CRYOVERSE: A FOUNDATION MODEL ACROSS CRYO-EM OBSERVATION TYPES AND SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Cryogenic electron microscopy (cryo-EM) reconstructs three-dimensional macromolecular structures from noisy two-dimensional observations spanning diverse data types and spatial scales. Existing representation-learning approaches, however, typically specialize in individual observation types, limiting knowledge transfer across single-particle analysis (SPA) and cryogenic electron tomography (cryo-ET) workflows. We introduce CryoVERSE (Cryo-EM Versatile Encoder for Representations across Scales), a unified foundation model jointly pretrained on four pre-reconstruction 2D cryo-EM observation types: SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles. CryoVERSE uses a scale-flexible shared encoder with pose-aware contrastive learning and masked patch-level prediction to capture transferable global and spatial representations across these observation types and scales. Across diverse downstream evaluations, CryoVERSE improves molecular structure classification, projection-pose classification, and particle quality ranking, transfers effectively to dense particle localization in micrographs, and preserves structural identity and acquisition-angle information in cryo-ET. These results demonstrate that joint pretraining across cryo-EM observation types and scales can yield transferable global and spatial representations within a shared backbone, providing a unified representation basis for diverse pre-reconstruction cryo-EM analysis tasks.

1 INTRODUCTION

Cryogenic electron microscopy (cryo-EM) has become a central technology for resolving macromolecular structures and conformational states at high resolution (Nogales, 2016; Cheng, 2018). Single-particle analysis (SPA) recovers structures by combining projections of isolated particles, whereas cryogenic electron tomography (cryo-ET) reconstructs specimens from tilt series and enables structural analysis in native cellular environments (Wan & Briggs, 2016). Both workflows ultimately recover three-dimensional (3D) information from noisy two-dimensional (2D) observations.

Recovering reliable structures from these observations necessitates a sequence of consequential computational processing (Scheres, 2012; Punjani et al., 2017). Particle detection identifies particle images for downstream processing. Image curation selects high-quality particle projections. Classification groups projections with similar structures or conformations. And pose estimation aligns the projections for three-dimensional reconstruction. These decisions are particularly difficult because radiation damage constrains the electron dose, producing observations with weak contrast and low signal-to-noise ratios (Baxter et al., 2009). Therefore, errors may propagate through the processing workflow and compromise the quality of the recovered density maps.

Representations learned directly from pre-reconstruction observations may streamline cryo-EM data processing (Shen et al., 2024; Yan et al., 2026a). Complementary foundation models have also been developed for reconstructed 3D density maps (Zhou et al., 2025; Fu et al., 2026). Existing cryo-EM representation models typically operate on a specific type or scale of data, ranging from micrographs to SPA particles or 3D density maps. Consequently, it remains unclear whether one encoder can learn transferable representations across multiple pre-reconstruction observation types, including SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles.

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.

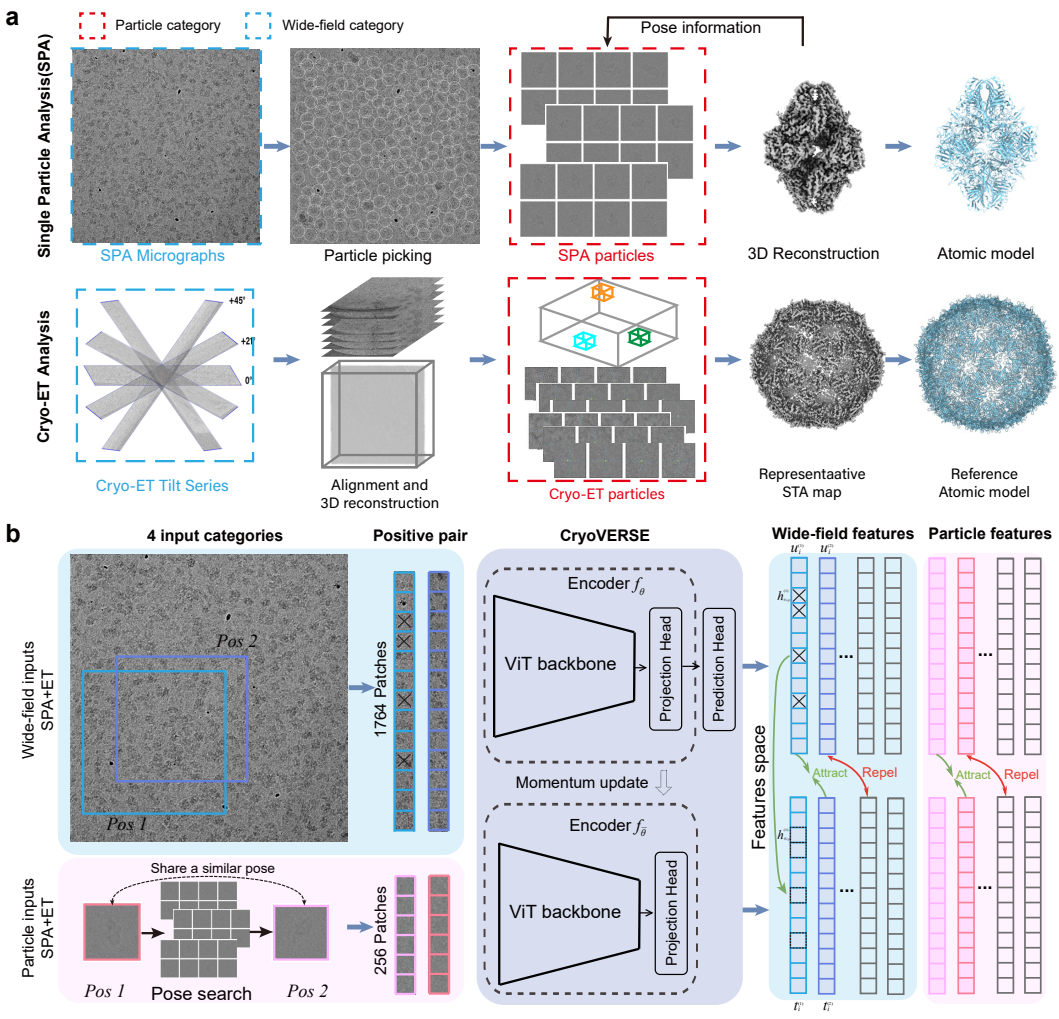

Figure 1: **Data and architecture of CryoVERSE pretraining.** **a**, Data sources used for CryoVERSE pretraining. The corpus comprises SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles. Blue and red dashed boxes denote wide-field and particle input categories, respectively. **b**, Architecture of CryoVERSE pretraining. Augmented crops from the same wide-field observation and particle images with neighboring views form positive pairs, whereas other pairs form negative ones. The masked image modeling (MIM) loss is applied to wide-field observations, including SPA micrographs and cryo-ET tilt images.

The four input categories differ markedly in spatial scale, signal density, and acquisition geometry, yet they are linked by the underlying imaging process (Figure 1a). SPA micrographs and particle crops differ in spatial scale. Cryo-ET tilt series comprise projections acquired at different orientations; individual tilt images and tilt-derived cryo-ET particles provide the corresponding pretraining observations (Wan & Briggs, 2016). Both SPA and cryo-ET observations are shaped by macromolecular structure, projection geometry, low-dose noise, and acquisition-dependent artifacts. Therefore, joint learning promises to transfer information across input categories.

We introduce CryoVERSE, a versatile cryo-EM encoder designed to learn transferable representations across observation types and spatial scales. The model is jointly pretrained on 143.7 million observations comprising SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles. Figure 1a situates these inputs within the SPA and cryo-ET workflows, while Figure 1b summarizes the architecture for CryoVERSE pretraining.

