Learning the Shared Latent Structure of Cryo-EM Inverse Problems

Anonymous authors
Paper under double-blind review

Abstract

Understanding biomolecular function requires characterizing three-dimensional structure and conformational variability, for which cryogenic electron microscopy (cryo-EM) is a powerful tool. Cryo-EM produces noisy two-dimensional particle images in which molecular structure, particle pose, and imaging conditions are jointly encoded. Recovering biological information from these observations spans inverse problems including pose estimation, heterogeneity analysis, and reconstruction. These tasks are usually addressed independently despite shared protein structure and image formation, motivating a common representation across inverse problems. Since biologically meaningful differences may be spatially localized, we study self-distillation with complementary global and masked patch objectives. Because experimental dataset discrimination can reflect acquisition-specific cues, we develop controlled synthetic and experimental evaluations and find that structural discrimination benefits from dense spatial features. We introduce the *EMber* family of vision transformers, pre-trained on >50 million experimental particle images from 214 EMPIAR datasets together with simulated projections from 268,000 molecular assemblies. *EMber* improves upon existing methods at conformational and compositional discrimination by 14.85 and 14.26 percentage points, estimates conformational populations to within 1.59pp error with a linear probe, and produces particle assignments yielding reference-quality reconstructions in fewer epochs. We demonstrate how to learn representations which capture biological information for use across multiple cryo-EM inverse problems.

1 Introduction

The functions of proteins are closely tied to their three-dimensional structure. Proteins, however, are dynamic objects. They occupy ensembles of conformations that result in markedly different function. Characterizing these structural ensembles is a central problem in structural biology.

Cryogenic electron microscopy (cryo-EM) enables the study of these ensembles by imaging large numbers of individual protein particles, each producing a two-dimensional projection of a three-dimensional molecular structure with noise levels one to three orders of magnitude larger than those encountered in natural images. Each particle image jointly reflects the underlying molecular conformation, particle pose, and acquisition-dependent imaging conditions. As such, recovering structural information requires reasoning from weak observations where the sources of variation are unknown and entangled.

Cryo-EM workflows address this ambiguity through a multitude of computational tasks, including particle classification, pose estimation, heterogeneity analysis, and 3D reconstruction. While these tasks have different objectives, they repeatedly recover and use overlapping structural and imaging information from the same particle images. Nevertheless, computational methods are largely applied on a per-experiment basis, while learned models are typically developed for a particular task or molecular system. Because proteins share common structural features, training a distinct model for each system can be redundant. This raises a broader question: **can general protein representations be learned from diverse cryo-EM data and reused across proteins and downstream tasks?**

This question encapsulates an under-specified representation learning problem. Ground truth labels describing the molecular state of individual particles are rarely available. At the same time, not all variation between particle images is equally useful. For example, acquisition-specific signals such as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sample preparation procedures can distinguish datasets without reflecting differences in molecular structure, while biologically meaningful conformational changes may affect only a small region of a projected molecule. A useful general protein representation must preserve molecular information in the presence of strong imaging variation while retaining spatially localized features that can capture subtle structural differences.

The scarcity of structural labels at a particle level makes self-supervised learning a natural learning setting. However, different self-supervised objectives, e.g. constrastive, reconstruction-based, or self-distillation, impose different biases on the information is preserved or discarded in a representation. We consider which learning objective is better suited to the qualities of cryo-EM data and the goal of learning generalizable protein representations.

We train the *EMber* family of vision transformers, ranging from 21M to 1.1B parameters, on >50 million experimental particles from 214 EMPIAR datasets together with simulated projections from 268,000 molecular assemblies. We construct self-supervised views directly through the cryo-EM forward model and combine exact structural correspondence from simulation with the diversity of experimental data (Sections 4). Across controlled synthetic and experimental evaluations, *EMber* improves conformational and compositional discrimination by up to 14.85 and 14.26 percentage points, respectively, while showing that dense spatial features retain structural information lost by global pooling (Sections 6.1). Frozen *EMber* representations further estimate conformational populations to within 1.59 percentage points using only a linear probe (Section 6.3), and can be inserted into a reconstruction pipeline to produce particle groupings yielding near-reference reconstructions after two pose-search epochs (Section 6.4). Together, these results show that a single pre-trained particle representation can be reused across several cryo-EM inverse problems. Anonymized code is available at `https://anonymous.4open.science/r/EMber-B221`.

