Sampling of the DRGN-AI latent space from the ryanodine receptor dataset shows two main clusters corresponding to RyR and non-RyR particles. *Left:* density maps sampled using k-means clustering with $K = 20$ on the latent space; *right:* latent space plots.

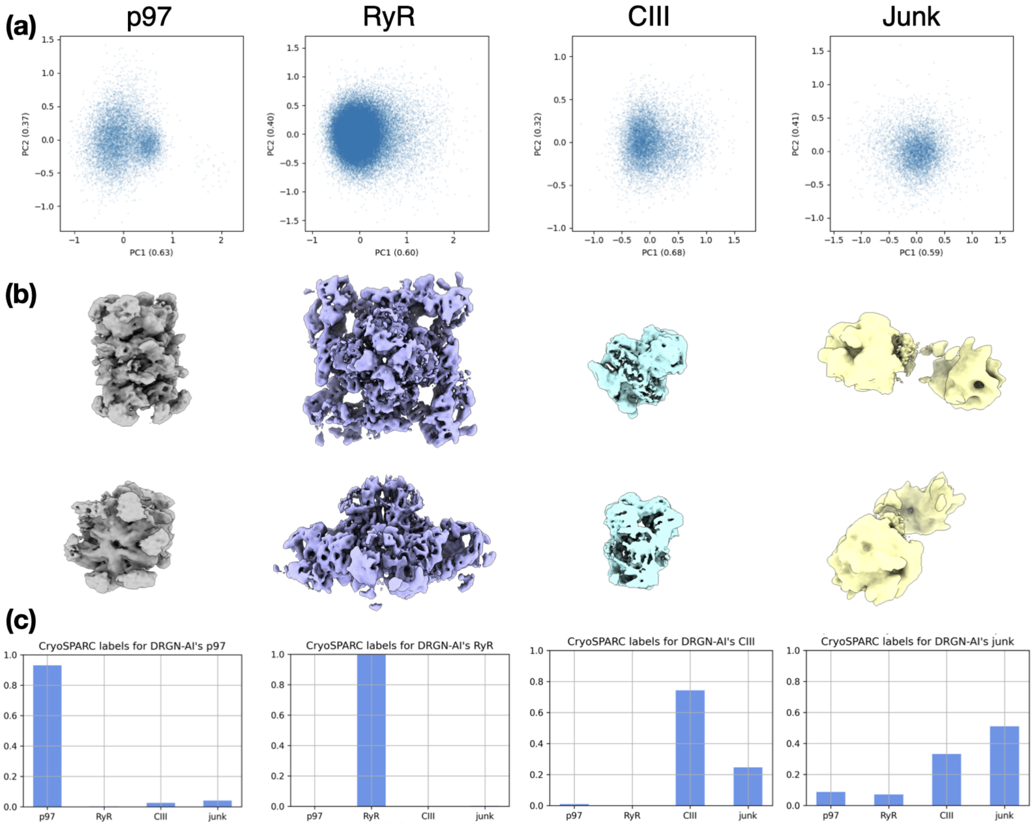


Fig. S6: Additional qualitative information for Hydra $K = 4$ on experimental ryanodine receptor dataset. **(a)** Latent space plots corresponding to each class from Hydra. **(b)** Additional top and side views of each density map generated from Hydra $K = 4$, with a cutaway view of CIII showing resolution of transmembrane helices. **(c)** Bar plots showing agreement between class assignments for Hydra $K = 4$ and cryoSPARC $K = 6$ heterogeneous refinement (particles from the three recovered RyR classes were combined for analysis), normalized by total number of particles in each Hydra class.