The main contributions of this work are summarized as follows:

- **A shared representation encoder for cryo-EM observations of different types and scales.** We introduce CryoVERSE, a shared vision transformer pretrained across four cryo-EM observation types: SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles. Input-aware view generation and scale-flexible positional encoding allow the same backbone to process different spatial scales and provide both global and spatial representations.
- **Pose-aware contrastive learning across distinct particle projections.** We introduce a pose-aware contrastive objective that relates distinct particle observations with neighboring orientations and weights their contributions according to angular distance. This objective incorporates cryo-EM projection geometry to augment conventional same-image contrastive learning.
- **Systematic evaluation of learning transfer across different cryo-EM data types and scales.** Pretrained on 143.7 million observations, CryoVERSE supports complementary downstream tasks requiring global or dense representations, including particle structure, pose and quality analysis, micrograph particle picking, and particle-type classification in cryo-ET. Controlled evaluations show cross-scale learning transfer and preservation of established particle-level capabilities relative to Cryo-IEF.

2 Related Work

Self-supervised visual representation learning. Modern self-supervised learning includes masked reconstruction, contrastive learning, and teacher–student distillation. MAE reconstructs masked image patches, MoCo v3 aligns augmented views, and iBOT predicts masked patch tokens under teacher supervision (He et al., 2022; Chen et al., 2021; Zhou et al., 2022). These objectives were developed for natural images and do not explicitly encode viewing-direction relationships among cryo-EM projections. CryoVERSE adapts global contrastive and masked patch-level learning to heterogeneous cryo-EM observations through input-aware views and pose-aware cross-particle positives.

Cryo-EM representation learning and particle-quality assessment. DRACO learns representations from cryo-EM movies and micrographs for micrograph curation and particle picking (Shen et al., 2024). Cryo-IEF learns SPA particle representations for structure, pose, and quality tasks, while CryoRanker adapts such representations for individual-particle quality prediction and ranking (Yan et al., 2026a). These approaches target particular observation types or processing stages rather than a unified representation across SPA and cryo-ET. CryoVERSE addresses this gap with a scale-flexible shared encoder jointly pretrained on SPA micrographs, SPA particles, cryo-ET tilt images, and tilt-derived cryo-ET particles. The shared backbone learns global and dense spatial representations across data scales, while the CryoRanker protocol tests whether broader pretraining preserves particle-quality information.

Particle picking and cryo-ET analysis. SPA particle-picking methods include Topaz, which uses positive–unlabeled classification, and CryoTransformer, which adopts a detection-transformer architecture (Bepler et al., 2019; Dhakal et al., 2024). Cryo-EMMAE instead clusters masked-autoencoder representations without coordinate annotations, while DRACO transfers pretrained micrograph features to particle localization (Zamanos et al., 2025; Shen et al., 2024). In cryo-ET, TomoTwin and DeepETPicker localize or classify macromolecules in reconstructed tomograms (Rice et al., 2023; Liu et al., 2024). These task-specific methods do not test whether one pre-reconstruction encoder can retain both wide-field spatial cues and localized structural information. CryoVERSE therefore targets representation transfer across 2D observation types rather than a new picking objective or a complete tomogram-analysis system.

3 Method

3.1 A shared encoder across cryo-EM observation types and scales

CryoVERSE jointly represents four pre-reconstruction cryo-EM observation types with a shared vision transformer. SPA particles and tilt-derived cryo-ET particles capture localized molecular

projections, whereas SPA micrographs and cryo-ET tilt images retain wide-field acquisition information. These inputs differ substantially in spatial extent, token-grid size, and the information useful for downstream tasks. Rather than introducing observation-specific encoder branches, CryoVERSE accommodates these differences through input-aware view generation and scale-flexible positional encoding. The shared backbone thereby preserves both global and dense patch representations across spatial regimes.

Let $m \in \mathcal{C}$ denote the input category, where $\mathcal{C} = \{\text{SPA particle}, \text{tilt-derived cryo-ET particle}, \text{SPA micrograph}, \text{cryo-ET tilt image}\}$. Two input-aware views $(x_i^{(1)}, x_i^{(2)})$ are processed by an online ViT-B/14 encoder f_θ and its momentum counterpart $f_{\bar{\theta}}$ (Figure 1b). Rotary position embeddings (RoPE) encode patch-grid coordinates without imposing a fixed grid size, allowing the shared backbone to process both particle crops and wide-field images (Su et al., 2024). Four learned register tokens further support the extraction of spatially resolved patch representations (Darcet et al., 2024).

The pretraining corpus contains 143,744,244 observations across the four input categories (Table 1). Because particle images substantially outnumber wide-field observations, we construct the training stream through batch-level hierarchical resampling. Particle and wide-field batches account for 65% and 35% of training batches, respectively, with a 4:1 SPA-to-cryo-ET ratio within each spatial regime. Scale-dependent batch sizes accommodate differences in token-sequence length. Detailed view-generation, batching, and sampling settings are provided in Appendix C.

Table 1: Pretraining corpus. Counts of collections/datasets and images in the pretraining corpus are presented. Data for batched training are resampled and the composition is indicated.

Input category	Collections/datasets	Images	Batch resampling
SPA particles	> 500	127,758,003	52%
Tilt-derived cryo-ET particles	97	15,571,232	13%
SPA micrographs	> 776	362,472	28%
Cryo-ET tilt images	42	52,537	7%
Total	—	143,744,244	100%

3.2 Pose-aware contrastive learning across projection views

Standard instance-level contrastive learning aligns two augmentations of the same image but does not exploit the known geometric relationships among cryo-EM projections. We extend this construction by allowing a particle observation to form a positive pair with a distinct particle whose viewing direction lies within a local angular neighborhood. This exposes the encoder to structural correspondence across independently observed projections rather than enforcing invariance only within one image. When reliable pose metadata or an eligible neighbor is unavailable, training falls back to the conventional same-image positive. The complete neighbor-selection procedure is provided in Appendix C.2.

Nearby viewing directions provide useful geometric supervision, but two distinct particles remain less reliable as a positive pair than two augmentations of the same observation because of conformational variation, image quality, and pose-estimation error. We therefore assign each cross-particle pair a confidence weight that decreases smoothly with its great-circle angular distance, while same-image pairs retain full weight. This formulation encourages locally smooth organization across projection directions while limiting the influence of uncertain cross-particle correspondences. The normalized bidirectional contrastive objective and weighting function are defined in Appendix C.2.

The pose-aware contrastive objective organizes global representations according to image identity and projection geometry. To additionally preserve spatially localized evidence in wide-field observations, we apply iBOT-inspired masked patch-level prediction (masked image modeling, MIM) to wide-field data including SPA micrographs and cryo-ET tilt images (Zhou et al., 2022). Register tokens complement this objective by reducing anomalous information concentration in spatial patch tokens (Darcet et al., 2024). Because both components were introduced to strengthen dense patch representations, we evaluate them jointly in the downstream ablation. KoLeo regularization further discourages crowding in the global representation space (Oquab et al., 2024). The complete

pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{KoLeo}}. \tag{1}$$

Here, $\mathcal{L}_{\text{MIM}}$ is applied only to wide-field-image batches and predicts momentum-encoder features at masked patches rather than reconstructing image pixels. Complete objective definitions and training settings are provided in Appendices C.2 and C.

3.3 Downstream adaptation at particle and micrograph scales

We adapt the same pretrained encoder to two supervised tasks with complementary spatial requirements. Particle-quality ranking requires a global representation of the structural quality of an individual particle, whereas micrograph particle picking requires dense features that preserve weak, spatially localized particle evidence. Both adaptations introduce task-specific prediction heads only during downstream fine-tuning, without adding input-specific backbone branches.

Particle-level quality ranking. For particle-quality ranking, a scalar prediction head maps the global representation of an SPA particle to a continuous quality score. Training combines binary classification and pairwise ranking objectives, encouraging the model to distinguish unambiguous low- and high-quality particles while preserving their relative ordering within an acquisition group. This adaptation tests whether the pretrained global representation supports reconstruction-relevant particle selection. Label construction, pair sampling, optimization, and training schedules are provided in Appendix D.1.

Micrograph-level particle picking. For particle picking, a lightweight dense prediction head operates on the spatial grid of final-layer ViT patch tokens to predict particle centers across a micrograph. Retaining the patch-grid organization tests whether joint pretraining across observation types preserves localized particle evidence in addition to global image-level information. The encoder and prediction head are fine-tuned jointly; the head architecture, training targets, tiling and coordinate-decoding procedures, post-processing, and optimization settings are detailed in Appendix D.2.