2 Cryo-EM as a Shared Inverse Problem

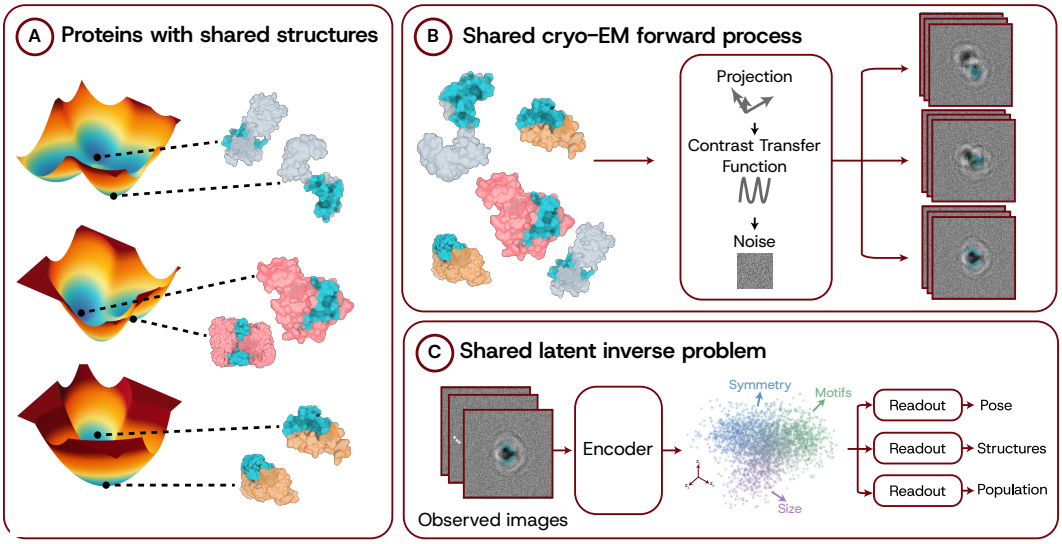


Figure 1: **Cryo-EM overview and shared inverse problems**. (1) Proteins share similar structural features across molecular systems. (2) Cryo-EM images are generated with an image formation process involving projection, imaging effects, and measurement noise. (3) Downstream tasks attempt to recover different variables of the same underlying generative process, relying on overlapping structural and imaging information.

A cryo-EM experiment forms noisy, two-dimensional projections of three-dimensional molecular structures. Let $X_k \in \mathbb{R}^{N\times 3}$ be a molecular state with population distribution $\pi(X_k)$. $\rho(X_k) : \mathbb{R}^3 \to \mathbb{R}$ is the corresponding electron density. For particle i, the image formation process can abstractly be written as

$$y_i = \mathrm{PSF}_i \star \mathcal{T}_{t_i} \mathcal{P}_{R_i} \rho(X_{k_i}) + \epsilon_i \tag{1}$$

where $R_i \in SO(3)$ is the particle's orientation, $\mathcal{P}_{R_i}$ rotates and projects the density along the viewing direction, $t_i \in \mathbb{R}^2$ is an in-plane translation applied by $\mathcal{T}_{t_i}$, PSF_i is the point spread function, and ϵ_i is measurement noise. The Fourier transform of PSF_i is the contrast transfer function (CTF), C_i.

The image formation process causes ambiguity in structural inference. Differences between any two images may be due to molecular state, pose, imaging conditions, or measurement noise. Therefore, recovering biological information from cryo-EM images necessitates reasoning over different latent components of the same underlying generative process.

Many cryo-EM tasks can be viewed as different inverse problems over shared latent variables. Even when variables such as molecular structure, pose, and imaging conditions are not the target of the task, they confound an estimator. For example, estimating the pose R_i requires determining which orientation of a candidate molecular structure is consistent with y_i. On the other hand, discriminating whether two particles arise from different molecular states requires distinguishing structural variation from differences in pose and imaging conditions. Reconstruction recovers $\rho(X_k)$ by combining particles according to both their orientations and their underlying structural states. Population estimation operates at the level of the structural ensemble, aggregating evidence about which molecular states are compatible with the observed particles in order to estimate $\pi(X_k)$.

This overlap in required information to solve a task motivates learning a common representation, z_i, with one encoder, $f_\theta(y_i)$ for every downstream task. Task-specific decoders and inference procedures can then re-organize learned concepts to solve each task. A useful representation should then preserve molecular information despite large variation in pose, imaging conditions, and noise, while remaining sensitive to biologically meaningful differences between molecular states. Moreover, these differences can be spatially localized. For example, a conformational change can be limited to a small region of a projected molecule. A global representation may discard small-scale variation in favor of larger elements of variation. We explore how to pose a learning task to create general protein representations such that they can be reused across diverse tasks and for any protein.

3 Related Works

Learning solutions to cryo-EM inverse problems. Machine learning has been used to solve individual tasks in cryo-EM, from supervised particle picking with Topaz and crYOLO (Bepler et al., 2019; Wagner et al., 2019) to direct structural inference. CryoDRGN (Zhong et al., 2021a) represents density, conditioned on a latent conformation, with an implicit neural representation, while e2gmm (Chen & Ludtke, 2021) represents heterogeneous structures with Gaussian mixtures; both assume known particle poses. CryoDRGN2 (Zhong et al., 2021b) and CryoDRGN-AI (Levy et al., 2025) extend neural heterogeneous reconstruction to unknown poses, while CryoSPHERE (Ducrocq et al., 2025) models heterogeneity through approximately rigid motions of learned structural segments. Other methods specialize in conformational inference, such as CryoSBI (Dingeldein et al., 2025), or pose estimation and reconstruction with CryoFastAR (Zhang et al., 2025). These approaches demonstrate the value of learning for cryo-EM, but are generally built for a specific inference problem. Our work aims to learn a representation that can be reused across proteins and downstream tasks.

Self-supervised representation learning. Modern self-supervised learning largely differs in how it defines agreement across views and what information the objective preserves. Contrastive methods such as MoCo (He et al., 2019) learn invariances by matching augmented views while contrasting different images. DINO (Caron et al., 2021) removes explicit negatives through self-distillation and produces spatially meaningful patch features in vision transformers. Masked-image methods such as MAE (He et al., 2021) reconstruct missing image content, while iBOT (Zhou et al., 2021) combines image-level self-distillation with masked prediction of patch representations. DINOv2 and DINOv3 (Oquab et al., 2024; Siméoni et al., 2025) scale this family of objectives to large datasets and models while retaining strong global and dense features. We adopt this self-distillation paradigm, but construct the training views and invariances using the cryo-EM forward model.