https://doi.org/10.52202/079017-1816

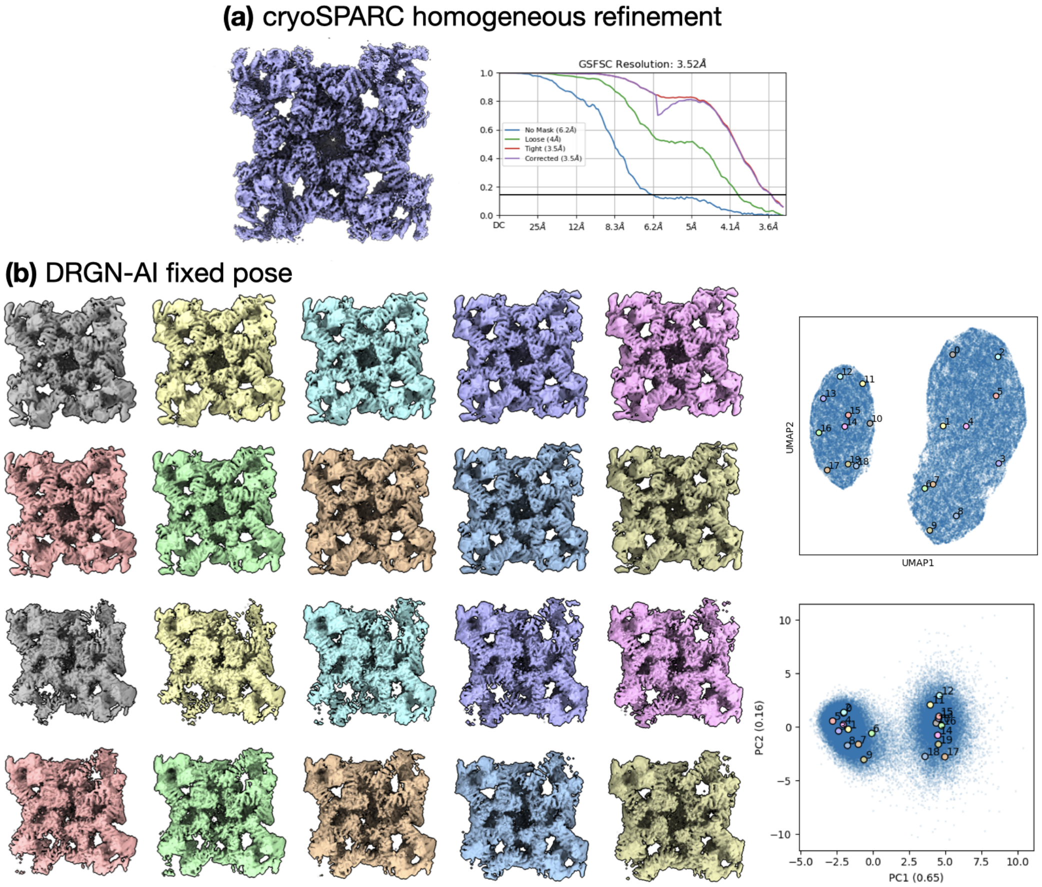


Fig. S7: The typical processing workflow of generating a consensus reconstruction followed by DRGN-AI heterogeneous reconstruction fails to capture the shape of non-RyR densities, as the cryoSPARC consensus reconstruction conceals compositional heterogeneity and yields a high-resolution density for RyR only. (**a**) Homogeneous refinement of the entire ryanodine receptor dataset against a cryoSPARC *ab initio* $K = 1$ alignment of the entire dataset (*left*); *right:* FSC curve. (**b**) single-class DRGN-AI fixed pose with poses from the cryoSPARC homogeneous refinement; *left:* densities from k-means 20 sampling of the latent space; *right:* latent space plots.