Together, the two adaptations test whether one pretrained backbone can support global particle assessment and dense acquisition-scale localization.

4 Experiments

4.1 Pose and structure awareness in SPA particle representations

To assess whether the broader pretraining scope of CryoVERSE preserves fine-grained SPA particle information, we evaluated pose awareness within each molecular structure and structure awareness across molecules. We used the 12 simulated SPA particle datasets from the Cryo-IEF study (Yan et al., 2026a). Similarity-weighted 5-nearest-neighbor (5-NN) classification was applied to normalized backbone features without fitting a task-specific prediction head. We compared CryoVERSE with the generic MAE and MoCo v3 baselines (He et al., 2022; Chen et al., 2021) and the particle-specific Cryo-IEF model. Appendix E.2 provides the complete evaluation protocol.

CryoVERSE achieved the highest accuracy for both projection pose and molecular identity (Figure 2 and Table 2). Relative to Cryo-IEF, pose accuracy increased by 4.53 pp to 95.67%, while structure accuracy increased by 2.00 pp to 98.94%. CryoVERSE outperformed Cryo-IEF on both tasks for every dataset. The three lowest-mass targets (69–195 kDa) showed the largest absolute pose gains, ranging from 9.22 to 13.27 pp. UMAP projections (McInnes et al., 2018) also showed more coherent local pose neighborhoods with CryoVERSE in representative datasets. Generic MAE and MoCo v3 representations performed substantially worse, indicating that natural-image self-supervision did not capture the same particle-level structural organization. Complete dataset-level accuracies and UMAP comparisons across all 12 datasets are reported in Appendix Table 7 and Appendix Figure 6, respectively.

Controlled ablations clarified the contributions of the pretraining data and pose-aware objective. Compared with CryoVERSE (w/o wide-field observations), the full model achieved similar structure accuracy (98.94% vs. 98.70%) and higher pose accuracy (95.67% vs. 94.83%). Thus, cross-scale pretraining did not dilute particle-level information. In contrast, CryoVERSE (w/o particle

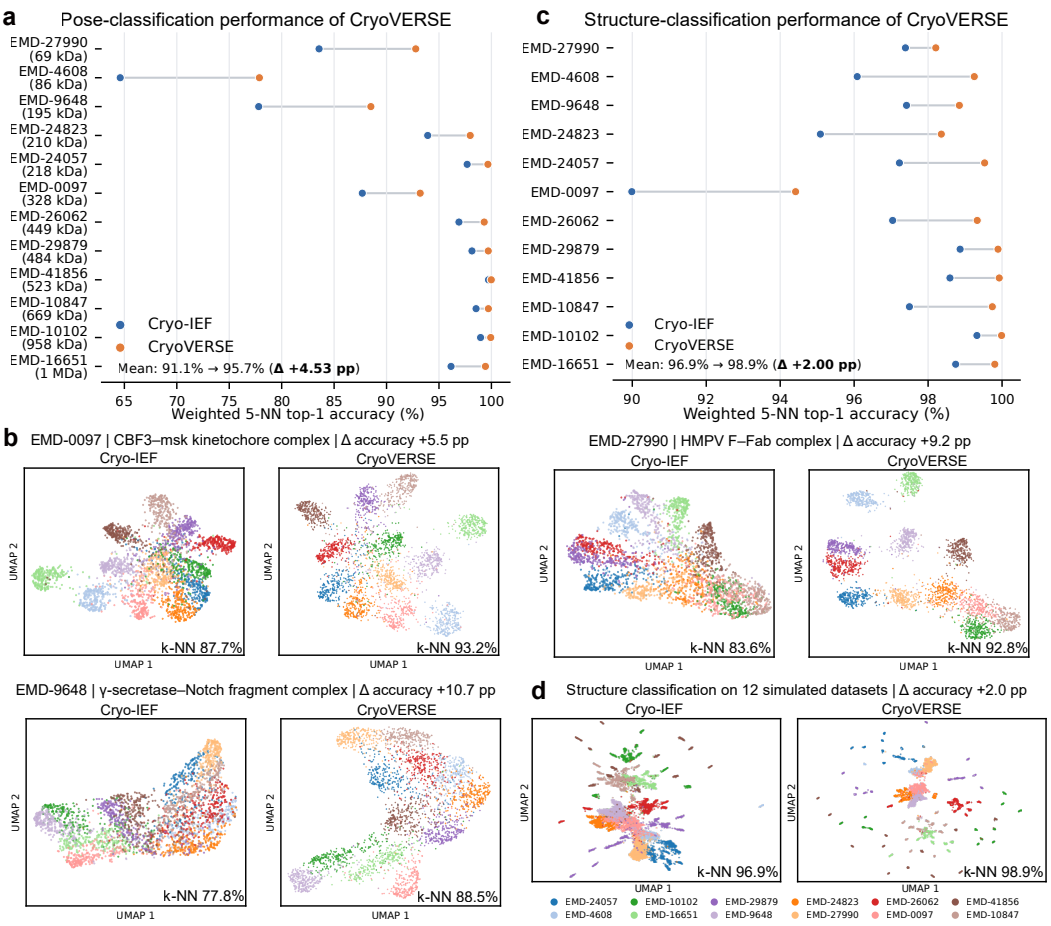

Figure 2: **CryoVERSE improves pose and structure classification of SPA particles over Cryo-IEF. a**, Weighted 5-NN pose classification accuracy across 12 simulated SPA particle datasets. **b**, UMAP projections comparing pose organization under Cryo-IEF and CryoVERSE for three representative datasets, colored by ground-truth pose classes. Precise dataset-level accuracies and the complete evaluation protocol are provided in Appendix Table 7 and Appendix E.2. **c**, Weighted 5-NN structure-classification accuracy across the 12 simulated datasets. Connected points compare Cryo-IEF and CryoVERSE on the same dataset, and the displayed means are unweighted macro-averages. **d**, UMAP projections comparing structure organization under Cryo-IEF and CryoVERSE, colored by ground-truth structure types.

observations) sharply degraded both tasks, confirming that particle observations remain essential for molecular and pose organization. The CryoVERSE (w/o cryo-ET observations) variant reduced pose and structure accuracy by 1.38 and 0.83 pp, respectively, while removing pose-aware positives reduced them by 2.32 and 1.09 pp. The larger reduction in pose accuracy shows that the pose-aware objective primarily strengthens projection-pose organization.

4.2 CryoVERSE improves particle-quality ranking and supports better reconstructions

To determine whether CryoVERSE improves particle quality estimation and supports better reconstructions, we evaluated particle-quality information at three complementary levels. We first examined the quality information encoded by CryoVERSE features, then assessed particle quality prediction after CryoRanker fine-tuning, and finally tested reconstructions using score-based selection of particle subsets. The CryoVERSE feature and CryoRanker evaluations used four held-out SPA datasets with curated continuous quality labels, whereas reconstruction experiments based on

Table 2: **Comparison and ablation of pretraining strategies for SPA pose and structure classification.** Values are unweighted macro-averages of weighted 5-NN top-1 accuracy across 12 simulated SPA particle datasets. Variant definitions are provided in Appendix C.3.