Representation learning for cryo-EM. Self-supervised learning has been applied across the cryo-EM pipeline. DRACO, CryoMAE, and cryo-EMMAE (Shen et al., 2024; Xu et al., 2024; Zamanos et al., 2025) learn from micrographs for particle localization and denoising, CryoLVM (Fu et al., 2026) learns representations of reconstructed density maps for restoration, and CryoSiam (Stojanovska et al.,

2025) operates on tomographic data. Closer to our setting, Cryo-IEF (Yan et al., 2026) pre-trains on picked single-particle images using a MoCo-style contrastive objective over augmented views, while CryoUNI (Dai et al., 2026) learns particle representations through paired denoising and reconstruction. These methods show transfer beyond their pre-training objectives, but their evaluations largely remain within one or two stages of the pipeline or set of particle-level properties. However, Cryo-IEF evaluates structural classification, pose clustering, and image-quality assessment. Our work further probes what biological information a general particle representation retains across conformational and compositional discrimination, population estimation, and particle grouping for reconstruction.

4 Learning through the Cryo-EM Forward Model

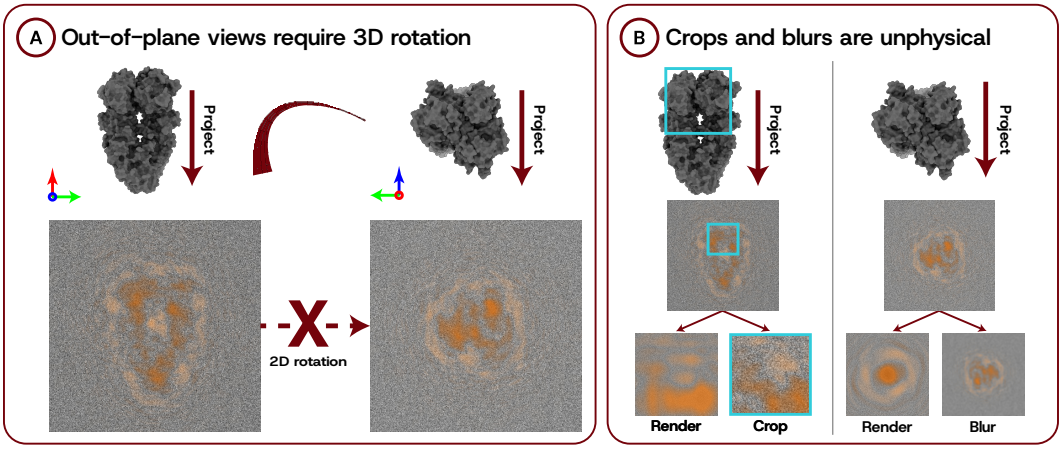


Figure 2: **Augmentations can be unphysical**. (A) 2D rotations of an image do not span all possible rotations of a 3D protein. (B) Image-space crops and blurs do not translate to the cryo-EM forward model.

Many self-supervised learning methods use multiple views of the same observation as a training signal, using augmentations that alter image appearance while preserving semantic content. When explicitly rendering new observations is impractical, as in natural scenes, image-space augmentations such as crops, rotations, color perturbations, and blur provide useful proxies for variation in the imaging process. For example, a 2D rotation corresponds to an in-plane camera rotation, while Gaussian blur can approximate optical or motion blur.

The same strategy, however, is less natural for cryo-EM. Transformations of an already projected particle image do not, in general, correspond to varying the latent variables of the cryo-EM forward model while holding structure fixed. For example, rotating a particle image, as in Cryo-IEF, changes only its in-plane orientation and cannot generate a different out-of-plane projection (Figure 2). Similarly, Gaussian blur does not reproduce a change in defocus, whose effect is governed by the oscillatory contrast transfer function. Thus, generic image-space augmentations may induce invariances that do not correspond to physically realizable variation in cryo-EM.

However, high-fidelity explicit forward models for cryo-EM already exist, such as CryoJAX (O'Brien et al., 2026). Instead of approximating changes to the image formation process with 2D augmentations, we can directly render alternate observations of the same structure:

$$y_i \sim p(y \mid X_k, R_i, C_i, t_i, \ldots) \tag{2}$$

while varying the latent quantities that determine how a fixed molecular structure is observed. We engineered CryoJAX to render these views fast enough to run inside the training loop. This lets us train models with SSL with grounded positive pairs without the cost generally associated with rendering.

Given this, the remaining question is which SSL objective should learn from these pairs. Contrastive objectives additionally require defining negative examples, which is difficult in cryo-EM. Two different particle images can arise from the same molecular structure, while their visual similarity or dissimilarity may instead be a function of their pose or imaging conditions. We use DINOv3 (Siméoni et al., 2025)

for image-level self-distillation and complement it with the iBOT masked-patch objective (Zhou et al., 2021) to supervise spatially resolved features without needing explicit negative pairs.

5 Pre-training Data and View Construction

Constructing the targets described in the previous section at scale presents a practical tradeoff. Simulated images provide precise control over structure and imaging conditions, but cannot capture the full acquisition statistics of experimental data; experimental particles provide this realism, but their underlying structures and correspondences are unknown. We pre-train on simulated data with marginal amounts of experimental images. Afterwards we introduce equal amounts of experimental and simulated data during a mid-training phase, using a different view-construction strategy for each.

Simulated and experimental training data. For simulated data, we use OpenProteinSet, a processed snapshot of the PDB (Berman et al., 2000; Ahdritz et al., 2023) released as part of OpenFold (Ahdritz et al., 2024). After filtering, our training set consists of 268k molecular assemblies determined across a range of experimental modalities, including X-ray crystallography and cryo-EM. We render particle images online using CryoJAX (O'Brien et al., 2026), independently sampling poses, translations, CTF parameters, noise levels, and Å/px values during training. We curate 51M picked single-particle images from 214 EMPIAR (Iudin et al., 2023) entries for the mid-training phase. The full filtering and preprocessing procedure for both data sources is described in Appendix A.

View Construction. The two data sources provide different information about correspondence between particle images. For simulated data, the underlying structure X_k is known, so we construct paired views by rendering the same X_k under independently sampled imaging conditions and noise, and with probability $p = 0.5$ independently sample the pose of each view. With this, we are able to retain exact structural correspondence even when two views represent substantially different projections of the molecule.