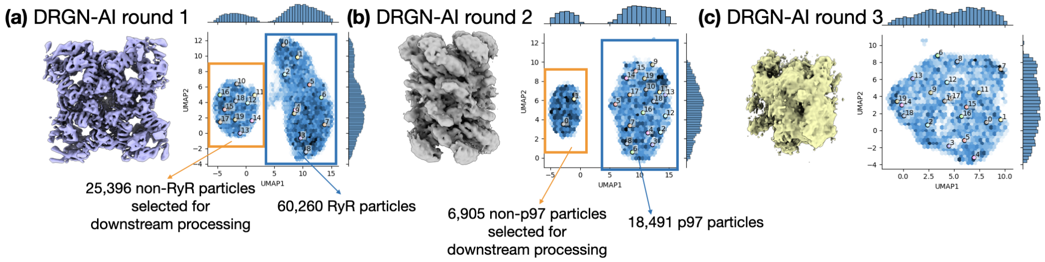


Fig. S8: Even when using a multi-shot heterogeneous reconstruction approach, DRGN-AI is only able to capture RyR and p97 densities, possibly due to the higher SNR of RyR and p97; the final density shows a mix of CIII and junk particles and no partitioning of the latent space. (**a**) DRGN-AI on the full experimental ryanodine receptor dataset; resulting RyR density generated from sampling the latent space cluster labeled in blue. (**b**) DRGN-AI on the subset of particles from the latent space cluster labeled in orange from part (**a**); p97 density sampled from latent space cluster labeled in blue. (**c**) DRGN-AI on the subset of particles from the latent space cluster labeled in orange from part (**b**).

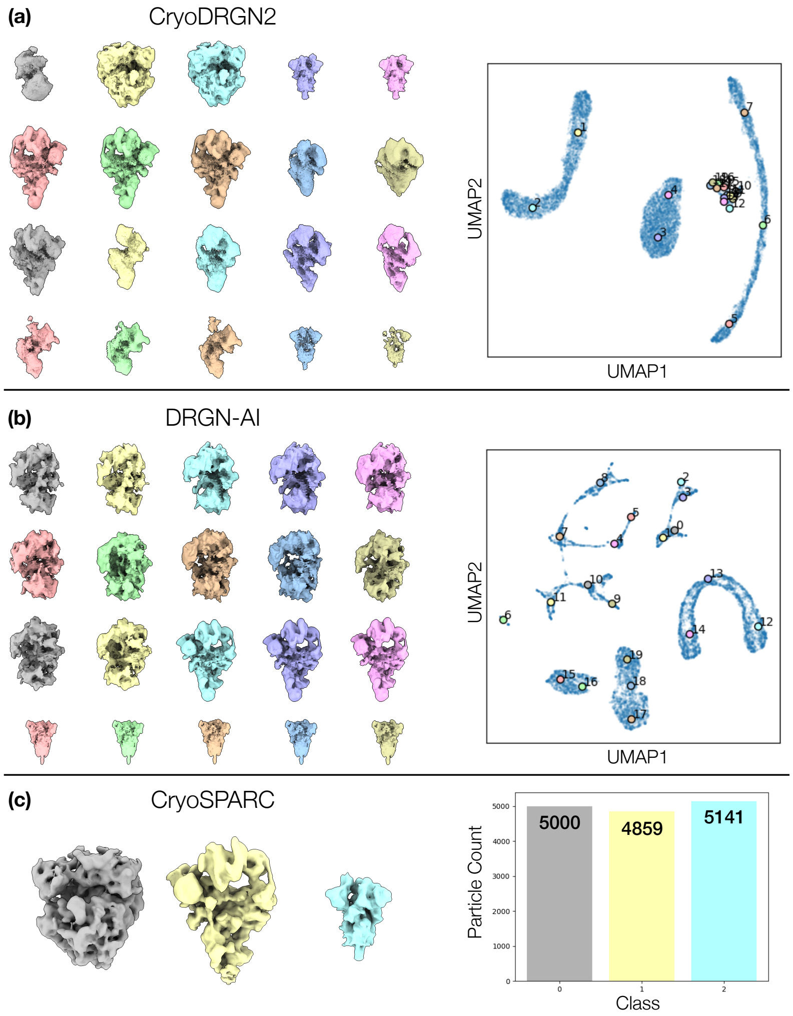


Fig. S9: Qualitative results of baselines on the synthetic `ribosplike` dataset containing compositional and conformational heterogeneity. (a)-(b) Conformational space plots and representative K-Means clustering volumes for CryoDRGN2 and single-class DRGN-AI. **(c)** Class volumes and particle counts for CryoSPARC.