Model	Pose classification (5-NN, %)	Structure classification (5-NN, %)
MAE ViT-B/16	9.51	30.10
MoCo v3 ViT-B/16	14.87	32.65
Cryo-IEF	91.14	96.94
CryoVERSE (w/o particle observations)	11.59	15.15
CryoVERSE (w/o wide-field observations)	94.83	98.70
CryoVERSE (w/o cryo-ET observations)	94.29	98.11
CryoVERSE (w/o pose-aware positives)	93.35	97.85
CryoVERSE	**95.67**	**98.94**

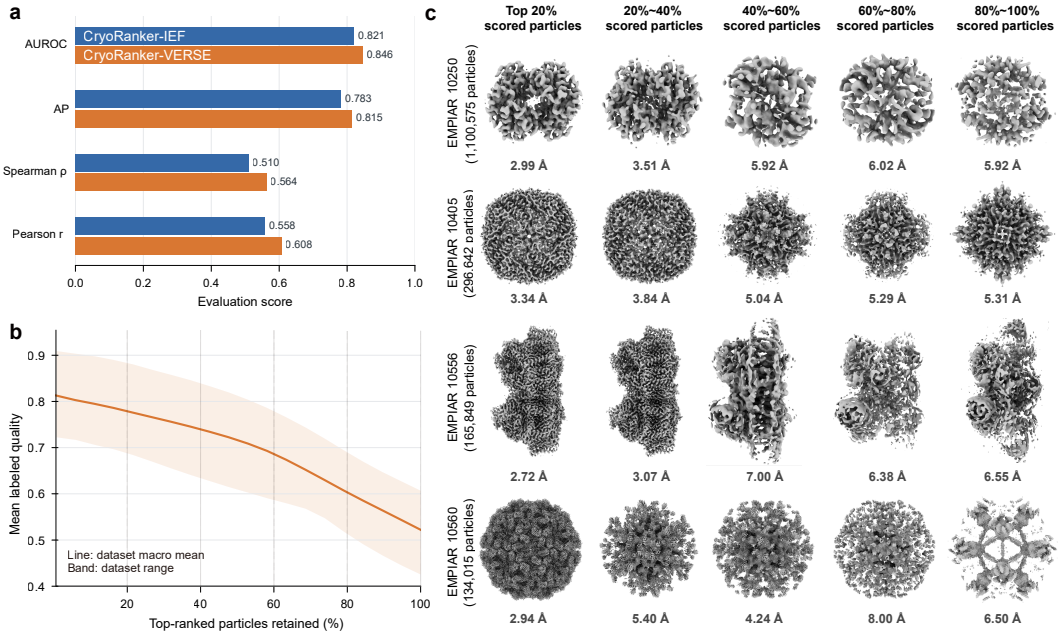

Figure 3: **CryoRanker-VERSE improves particle quality assessment and supports higher-resolution reconstructions over CryoRanker-IEF. a**, Evaluation of particle quality assessment by CryoRanker-IEF and CryoRanker-VERSE. Values are macro-averages over four held-out SPA datasets. **b**, Correlation between labeled data quality and CryoRanker-VERSE assigned quality scores. The line denotes the macro-average of labeled quality scores across the four validation datasets and the shaded band denotes the score ranges. **c**, Evaluation for particle quality assessment by CryoRanker-VERSE via structural reconstruction. Particles of four held-out test datasets were ranked using CryoRanker-VERSE and partitioned into five equal-sized, non-overlapping score quintiles. Reconstructions of each quintile were conducted using cryoSPARC Non-Uniform Refinement. Nominal resolution is reported below each reconstruction. Complete evaluation protocols are provided in Appendix E.3.

quality-guided particle selection used a separate panel of four experimental datasets spanning diverse molecular sizes and sample compositions. Dataset construction and complete protocols are provided in Appendix A.3 and Appendix E.3.

CryoVERSE features preserved particle-quality information and modestly improved all four macro-average readout metrics relative to Cryo-IEF (Appendix Figure 7). These consistent gains show that broader pretraining across SPA and cryo-ET observations did not compromise quality-related infor-

mation. Because this leave-one-out analysis retrieves labels from neighboring evaluation particles, it serves only as a representation diagnostic rather than a standalone quality predictor.

Under the matched fine-tuning protocol, CryoRanker-VERSE improved both binary quality discrimination and continuous particle ranking relative to CryoRanker-IEF (Figure 3a). The largest gains were observed in average precision and Spearman correlation, which increased by 0.032 and 0.054, respectively. The two models shared the fine-tuning data, prediction head, objectives, sampling procedure, and optimization settings. Their performance difference therefore reflects the combined effect of encoder design and pretrained initialization rather than the downstream training protocol.

The predicted scores of CryoRanker-VERSE also produced better reconstructions. Higher-ranked subsets had greater curated quality, and their mean quality decreased progressively as more particles were retained (Figure 3b). In independent reconstructions, the highest-scoring quintile achieved the best nominal resolution in all four datasets (Figure 3c). At the matched top-20% retention level, CryoRanker-VERSE yielded better resolutions than CryoRanker-IEF (Appendix Table 8). These results show that CryoVERSE provides more informative particle-quality representations and supports reconstruction-relevant particle selection after task-specific adaptation.

4.3 Dense learning transfer at the micrograph scale

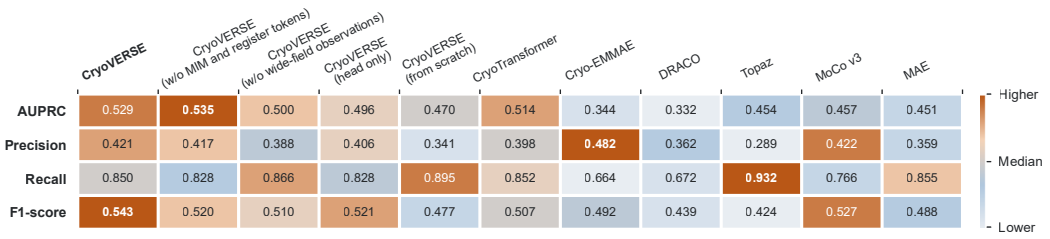

Figure 4: **Particle-picking performance across different models.** Cell values are macro-averages over five held-out CryoPPP test accessions. Colors encode rank within each metric from lower (blue) to higher (orange), and bold values denote the best result for each metric.

At the micrograph scale, we evaluated whether the spatially resolved patch representations learned by CryoVERSE transfer to dense particle localization. CryoVERSE was evaluated on five held-out CryoPPP accessions using a lightweight prediction head over its final patch-token grid (Dhakal et al., 2023). The evaluated baselines included MAE ViT-B/16 (He et al., 2022), MoCo v3 ViT-B/16 (Chen et al., 2021), Topaz (Bepler et al., 2019), DRACO (Shen et al., 2024), Cryo-EMMAE (Zamanos et al., 2025), and CryoTransformer (Dhakal et al., 2024). Full CryoVERSE achieved the highest macro-average F1-score of 0.5434 while retaining the second-highest macro-average AUPRC of 0.5294 (Figure 4). Its F1-score exceeded those of the strongest competing method and the same architecture trained from scratch. Removing wide-field observations from pretraining also reduced F1-score, supporting their contribution to dense particle localization. Representative predictions across diverse micrograph conditions are shown in Appendix Figure 8. Together, these results show that the shared pretrained encoder supports dense localization at the micrograph scale. Complete experimental protocols and implementation details are provided in Appendix Section E.4.

4.4 CryoVERSE preserves structural identity and acquisition-angle information in cryo-ET

Cryo-ET acquires multiple low-dose projections of the same specimen region at known tilt angles. Distributing the electron dose across the tilt series, together with the increased effective specimen thickness at high tilt, produces weak and angle-dependent signals (Wan & Briggs, 2016). We therefore evaluated CryoVERSE representations at two complementary levels: structural identity in particle views and acquisition-angle information in complete tilt images. Both evaluations used similarity-weighted 5-nearest-neighbor (5-NN) analysis without fitting a task-specific prediction head.

Table 3: **Macro-averaged particle-type classification and tilt-angle estimation results across four datasets.** Values are unweighted averages over CZI-10301, CZI-10453, CZI-10468, and CZI-10469. **Bold** and underlined values denote the best and second-best results, respectively. A dash indicates that the method is not applicable to the corresponding task.