For experimental particles, the underlying X_k is unknown and exact structural pairs cannot be constructed. Instead, with probability $p = 0.5$ we pair a particle with another particle from the same EMPIAR entry. This is a weaker correspondence than we can achieve with synthetic data. Particles from the same entry generally arise from the same molecular system, but may differ in conformation or composition. Overall, these datasets are complementary. Simulated views provide strong structural correspondence, while experimental views expose the model to realistic variation that can be absent from the simulator.

Mixing simulated and experimental data. We find that introducing experimental data at the start of training destabilizes optimization and can lead to representation collapse. Instead, we first pre-train simulated data with marginal amounts of experimental images and gradually introduce experimental particles. The first 50k steps of training use a 93.75:6.25 ratio between simulated particles and experimental particles. Over the next 50k steps of training the experimental fraction is increased on a cosine schedule to a stochastic 50:50 mixture, which is then used for the remainder of training.

Pre-training objective and architecture. We instantiate the learning procedure described in the previous section using DINOv3 with the iBOT masked-patch objective. The student and EMA teacher are vision transformers with a patch size of 16×16 pixels, a learnable `[CLS]` token, and four register tokens (Darcet et al., 2024). We train ViT-S (21.4M), ViT-B (85.3M), ViT-L (304.2M), and ViT-g (1.1B) variants, collectively called the EMber family. Complete loss definitions, architectural details, regularization terms, and optimization hyperparameters are provided in Appendix B.

6 Evaluating What *EMber* Learns

We aim to evaluate *EMber* comprehensively so we can understand what information is accessible from the learned representations, how that information can be reused, and ultimately, performance on downstream tasks. Our evaluations span non-parametric use of frozen features, linear probes, full fine-tuning, and integration into a reconstruction pipeline. First, we use experimental data to assess if frozen representations contain molecular information and if that information is global or

spatially distributed. Then, we test how the representation can be adapted to discriminate between subtle conformational changes, if representations learning at an image level can be aggregated to recover conformational populations, and whether the learned representation can be reused for particle grouping during reconstruction. For each evaluation, additional details and hyperparameters are provided in Appendix C.

6.1 Probing Structural Information in Experimental Images

We first test if the representations learned by *EMber* can organize experimental particles consistently enough to recover dataset identity without learning task-specific readouts. Following prior work, we build a k-nearest neighbor classifier over 29 experimental particle datasets from CryoPPP (Dhakal et al., 2023). As shown in Figure 6, models are able to easily discriminate between EMPIAR datasets in latent space. *EMber* ViT-S reaches 92.17% Top-1 accuracy and 99.77% Top-5 accuracy, with larger models saturating quickly. At comparable model sizes, *EMber* improves over Cryo-IEF and CryoUNI.

However, this result is difficult to interpret biologically. EMPIAR datasets vary the molecule being imaged and acquisition-specific factors. This experiment shows that the representation captures strong and reproducible signal in experimental particles, but not which part of that signal is structural. Next, we design a more controlled experimental setting to isolate molecular composition from imaging conditions.

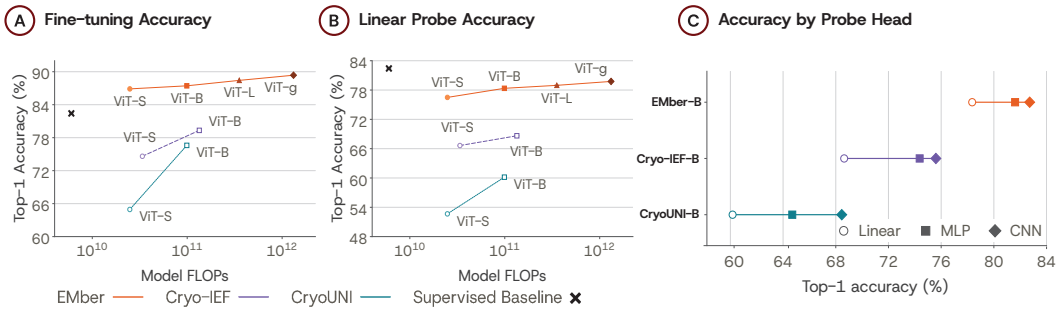


Figure 3: **Compositional heterogeneity classification on CAHRA experimental data.** (A) Fine-tuning accuracy versus model FLOPs, with a supervised ResNet-18 reference. (B) Accuracy versus model FLOPs when linear probing frozen backbones, with a reference supervised ResNet-18. (C) Pooled linear/MLP readouts with a convolutional head that preserves spatial information.

Compositional discrimination under controlled imaging conditions. To directly test whether frozen representations retain molecular information, we use the CAHRA compositional heterogeneity challenge (Feathers et al., 2026) which contains labeled particles from six proteins collected under similar microscopy conditions which are not seen during pre-training. We fine-tune the pre-trained backbones for this task, as well as multiple probes on frozen features. On pooled features, we train linear and MLP probes, and on the unpooled two-dimensional patch features we train a convolutional probe. Full details are in Appendix C.

The results when fine-tuning the encoders is shown in Fig. 3. Across all model sizes *EMber* is the only model which outperforms the baseline tuned ResNet-18. Fig. 3B shows that compositional information is already linearly accessible from frozen representations. *EMber*-g reaches 79.76% with a linear probe, only 2.65 percentage points below the fully supervised ResNet-18, and outperforms Cryo-IEF and CryoUNI at comparable model sizes. Fig. 3B tests whether pooling discards useful information. Across all pre-trained backbones, the parameter-matched convolutional probe outperforms the MLP. *EMber*-B reaches 82.72%, comparable to the supervised baseline. As the trend is similar for Cryo-IEF and CryoUNI, this result supports retaining the spatial organization of patch features rather than attributing the gain solely to our patch-level objective.