J Additional Tables

		All	PDB `6up6`	PDB `6id1`	PDB `4cr2`
Model		**ARI ↑**	**Image-FSC ↑**	**Image-FSC ↑**	**Image-FSC ↑**
Hydra $K = 3$	Replica 1	**1.00**	0.24 ± 0.03	0.393 ± 0.001	0.364 ± 0.001
	Replica 2	**1.00**	0.25 ± 0.03	0.394 ± 0.001	**0.367 ± 0.001**
	Replica 3	0.75	0.15 ± 0.02	0.388 ± 0.001	**0.367 ± 0.001**
Hydra $K = 5$	Replica 1	**1.00**	0.25 ± 0.03	**0.396 ± 0.001**	**0.367 ± 0.001**
	Replica 2	**1.00**	0.27 ± 0.02	0.394 ± 0.001	0.363 ± 0.001
	Replica 3	0.89	0.18 ± 0.02	**0.396 ± 0.001**	0.365 ± 0.001
DRGN-AI $K = 1$	Replica 1	0.59	0.0242 ± 0.003	0.0396 ± 0.005	0.042 ± 0.004
	Replica 2	0.49	0.022 ± 0.002	0.040 ± 0.004	0.040 ± 0.004
	Replica 3	0.53	0.026 ± 0.005	0.038 ± 0.004	0.041 ± 0.005
CryoDRGN2	—	0.36	0.08 ± 0.01	0.070 ± 0.006	0.3 ± 0.1
CryoSPARC	—	**1.00**	**0.284**	0.367	0.338

Table 3: Extension of Table 1 The classification accuracy is evaluated for each method using the adjusted Rand index (ARI). To evaluate each method's reconstruction quality, we use the mean area under the Fourier shell correlation (FSC) curve for 20 images per class (we report ±1 standard deviation). We **bold** the best result, and underline the best result for the second best method.

	Rotation Error ↓			**Translation Error ↓**			**Resolution at 0.5 FSC ↑**		
K	`6up6`	`6id1`	`4cr2`	`6up6`	`6id1`	`4cr2`	`6up6`	`6id1`	`4cr2`
1	122.73	80.07	1.21	4.754	19.144	0.005	44.31	82.29	9.14
3	114.82	1.19	1.23	2.936	0.022	0.013	9.29	9.14	9.14
5	123.43	1.18	1.22	2.812	0.020	0.016	9.29	9.14	9.14

Table 4: Quantitative results on the `tomotwin` dataset (30k particles) for Hydra with varying K. Metrics include median rotational errors (in degrees), median translation errors (in pixels), and resolution in Å at an FSC cutoff of 0.5, where we compare the ground truth volume to a backprojected volume using the predicted poses (Nyquist limit is at 9.00 Å). We note high pose errors for structure `6up6` despite its good reconstruction quality, likely indicative of pose ambiguity induced by the symmetry of the molecule.

	ARI ↑	**Rotation Error ↓**			**Translation Error ↓**			**Per-Image AUC ↑**		
Method		Splice	Ribo	Spike	Splice	Ribo	Spike	Splice	Ribo	Spike
Hydra ($K = 3$)	**0.997**	**1.69**	0.67	**0.85**	**0.141**	**0.003**	**0.003**	**0.373**	**0.429**	**0.441**
DRGN-AI ($K = 1$)	0.994	1.87	34.18	100.09	0.155	0.494	0.023	0.353	0.064	0.206
CryoDRGN2	0.986	2.06	**0.59**	0.99	0.611	0.006	0.011	0.357	0.424	0.416
CryoSPARC	0.972	1.94	1.34	1.45	0.367	0.018	0.020	0.324	0.355	0.353

Table 5: Quantitative results on the `ribosplike` dataset. Metrics include particle classification accuracy (Adjusted Rand Index, ARI), median rotational error (in degrees), median translation error (in pixels), and reconstruction quality (per-image area under the FSC curve). Hydra outperforms state-of-the-art methods at jointly capturing conformational and compositional heterogeneity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduce a new neural-based model to cope with compositional and conformational heterogeneity in cryo-EM. Our contributions are supported by results on synthetic and experimental datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are mentioned in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-1816

Answer: [NA]

Justification: The paper does not introduce new theoretical results requiring proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: `https://hydra.cs.princeton.edu/`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details are provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the results obtained with 3 replica for each experiment in Table 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

https://doi.org/10.52202/079017-1816

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information in experiments section and in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of the authors' knowledge, the paper respects the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We do not foresee any potential malicious applications or uses of the work described in this paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-1816

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The structures from the PDB are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-1816

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1816