	Particle-type classification			Tilt-angle estimation			
Method	Type-macro acc. ↑	Mean particle acc. ↑	View macro-F1 ↑	MAE (°) ↓	Spearman ρ ↑	AUROC ↑	AP ↑
Generic visual models							
MAE	0.6985	0.8123	0.6994	19.4736	0.4113	0.6618	0.5990
MoCo v3	0.6972	0.8051	0.7033	19.0160	0.4580	0.6830	0.6191
Cryo-EM models							
Cryo-IEF	0.6556	0.7502	0.6656	–	–	–	–
DRACO	–	–	–	16.7953	0.5707	0.7697	0.7123
CryoVERSE (w/o cryo-ET observations)	0.7007	0.7742	0.7102	10.0043	0.8440	0.8573	0.8322
CryoVERSE (w/o pose-aware positives)	0.7396	0.7940	0.7476	9.6557	0.8519	0.8677	0.8421
CryoVERSE	**0.7894**	**0.8433**	**0.7984**	**8.5237**	**0.8800**	**0.8758**	**0.8471**

CryoVERSE improved particle-type classification across all four cryo-ET datasets. For each query view, we excluded every alternative view originating from the same physical particle before neighbor retrieval, preventing repeated observations of one object from creating trivial correspondences. Under this constraint, particle-level type-macro accuracy increased from 65.56% with Cryo-IEF to 78.94% with CryoVERSE (Appendix Figure 9a), showing that CryoVERSE better organizes particle views by structural identity across distinct physical particles. Complete dataset-level and particle-type-specific results are reported in Appendix Tables 10 and 11, respectively.

Frozen CryoVERSE tilt-image representations retained substantial acquisition-angle information. Similarity-weighted 5-NN regression reduced the macro-average MAE from 16.79° with DRACO to 8.52° with CryoVERSE. This corresponds to an 8.27° reduction in angular error. CryoVERSE achieved lower MAE on all four datasets (Appendix Figure 9b and Appendix Table 9). The largest reduction occurred on CZI-10301, where MAE decreased from 16.317° to 6.278°.

Together with the particle-view results, these findings show that CryoVERSE preserves structural identity at the local particle level while retaining physically meaningful projection geometry at the global tilt-image level.

5 Conclusion

We introduced CryoVERSE, a unified representation model spanning four cryo-EM pre-reconstruction 2D observation types across SPA and cryo-ET. Its defining feature is a scale-flexible shared encoder that combines pose-aware global learning with patch-level representation learning. This design allows one backbone to process both particle-scale and wide-field observations without observation-specific branches.

Across the evaluated tasks, the global representations retained information about particle structure, pose, and quality, while the spatial representations transferred to dense particle localization. The same encoder also organized cryo-ET particle views by structural identity and retained acquisition-angle information in complete tilt images. After downstream adaptation, the representations further supported particle quality ranking and better reconstructions. Together, these results show that a shared pretrained backbone can provide reusable features across complementary stages of pre-reconstruction cryo-EM analysis.

The current evidence is limited to individual 2D observations. Future work could aggregate related observations within an acquisition and extend the framework to complete tilt series, reconstructed tomograms, and subtomogram-level analysis. Such extensions would move beyond reusable 2D features toward coordinated representations across the cryo-EM data processing hierarchy.

AI use statement

Generative AI tools were used to edit the manuscript for clarity, readability, and grammar, and to assist with debugging research code by explaining error messages and suggesting local code corrections. The authors reviewed all AI-assisted edits and independently tested the modified code

for correctness. Generative AI tools were not used to generate synthetic data, develop theoretical or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design the research methodology or experiments, clean or reformat datasets, conduct qualitative or thematic analysis, or interpret the results. The authors take full responsibility for the final content, code, and results of this work.

Reproducibility statement

We support reproducibility by providing a complete account of dataset provenance, preprocessing, dataset-level splits, and leakage controls in Appendix A.1–B; the pretraining objectives, batch construction, and optimization settings in Appendix C; and the downstream adaptation and frozen-feature evaluation protocols in Appendix D.1–E.5. The 12 simulated SPA particle datasets and four experimental SPA particle datasets reused from Cryo-IEF are publicly available at `https://zenodo.org/records/17066236` and `https://zenodo.org/records/17066297`, respectively; the remaining public evaluation datasets are identified by repository accession in Appendix A.3. Source code and pretrained model weights will be released at `https://anonymous.4open.science/r/CryoVERSE-446C`.

References

William T. Baxter, Robert A. Grassucci, Haixiao Gao, and Joachim Frank. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *Journal of Structural Biology*, 166(2):126–132, 2009. doi: 10.1016/j.jsb.2009.02.012.

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*, 16:1153–1160, 2019. doi: 10.1038/s41592-019-0575-8.

Qihe Chen et al. A large-scale curated and filterable dataset for cryo-EM foundation model pretraining. *Scientific Data*, 12:960, 2025. doi: 10.1038/s41597-025-05179-2.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Yifan Cheng. Single-particle cryo-EM—how did it get here and where will it go. *Science*, 361 (6405):876–880, 2018. doi: 10.1126/science.aat4346.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2dnO3LLiJ1`.

Ashwin Dhakal, Rajan Gyawali, Liguo Wang, Jianlin Cheng, et al. A large expert-curated cryo-EM image dataset for machine learning protein particle picking. *Scientific Data*, 10:392, 2023. doi: 10.1038/s41597-023-02280-2.

Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. CryoTransformer: a transformer model for picking protein particles from cryo-EM micrographs. *Bioinformatics*, 40(3):btae109, 2024. doi: 10.1093/bioinformatics/btae109.

Utz Ermel, Anchi Cheng, Jun Xi Ni, et al. A data portal for providing standardized annotations for cryo-electron tomography. *Nature Methods*, 21:2200–2202, 2024. doi: 10.1038/s41592-024-02477-2.

Weining Fu, Kai Shu, Kui Xu, and Qiangfeng Cliff Zhang. CryoLVM: Self-supervised learning from cryo-EM density maps with large vision models. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=9xcvEF2BRi`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16000–16009. IEEE, 2022.

Dari Kimanius, Liyi Dong, Grigory Sharov, Takanori Nakane, and Sjors H. W. Scheres. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal*, 478(24): 4169–4185, 2021. doi: 10.1042/BCJ20210708.

James R. Kremer, David N. Mastronarde, and J. Richard McIntosh. Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology*, 116(1):71–76, 1996. doi: 10.1006/jsbi.1996.0013.

Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, Ge Yang, et al. DeepETPicker: fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15:2090, 2024. doi: 10.1038/s41467-024-46041-0.

David N. Mastronarde. Automated electron microscope tomography using robust prediction of specimen movements. *Journal of Structural Biology*, 152(1):36–51, 2005. doi: 10.1016/j.jsb.2005.07.007.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Eva Nogales. The development of cryo-EM into a mainstream structural biology technique. *Nature Methods*, 13(1):24–27, 2016. doi: 10.1038/nmeth.3694.

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 Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

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. doi: 10.1038/nmeth.4169.

Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniela Prumbaum, and Stefan Raunser. TomoTwin: generalized 3d localization of macromolecules in cryo-electron tomograms with structural data mining. *Nature Methods*, 20:871–880, 2023. doi: 10.1038/s41592-023-01878-z.

Alexis Rohou and Nikolaus Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221, 2015. doi: 10.1016/j.jsb.2015.08.008.

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

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/2410.11373. arXiv:2410.11373 [cs.CV].

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Dimitry Tegunov and Patrick Cramer. Real-time cryo-electron microscopy data preprocessing with Warp. *Nature Methods*, 16:1146–1152, 2019. doi: 10.1038/s41592-019-0580-y.

William Wan and John A. G. Briggs. Cryo-electron tomography and subtomogram averaging. In *Methods in Enzymology*, volume 579, pp. 329–367. Elsevier, 2016. doi: 10.1016/bs.mie.2016.04.014.

wwPDB Consortium. EMDB—the electron microscopy data bank. *Nucleic Acids Research*, 52 (D1):D456–D465, 2024. doi: 10.1093/nar/gkad1019.

Yang Yan, Shiqi Fan, Fajie Yuan, and Huaizong Shen. A comprehensive foundation model for cryo-EM image processing. *Nature Methods*, 23:88–95, 2026a. doi: 10.1038/s41592-025-02916-8.