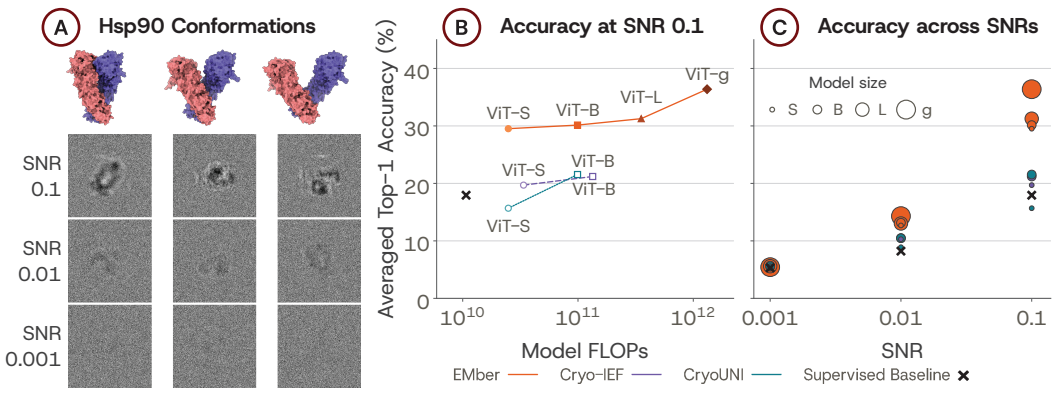


Figure 4: **Fine-tuning for Hsp90 conformation classification.** We discretize Hsp90 conformations according to CMA and CMB chain-opening angles, each with 20 classes, and average accuracy across the two prediction heads (chance: 5%). (A) Representative conformations and rendered particles at different SNRs. (B) Top-1 accuracy at SNR 0.1 versus model forward FLOPs. (C) Accuracy across SNRs. Model differences largely disappear at SNR 0.001, where all methods approach chance.

6.2 Adapting Representations to Conformational Variation

The previous experiments ask what molecular information can be extracted from frozen representations. We now evaluate whether pre-training provides a useful initialization when the full representation is adapted to distinguish subtle conformational changes within the same molecule. Since pose is unknown, this is a difficult task. Methods based on template search or simulation-based inference require dense coverage of pose space (Giammar et al., 2025; Dingeldein et al., 2025). We purposefully evaluate in an under-sampled pose regime.

We study heat shock protein 90 (Hsp90), which is a standard system for studying conformational heterogeneity (Seitz et al., 2019; Dingeldein et al., 2025). Two opening angles, CMA and CMB, parameterize the transition from open to closed states. We discretize this continuous transition spectrum into 20 bins, yielding 400 conformations, and jointly fine-tune the backbone with one readout head per angle. The training set contains 200,000 images—only 500 per conformation—so each state is observed under a sparse set of poses. We create validation sets at SNR 0.1, 0.01, and 0.001 which measure how performance varies as signal is removed.

At SNR 0.1, nearly every pre-trained model outperforms the ResNet-18 trained from scratch, demonstrating that pre-training excels when task-specific data sparsely sample pose space (Fig. 4B). *EMber*-S reaches 29.52% accuracy compared with 17.95% for the ResNet-18 baseline, while scaling to *EMber*-g raises accuracy to 36.39%. The full model comparison is shown in Fig. 4B.

However, representation quality is limited by the epistemic identifiability of noisy images. As SNR decreases, all models degrade sharply and approach chance at SNR 0.001 (Fig. 4C). Thus, pre-training makes fine-grained conformational information easier to learn from sparse pose coverage, but cannot recover structural signal once the observations themselves are noise-dominated.

6.3 From Particle Representations to Conformational Populations

Particle-level classification is not always the quantity of biological interest. Since both pose and noise can make individual images ambiguous Evans et al. (2025), we measure whether representations across many particles can be aggregated to recover a property of the structural ensemble. Specifically, we estimate the conformational population distribution $\pi(X_k)$ from image-level predictions.

We use the SARS-CoV-2 Spike molecular dynamics dataset from CryoBench Wieczór et al. (2023); Jeon et al. (2024) and group the conformational ensemble into Spike-Up and Spike-Down states (Appendix Fig. 16). We train a linear classifier on frozen features using 50k simulated images, calibrate on held-out data, and pass the probe's outputs to the expectation-maximization procedure from Evans et al. (2025). Since a linear probe's logits are not calibrated likelihoods, we perform Platt

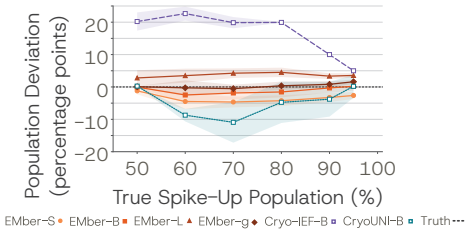


Figure 5: **Estimating Spike populations from cryo-EM images.** The plot shows estimated minus true Spike-Up population in the mixed-SNR setting. Each point is an average across three probe training seeds and 10 population sampling seeds, with shaded regions showing ± 1 standard deviation. *EMber* tracks the true population more closely than Cryo-IEF and CryoUNI.

Model	Top-1 (%)	Top-5 (%)
EMber-S	92.17	99.77
EMber-B	94.47	**99.94**
EMber-L	93.63	99.60
EMber-G	**95.77**	99.86
Cryo-IEF-S	81.78	98.59
Cryo-IEF-B	81.12	98.33
Cryo-UNI-S	82.68	98.79
Cryo-UNI-B	85.88	98.98

Figure 6: **k-nearest neighbor EMPIAR dataset discrimination.** Top-1 and Top-5 accuracies (%) across 29 experimental datasets from CryoPPP. High dataset classification accuracy may reflect both molecular structure and acquisition-specific information.

scaling on a held-out calibration set before the EM procedure. The evaluation populations contain images spanning SNR 0.1, 0.01, and 0.001. Results stratified by SNR, additional training sizes, and the complete EM setup are provided in Appendix C.4.