Yang Yan, Yanwanyu Xi, Shiqi Fan, Yifei Wang, Ziyun Tang, Fajie Yuan, and Huaizong Shen. CryoDECO: Deconstructing extreme compositional and conformational heterogeneity in cryo-EM via foundation model priors. *LangTaoSha Preprint Server*, 2026b. doi: 10.65215/LTSpreprints.2025.12.30.000075. Preprint.

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. doi: 10.1016/j.crmeth.2025.101089.

Andreas Zamanos, Fotis L. Kyrilis, Panagiotis Koromilas, Panagiotis L. Kastritis, and Yannis Panagakis. A 37-million-particle dataset from over 250 experiments to accelerate data-driven cryo-EM analysis. *bioRxiv*, 2026. doi: 10.64898/2026.04.29.720997.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. doi: 10.48550/arXiv.1904.07850. URL https://arxiv.org/abs/1904.07850.

Yi Zhou, Yilai Li, Jing Yuan, and Quanquan Gu. CryoFM: A flow-based foundation model for cryo-EM densities. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=T4sMzjy7fO.

A Dataset inventory, splits, and leakage control

A.1 Pretraining corpus

A.1.1 Micrographs

The micrograph corpus comprises data from every EMPIAR accession listed in the Cryo-IEF and CryoDECO datasets, the in-house datasets used by Cryo-IEF, and the CryoCRAB dataset (Yan et al., 2026a;b; Chen et al., 2025). For each eligible EMPIAR accession, we sampled up to 600 micrographs before quality control and preprocessing. When an accession provided dose-fractionated movies rather than motion-corrected micrographs, we generated micrographs using Patch Motion Correction in cryoSPARC.

A.1.2 SPA particles

SPA particle images were obtained either by processing the movies or micrographs described above or directly from datasets that released extracted particles. Following the cryoSPARC-based workflow of Cryo-IEF, dose-fractionated movies were first processed by Patch Motion Correction, whereas this step was omitted for motion-corrected micrographs (Yan et al., 2026a). Contrast transfer function parameters were estimated using CTFFIND4 (Rohou & Grigorieff, 2015); candidate particles were identified using Blob Picker, and particle images were extracted using the corresponding cryoSPARC tools (Punjani et al., 2017). Datasets that already provided extracted particle images bypassed these preprocessing steps. We further supplemented the corpus with particle images from cryoPANDA, retaining only particles whose EMPIAR accession was not already represented in our micrograph-derived particle corpus (Zamanos et al., 2026).

For each dataset, particles were curated using either CryoRanker scores or class averages generated by cryoSPARC 2D Classification. Selected particles were subjected to cryoSPARC Ab-Initio Reconstruction followed by Non-Uniform Refinement. When the resulting map reached a nominal resolution better than 4.2 Å, reconstructed particle poses were used to select pretraining positives. Otherwise, two independently augmented views of the same particle image formed a positive pair. We additionally retained junk particles corresponding to 10% of each dataset; these particles

were excluded from reconstruction and used only with image-level augmented positives. A particle dataset was discarded when 2D Classification did not yield interpretable class averages or when 3D reconstruction did not produce a reliable map.

A.1.3 CRYO-ET TILT IMAGES AND PARTICLE VIEWS

The cryo-ET corpus was obtained from the CryoET Data Portal (Ermel et al., 2024) and contains 52,537 successfully processed tilt images and 15,571,232 tilt-derived particle views. The provenance-resolution and particle-view construction procedure is described in Appendix B.

A.2 DOWNSTREAM TRAINING DATASETS

A.2.1 CRYORANKER FINE-TUNING

We reused the CryoRanker fine-tuning corpus introduced with Cryo-IEF, retaining its particle composition and continuous quality labels (Yan et al., 2026a). The corpus contains 42,388,319 SPA particle images from EMPIAR, CryoPPP, and 50 in-house datasets. Movies and micrographs were processed into particles using the workflow described in Appendix A.1, whereas datasets that already provided extracted particles bypassed this processing.

To obtain particle-quality labels, particles were grouped using cryoSPARC 2D Classification, and the resulting class averages were assigned continuous quality scores using the RELION Class Ranker (Kimanius et al., 2021). The automatically assigned scores were manually inspected and corrected when evident scoring errors were identified, and every particle inherited the score of its corresponding 2D class. Scores ranged from 0 for noise or ice to 1 for clearly resolved structural features. Datasets whose 2D class averages did not support reliable quality assignment were excluded, and the retained corpus was balanced to limit strong overrepresentation of either high- or low-quality particles. The dataset-separation procedure described above was applied before the final fine-tuning corpus was constructed. Model initialization, objectives, optimization, and matched CryoRanker-IEF and CryoRanker-VERSE training settings are reported in Appendix D.1.

A.2.2 PARTICLE-PICKING FINE-TUNING

We used the raw micrographs and released particle-center annotations from CryoPPP for supervised particle-picking adaptation. The five validation and five test accessions specified in Appendix A.3 were excluded from fine-tuning, and the remaining 24 accession-level datasets formed the training split. Following the dataset organization provided by CryoPPP, EMPIAR-10389 is represented as two predefined subsets, EMPIAR-10389A and EMPIAR-10389B. Consequently, the 24 training accessions correspond to 25 loader-defined training subsets. All released particle coordinates were retained without confidence-based annotation filtering.

CryoPPP was used only for downstream supervised adaptation and was not a source dataset for CryoVERSE encoder pretraining. The pretrained encoder was combined with the dense prediction head described in Section 3.3 and fine-tuned end to end. Sampling, input preprocessing, optimization, and compute settings are provided in Appendix D.2.

A.3 EVALUATION DATASETS AND SPLITS

A.3.1 SPA PARTICLE EVALUATION

We used the same two SPA particle evaluation panels as in the original Cryo-IEF study (Yan et al., 2026a). To evaluate structural awareness, we used 12 labeled simulated datasets generated from high-resolution single-particle cryo-EM density maps downloaded from the EMDB (wwPDB Consortium, 2024). The source structures span molecular masses of approximately 69 kDa to 1 MDa. For each density map, simulated particle images were generated in cryoSPARC with known structure identity and projection pose (Punjani et al., 2017). These labels allow us to evaluate whether the learned representations distinguish molecular structures and organize particles according to viewing direction.

To evaluate particle-quality awareness and ranking, we used the four experimental SPA datasets from the original CryoRanker evaluation. Particles in each dataset were grouped by cryoSPARC

2D Classification, and the resulting class averages were assigned continuous quality scores ranging from 0 to 1, where 0 represents noise or ice and 1 represents clearly resolved structural features. Scores were initially obtained using the RELION Class Ranker (Kimanius et al., 2021) and were manually reviewed for evident scoring errors. Each particle inherited the score of its corresponding class. These datasets were used both to assess the readability of particle quality in frozen feature space and to evaluate the continuous quality predictions and particle rankings produced by fine-tuned CryoRanker models. None of the 12 simulated or four experimental evaluation datasets, or datasets derived from the same or highly similar specimens, was included in model pretraining or CryoRanker fine-tuning.

A.3.2 Reconstruction evaluation

The four reconstruction datasets were deliberately selected to cover a broad range of molecular sizes and sample compositions rather than sampled at random. EMPIAR-10250 contains a 64-kDa target (1,100,575 particles), EMPIAR-10405 a 440-kDa target (296,642 particles), EMPIAR-10556 a 725-kDa target (165,849 particles), and EMPIAR-10560 a nucleic-acid-containing assembly of approximately 16 MDa (134,015 particles). This panel ranges from a small protein target to a very large nucleoprotein assembly, providing a test of score-based particle selection across substantially different reconstruction regimes. EMPIAR-10250, EMPIAR-10405, and EMPIAR-10556 were also evaluated in the published CryoRanker-IEF reconstruction experiment (Yan et al., 2026a). Because EMPIAR-10560 was not included in that experiment, we reproduced CryoRanker-IEF on this dataset under the same protocol for the matched comparison reported here.

A.3.3 Micrograph particle-picking evaluation

We evaluated particle picking on the CryoPPP benchmark (Dhakal et al., 2023), with datasets split at the EMPIAR-accession level. EMPIAR-10184, EMPIAR-10671, EMPIAR-10737, EMPIAR-10077, and EMPIAR-10947 formed the validation set. For CryoVERSE, these validation accessions were used to select training hyperparameters, the checkpoint epoch, and the operating threshold for threshold-dependent metrics. After these choices were fixed, final performance was evaluated on five held-out test accessions: EMPIAR-10028, EMPIAR-10532, EMPIAR-10093, EMPIAR-10345, and EMPIAR-11056. The operating threshold was selected on the validation accessions and then fixed for test evaluation as described in Appendix E.4. CryoPPP was not used during CryoVERSE encoder pretraining, making particle picking a downstream transfer evaluation on separately curated micrographs.

A.3.4 Cryo-ET evaluation

We evaluated cryo-ET transfer using four publicly available CZI datasets from the CryoET Data Portal (Ermel et al., 2024): CZI-10301, CZI-10453, CZI-10468, and CZI-10469. Datasets were retained only when particle annotations could be linked unambiguously to the corresponding tomograms, tilt series, and alignment metadata, allowing annotated three-dimensional particle locations to be projected consistently into the acquired tilt images. The complete provenance-resolution and particle-construction procedure is described in Appendix B.

The resulting evaluation data comprise two complementary representation levels. At the *particle-view level*, annotation-linked image crops were extracted for physical particles across the available tilt images. Each particle view retains its physical-particle identity, particle-type annotation, source tomogram, tilt index, and acquisition angle. Multiple particle views may therefore correspond to the same physical particle observed under different projection geometries. At the *tilt-image level*, individual tilt images were retained together with their recorded acquisition angles, providing image-level samples for evaluating whether frozen representations preserve acquisition geometry.

Particle-type annotations are used in the particle-view protocol described in Appendix E.5.3, while acquisition angles are used in the tilt-image protocol described in Appendix E.5.2. Physical-particle sampling is defined only in Appendix E.5.1.

A.4 Dataset separation and leakage control

For all training stages, including foundation-model pretraining and downstream fine-tuning, data overlapping the corresponding validation or test sets were excluded at the source-dataset level. For public SPA data, separation was enforced at the EMPIAR-accession level; analogous source identifiers were used for simulated, in-house, and cryo-ET datasets. The 12 simulated and four experimental SPA evaluation datasets, together with datasets derived from the same or highly similar specimens, were excluded from both pretraining and CryoRanker fine-tuning. CryoPPP micrographs were not used for CryoVERSE encoder pretraining and were used only for downstream particle-picking adaptation. The five CryoPPP validation and five test accessions were excluded from particle-picking fine-tuning. The four cryo-ET evaluation datasets were excluded from both encoder pretraining and supervised fine-tuning and were used only for frozen-feature evaluation.

B Cryo-ET data processing and tilt-derived particle construction

Provenance resolution and dataset validation. Each particle annotation was traced through the CryoET Data Portal provenance graph to its associated tomogram, acquisition run, tilt series, and alignment records. We retained only annotations for which the annotation and alignment records referred to the same run and identified a consistent tomogram. This validation prevented particle coordinates from being linked to an unrelated reconstruction or tilt-series acquisition.

Tilt-series metadata synchronization. For each validated run, we collected the downloaded tilt-series stack, acquisition-angle list, SerialEM MDOC metadata (Mastronarde, 2005), and the corresponding IMOD alignment transforms (Kremer et al., 1996). Because these records may contain different projection subsets or ordering conventions, we matched angles in the `.rawtlt` file to the corresponding `TiltAngle` entries in the MDOC file within a predefined angular tolerance. Unmatched MDOC entries were removed, and the retained MDOC records were reordered to follow the tilt-stack order. These checks established an index-wise correspondence among each tilt image, acquisition-angle entry, alignment transform, and MDOC record.

Warp initialization and tilt-geometry import. The motion-corrected per-tilt frame sums provided by the CryoET Data Portal were imported into Warp as precomputed tilt sums (Tegunov & Cramer, 2019). Pixel spacing was obtained from the MDOC metadata and MRC headers and verified for consistency, while per-tilt exposure was derived from the acquisition metadata. Because the downloaded images were already motion corrected, no additional movie-level motion correction was performed. The validated tilt geometry and alignment parameters were then imported into Warp for particle projection and extraction.

Coordinate transformation and tilt-derived particle extraction. Particle coordinates were obtained from CZI point annotations associated with each validated tomogram. Annotation coordinates were interpreted in the voxel coordinate system of the corresponding tomogram and mapped into the coordinate convention used for Warp-based projection.

Coordinates outside the annotation-linked tomogram were discarded. The coordinate scale was defined using the tomogram voxel spacing, and the unbinned reconstruction dimensions were determined from the ratio of the tomogram voxel spacing to the tilt-series pixel spacing.

The mapped coordinates were projected into the corresponding tilt images using the imported geometry and alignment parameters. Warp then generated tilt-specific particle crops with a predefined output pixel size and a box size selected to cover the physical particle diameter. The observations associated with physical particle i are

$$\mathcal{P}_i = \{x_i^{(\ell)}\}_{\ell \in \mathcal{T}_i}, \tag{2}$$

where ℓ indexes the retained tilt images and $x_i^{(\ell)}$ denotes the two-dimensional observation of particle i at acquisition angle ϑ_ℓ. Thus, all observations in $\mathcal{P}_i$ correspond to the same annotated physical particle under different projection geometries.

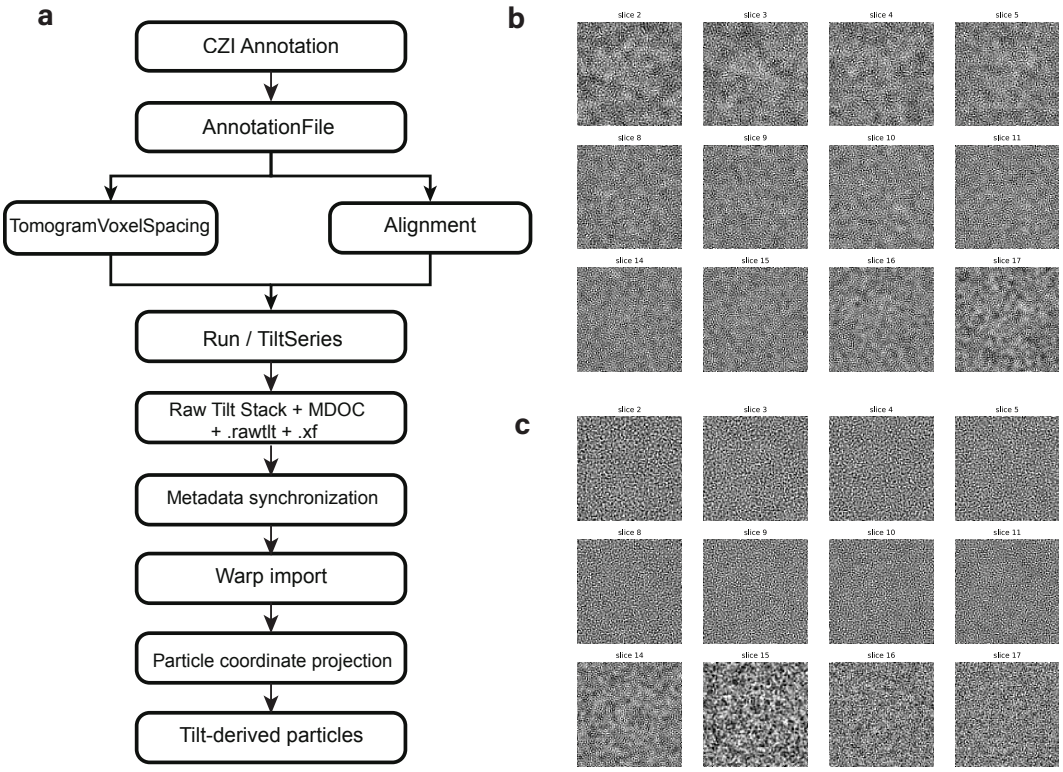

Figure 5: **Construction of cryo-ET particle images. a**, Cryo-ET data processing workflow. Annotation, tilt-series, alignment, and tomogram records are resolved before three-dimensional particle coordinates are projected into aligned tilt images. **b,c**, An illustrative construction example showing multiple views of cytosolic ribosomes (**b**) and PP7 capsids (**c**) from CZI-10440, run TS_5_4. Each row contains tilt views of the same annotated physical particle.