Population estimates from evaluated models reveal a larger separation between representations than particle-level accuracy alone would suggest. In our mixed-SNR setting, *EMber*-B reaches a mean absolute population error of 1.59 ± 0.26 percentage points, compared with 16.28 ± 0.61 for Cryo-IEF-B and 5.64 ± 1.73 for CryoUNI-B (Fig. 5). *EMber*-g reaches 1.11 percentage points, although this scaling trend is not consistent across the other training set sizes reported in Appendix C.4. Cryo-IEF systematically overestimates the Spike-Up fraction while CryoUNI underestimates it. Across all tested mixtures, *EMber* remains much closer to the true population across the tested mixtures. These results demonstrate that representations derived from our pre-training setup can be useful even when individual particles are ambiguous, and individually weak image-level evidence can be aggregated to recover an ensemble-level quantity.

6.4 Representations for Heterogeneous Reconstruction

Finally, we measure whether representations learned with our objective can be reused inside a coupled cryo-EM inverse problem. Heterogeneous reconstruction requires estimating both particle pose and grouping particles by structure. Since reconstruction quality by itself does not reveal which part of the inference failed, we design a two-phase protocol to isolate particle grouping.

In Phase 1, we run CryoDRGN-Ab-initio (Levy et al., 2025) with unknown poses while replacing its encoder with *EMber* or a baseline encoder. In Phase 2, we cluster the learned latent space with k-means++ (Arthur & Vassilvitskii, 2007; Pedregosa et al., 2011) and reconstruct each group in RELION (Scheres, 2012) with ground truth poses. As pose is unknown during pre-training and fixed during the final reconstruction, differences in reconstruction quality reflect how well each representation organizes particles by structure.

We compare frozen ViT-B backbones from *EMber*,Cryo-IEF, and CryoUNI with the default CryoDRGN CNN encoder and its lookup table baseline. All methods use the same CryoDRGN-Ab-initio pose-search procedure, latent dimensionality, particle clustering, and RELION reconstruction protocol. Oracle reconstructions use ground truth structure assignments and poses. Complete details on head architectures, clustering settings, and FSC curves are provided in Appendix C.5.

Tomotwin-8. We evaluate compositional heterogeneity using eight structures from the CryoBench Tomotwin system (Rice et al., 2023; Jeon et al., 2024). After a single pose-search epoch, *EMber* already produces clusters that reconstruct to 3.74 ± 0.11 Å—only 0.18 Å from the Oracle resolution of 3.56 Å (Fig. 7). By epoch two, it reaches 3.57 ± 0.03 Å, effectively matching the reference. Cryo-IEF reaches comparable resolution after two epochs, while the CNN encoder requires three. The full distribution across reconstructed structures can be seen in Fig. 7A. This results shows

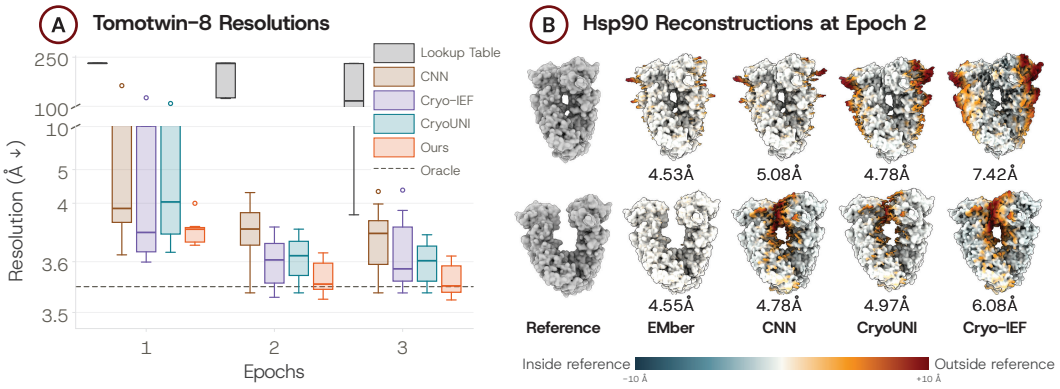


Figure 7: **Particle grouping measured through reconstruction resolution.** After each epoch of CryoDRGN-AI pose search, particles are clustered in the learned latent space and reconstructed using ground truth poses. (A) Tomotwin-8: Boxes show the spread in FSC=0.5 resolution across eight structures. The dashed line is an Oracle reconstruction obtained with ground truth group assignments. Notably, *EMber* is already near the Oracle after one epoch. *EMber* reaches near reference resolution within two epochs, and has less variation across the reconstructed clusters compared to baseline methods. (B) Hsp90: Reconstructions visualized after 2 epochs of CryoDRGN-AI pose search. Surfaces are colored by signed distance to the reference surface (negative inside, positive outside).

that the *EMber* pre-training setup produces a representation that is already close to the underlying compositional grouping without any supervision applicable variables.

Hsp90. Finally, we revisit the open and closed Hsp90 states from Section 6.2. This provides a conformational counterpart to Tomotwin-8, where the structural difference is more localized. After two epochs, *EMber* reconstructs the closed and open states at 4.53 Å and 4.55 Å, respectively, within 0.05 Å of the corresponding Oracle resolutions. Although CryoUNI produces the best reconstruction after the first epoch, its grouping does not converge to the Oracle solution by epoch two, visualized in Fig. 7B. All methods recover the closed state while only *EMber* recovers the open state at near-reference resolution while visibly resolving the gap between the two Hsp90 arms.

Across both compositional and conformational heterogeneity, representations learned with *EMber* organize particles in a way that can be directly reused by a downstream reconstruction procedure with minimal re-training.

7 Conclusion

We introduce *EMber*, a family of general cryo-EM representations learned with self-supervision through the cryo-EM forward model. Our evaluations provide a framework for studying what cryo-EM representations retain: distinguishing dataset discrimination from evidence of molecular information, comparing global and spatial readouts, and testing whether those representations support downstream structural inference. Across frozen probing, fine-tuning, population estimation, and reconstruction, *EMber* sets a new performance-compute frontier across several structural inference problems, while our analyses show that biologically relevant information is retained in both global and spatial features. Our results provide a practical recipe for scaling representation learning on cryo-EM data: use the forward model to construct physically meaningful supervision, combine structurally precise simulated data with diverse experimental particles, and reuse the resulting representation across downstream analyses. There remains substantial room to improve how simulated and experimental data are mixed, how spatial features are supervised under extreme noise, and how much experimental diversity can be recovered from currently deposited particle data. Nevertheless, the breadth of transfer we observe suggests that representation learning on noisy cryo-images alone can learn to solve multiple inverse problems without explicit supervision, and opens the door to extending the same approach to micrographs, tomography, and other structural inference problems.

References

Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Daniel Berenberg, Ian Fisk, Andrew M. Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. OpenProteinSet: Training data for structural biology at scale, August 2023. URL http://arxiv.org/abs/2308.05326. arXiv:2308.05326 [q-bio.BM].

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Andrew Nowaczynski, Bowen Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Akintunde Ojewole, Murat E. Guney, Stella Biderman, Andrew M. Watkins, Sae Ra, Pablo R. Lorenzo, Lucas Nivon, Brian Weitzner, Y. A. Ban, Shiyang Chen, Ming Zhang, Cheng Li, Seonhee L. Song, Yuning He, Peter K. Sorger, Emad Mostaque, Zhen Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, August 2024. doi: 10.1038/s41592-024-02272-z.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pp. 1027–1035, USA, January 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-89871-624-5. URL https://dl.acm.org/doi/10.5555/1283383.1283494.

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature methods*, 16(11):1153–1160, 2019.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

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.

Haizhao Dai, Qihe Chen, Lingqi Li, Yingjun Shen, Zhenyang Xu, Minzhang Li, Yufan Xie, Jingjing Zheng, Zhijie Liu, Liping Sun, Yuan Pei, Jiakai Zhang, and Jingyi Yu. Cryo-EM as latent structural landscape microscopy. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717737. URL https://www.biorxiv.org/content/10.64898/2026.04.10.717737v2.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers, April 2024. URL http://arxiv.org/abs/2309.16588. arXiv:2309.16588 [cs.CV].

Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. CryoPPP: A Large Expert-Labelled Cryo-EM Image Dataset for Machine Learning Protein Particle Picking. *bioRxiv*, pp. 2023.02.21.529443, February 2023. ISSN 2692-8205. doi: 10.1101/2023.02.21.529443. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9980126/.

Lars Dingeldein, David Silva-Sánchez, Luke Evans, Edoardo D'Imprima, Nikolaus Grigorieff, Roberto Covino, and Pilar Cossio. Amortized template matching of molecular conformations from cryoelectron microscopy images using simulation-based inference. *Proceedings of the National Academy of Sciences*, 122(23):e2420158122, June 2025. doi: 10.1073/pnas.2420158122. URL https://www.pnas.org/doi/10.1073/pnas.2420158122.

Gabriel Ducrocq, Lukas Grunewald, Sebastian Westenhoff, and Fredrik Lindsten. cryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=n8O0trhost.

Luke Evans, Lars Dingeldein, Roberto Covino, Marc Aurèle Gilles, Erik Thiede, and Pilar Cossio. Counting particles could give wrong probabilities in Cryo-Electron Microscopy, March 2025. URL https://www.biorxiv.org/content/10.1101/2025.03.27.644168v1. Pages: 2025.03.27.644168 Section: New Results.

J. Ryan Feathers, Robert C. Heeter, Geoffrey Woollard, Sonya M. Hanson, Pilar Cossio, Joel Greer, Tom Burnley, and Ellen D. Zhong. The CAHRA Challenge: A Community-Wide Assessment of Cryo-EM Heterogeneous Reconstruction Algorithms. *bioRxiv*, 2026. doi: 10.64898/2026.09.15.751515. URL https://www.biorxiv.org/content/early/2026/09/17/2026.09.15.751515.

Weining Fu, Kai Shu, Kui Xu, and Qiangfeng Cliff Zhang. CryoLVM: Self-supervised Learning from Cryo-EM Density Maps with Large Vision Models, February 2026. URL http://arxiv.org/abs/2602.02620. arXiv:2602.02620 [q-bio.QM].

Matthew David Giammar, Joshua Luke Dickerson, Laina Nicole Hall, and Bronwyn Ayla Lucas. Leopard-EM: An extensible 2DTM package to accelerate in situ structural biology. *bioRxiv*, 2025. doi: 10.1101/2025.08.26.672452. URL https://www.biorxiv.org/content/early/2025/08/29/2025.08.26.672452. _eprint: https://www.biorxiv.org/content/early/2025/08/29/2025.08.26.672452.full.pdf.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, November 2019. URL https://arxiv.org/abs/1911.05722v3.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs.CV].

Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023.

Minkyu Jeon, Rishwanth Raghu, Miro A. Astore, Geoffrey Woollard, J. Ryan Feathers, Alkin Kaz, Sonya M. Hanson, Pilar Cossio, and Ellen D. Zhong. CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM. In *Advances in Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=kP92Fyc6ry#discussion.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Axel Levy, Rishwanth Raghu, J. Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D. Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D. Zhong. CryoDRGN-AI: neural ab initio reconstruction of challenging cryo-EM and cryo-ET datasets. *Nature Methods*, 22(7):1486–1494, July 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02720-4. URL https://www.nature.com/articles/s41592-025-02720-4.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is Scalable for LLM Training, February 2025. URL http://arxiv.org/abs/2502.16982. arXiv:2502.16982 [cs.LG].

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs.LG].

M. J. O'Brien, D. Silva-Sánchez, G. Woollard, K. Je, S. M. Hanson, D. J. Needleman, P. Cossio, E. H. Thiede, and M. A. Astore. CryoJAX: a cryo-electron microscopy image-simulation library in JAX. *Acta Crystallographica Section D: Structural Biology*, 82(3):155–167, March 2026. ISSN 2059-7983. doi: 10.1107/S2059798326000550. URL https://journals.iucr.org/d/issues/2026/03/00/zi5008/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

L.-M. Peng, G. Ren, S. L. Dudarev, and M. J. Whelan. Robust Parameterization of Elastic and Absorptive Electron Atomic Scattering Factors. *Acta Crystallographica Section A*, 52 (2):257–276, 1996. ISSN 1600-5724. doi: 10.1107/S0108767395014371. URL https://onlinelibrary.wiley.com/doi/abs/10.1107/S0108767395014371.

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, June 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01878-z. URL https://www.nature.com/articles/s41592-023-01878-z.

Peter B. Rosenthal and Richard Henderson. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745, 2003. ISSN 0022-2836. doi: 10.1016/j.jmb.2003.07.013. URL https://www.sciencedirect.com/science/article/pii/S0022283603010222.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Sjors H. W. Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, December 2012. ISSN 1095-8657. doi: 10.1016/j.jsb.2012.09.006.

Evan Seitz, Francisco Acosta-Reyes, Peter Schwander, and Joachim Frank. Simulation of cryo-EM ensembles from atomic models of molecules exhibiting continuous conformations, December 2019. URL https://www.biorxiv.org/content/10.1101/864116v1. Pages: 864116 Section: New Results.

Yingjun Shen, Haizhao Dai, Qihe Chen, Yan Zeng, Jiakai Zhang, Yuan Pei, and Jingyi Yu. DRACO: A Denoising-Reconstruction Autoencoder for Cryo-EM, October 2024. URL https://arxiv.org/abs/2410.11373v2.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. URL https://arxiv.org/abs/2508.10104v1.

Frosina Stojanovska, Ricardo M Sanchez, Rasmus K Jensen, Julia Mahamid, Anna Kreshuk, and Judith B Zaugg. Cryosiam: self-supervised representation learning for automated analysis of cryo-electron tomograms. *bioRxiv*, pp. 2025–11, 2025.

Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, et al. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communications biology*, 2(1):218, 2019.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective, December 2024. URL http://arxiv.org/abs/2410.05192. arXiv:2410.05192 [cs.LG].

Miłosz Wieczór, Phu K. Tang, Modesto Orozco, and Pilar Cossio. Omicron mutations increase interdomain interactions and reduce epitope exposure in the SARS-CoV-2 spike. *iScience*, 26 (2):105981, February 2023. ISSN 2589-0042. doi: 10.1016/j.isci.2023.105981. URL https://www.sciencedirect.com/science/article/pii/S2589004223000585.

Chentianye Xu, Xueying Zhan, and Min Xu. CryoMAE: Few-Shot Cryo-EM Particle Picking with Masked Autoencoders, April 2024. URL http://arxiv.org/abs/2404.10178. arXiv:2404.10178 [q-bio.BM].

Yang Yan, Shiqi Fan, Fajie Yuan, and Huaizong Shen. A comprehensive foundation model for cryo-EM image processing. *Nature Methods*, 23(1):88–95, January 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02916-8. URL https://doi.org/10.1038/s41592-025-02916-8.

Andreas Zamanos, Panagiotis Koromilas, Giorgos Bouritsas, Panagiotis L. Kastritis, and Yannis Panagakis. Self-supervised learning for generalizable particle picking in cryo-EM micrographs. *Cell Reports Methods*, 5(7):101089, 2025. ISSN 2667-2375. doi: 10.1016/j.crmeth.2025.101089. URL https://www.sciencedirect.com/science/article/pii/S2667237525001250.

Jiakai Zhang, Shouchen Zhou, Haizhao Dai, Xinhang Liu, Peihao Wang, Zhiwen Fan, Yuan Pei, and Jingyi Yu. CryoFastAR: Fast Cryo-EM Ab Initio Reconstruction Made Easy, June 2025. URL https://arxiv.org/abs/2506.05864v2.

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, February 2021a. ISSN 1548-7105. doi: 10.1038/s41592-020-01049-4. URL https://www.nature.com/articles/s41592-020-01049-4.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4046–4055, Montreal, QC, Canada, October 2021b. IEEE. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00403. URL https://ieeexplore.ieee.org/document/9710804/.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, November 2021. URL https://arxiv.org/abs/2111.07832v3.

AI USE STATEMENT

In this work, we used generative AI tools to edit our writing, verify that the paper reads cleanly, to write visualization scripts, and to launch sweeps across disparate compute clusters. Any and all LLM-generated code, which was largely limited to visualizations for debugging and infrastructure-related scripts, were manually reviewed by at least two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility during peer review, we have released all the code for our project via Anonymous GitHub. This includes code to generate synthetic data, process experimental datasets, train the model, and all evaluation code. Model weights will be released upon publication of this work.