C Pretraining input processing, objectives, and optimization

C.1 Architecture, input processing, and batch construction

The shared backbone is a single-channel ViT-B/14 with rotary position embeddings and register tokens. The online branch uses a three-layer projector and a two-layer predictor, while the momentum branch omits the predictor. View-generation probabilities are reported in Table 4; resized crops use input-category-specific settings. Gradient checkpointing is enabled.

Pretraining uses dataset-capped, source-homogeneous, hierarchical batch sampling. The first level selects the particle or wide-field regime, and the second selects SPA or cryo-ET within that regime. Each batch contains one input category from one data source. Sampling and batch order are refreshed each epoch. The probabilities in Table 4 refer to numbers of batches; wide-field batches contain fewer images because of their longer token sequences.

C.2 Pretraining objectives and optimization

Objective overview. As defined in Equation 1, pretraining combines a weighted momentum-contrastive objective, masked patch-level prediction, and KoLeo regularization. All active terms have unit coefficients. The contrastive and KoLeo terms are applied to all four input categories, whereas masked patch-level prediction is applied only to wide-field-image batches, comprising SPA micrographs and cryo-ET tilt images.

For training example i, let a_i denote the anchor image and x_i^+ its selected positive image. For an ordinary instance-level pair, $x_i^+ = a_i$; for a pose-aware pair, x_i^+ is a distinct particle image selected

using the procedure below. Independently sampled augmentations produce

$$x_i^{(1)} = \mathcal{A}_1(a_i), \qquad x_i^{(2)} = \mathcal{A}_2(x_i^+). \tag{3}$$

For view $v \in \{1, 2\}$, the normalized online and momentum representations are

$$\begin{aligned} u_i^{(v)} &= \mathrm{norm}\Big(q_\theta\Big(g_\theta\Big(f_\theta\Big(x_i^{(v)}\Big)\Big)\Big)\Big), \\ t_i^{(v)} &= \mathrm{norm}\Big(g_{\bar{\theta}}\Big(f_{\bar{\theta}}\Big(x_i^{(v)}\Big)\Big)\Big). \end{aligned} \tag{4}$$

Gradients propagate only through the online branch. The momentum parameters are updated by an exponential moving average of the online parameters following MoCo v3 (Chen et al., 2021).

Pose-aware positive construction. For each particle with an available pose estimate, we convert its rotation vector to a rotation matrix $R_i \in \mathrm{SO}(3)$. We therefore express the microscope beam axis $\mathbf{e}_z = (0, 0, 1)^\mathsf{T}$ in particle coordinates and normalize it to obtain the viewing-direction vector

$$\mathbf{n}_i = \frac{R_i^\mathsf{T}\mathbf{e}_z}{\left\|R_i^\mathsf{T}\mathbf{e}_z\right\|_2} \in \mathbb{S}^2. \tag{5}$$

This representation discards in-plane rotation without introducing an Euler-angle parameterization or an azimuthal discontinuity. For particles i and j, we define the viewing-direction distance as the great-circle angle

$$d_{ij} = \arccos\big[\mathrm{clip}\big(\mathbf{n}_i^\mathsf{T}\mathbf{n}_j, -1, 1\big)\big] \in [0, \pi]. \tag{6}$$

For a structure with known rotational symmetry group $G \subset \mathrm{SO}(3)$, symmetry-equivalent viewing directions are identified by instead using

$$d_{ij}^{\mathrm{sym}} = \min_{g \in G} \arccos\big[\mathrm{clip}\big(\mathbf{n}_i^\mathsf{T} g\mathbf{n}_j, -1, 1\big)\big]. \tag{7}$$

We write $\delta_{ij} = d_{ij}^{\mathrm{sym}}$ when a rotational symmetry group is specified for the structure and $\delta_{ij} = d_{ij}$ otherwise.

For a particle with pose metadata, candidate neighbors must lie within the angular threshold

$$\theta_{\max} = 20^\circ = \frac{\pi}{9}\ \mathrm{rad}. \tag{8}$$

For an SPA anchor i, let

$$\mathcal{C}_i^{\mathrm{SPA}} = \{j \neq i : \mathrm{dataset}(j) = \mathrm{dataset}(i),\ \delta_{ij} \leq \theta_{\max}\}. \tag{9}$$

If more than 100 particles are eligible, 100 candidates are randomly selected and retained. For a tilt-derived cryo-ET anchor, let

$$\mathcal{C}_i^{\mathrm{ET}} = \{j \neq i : \mathrm{stack}(j) = \mathrm{stack}(i),\ \delta_{ij} \leq \theta_{\max}\}. \tag{10}$$

If more than three particles are eligible, three candidates are randomly selected and retained. We sample the positive uniformly from the corresponding retained set. Let $j(i)$ denote the sampled neighbor and let $d_i = \delta_{i,j(i)}$. If pose-aware sampling is not selected or the retained eligible set is empty, we use the anchor itself as the positive, $x_i^+ = a_i$.

Viewing-direction-dependent pair confidence. Same-image pairs are assigned full confidence, $\omega_i = 1$. Because two distinct particles can differ in conformation, image quality, or pose-estimation accuracy even when their estimated viewing directions are close, pose-neighbor pairs receive a lower, distance-dependent weight:

$$\omega_i = \lambda_{\mathrm{cross}} \exp\left(-\frac{d_i^2}{2\sigma_{\mathrm{pose}}^2}\right), \qquad d_i \leq \theta_{\max}, \tag{11}$$

where $\lambda_{\mathrm{cross}} = 0.95$ is the maximum confidence assigned to a cross-particle pair and $\sigma_{\mathrm{pose}} = 10^\circ = \pi/18$ rad is the angular decay scale. At $d_i = \sigma_{\mathrm{pose}}$, the confidence is $\lambda_{\mathrm{cross}} \exp(-1/2) \approx 0.576$. Over the eligible interval, the weight decreases smoothly from 0.95 at $d_i = 0^\circ$ to approximately 0.129 at $d_i = 20^\circ$. Thus, a cross-particle pair with a nearly identical viewing direction receives close to full confidence while remaining slightly less reliable than a same-image pair; confidence then decreases monotonically with viewing-direction distance.

Confidence-weighted momentum contrast. For the direction from online view 1 to momentum view 2, the per-example contrastive cross-entropy is

$$\ell_{\mathrm{con},1\to2,i} = -\log \frac{\exp\Big((u_i^{(1)})^{\mathsf{T}} t_i^{(2)}/\tau\Big)}{\sum_{j=1}^{B_{\mathrm{all}}} \exp\Big((u_i^{(1)})^{\mathsf{T}} t_j^{(2)}/\tau\Big)}, \tag{12}$$

where $\tau = 0.08$ and B_{all} includes momentum targets gathered across distributed processes. The implementation computes the normalized weighted mean

$$\mathcal{L}_{1\to2} = 2\tau \frac{\sum_{i=1}^{B} \omega_i \ell_{\mathrm{con},1\to2,i}}{\sum_{i=1}^{B} \omega_i + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{13}$$

The factor 2τ is retained from the implemented loss scaling. The reverse-direction term is obtained by exchanging the two views:

$$\ell_{\mathrm{con},2\to1,i} = -\log \frac{\exp\Big((u_i^{(2)})^{\mathsf{T}} t_i^{(1)}/\tau\Big)}{\sum_{j=1}^{B_{\mathrm{all}}} \exp\Big((u_i^{(2)})^{\mathsf{T}} t_j^{(1)}/\tau\Big)}, \tag{14}$$

$$\mathcal{L}_{2\to1} = 2\tau \frac{\sum_{i=1}^{B} \omega_i \ell_{\mathrm{con},2\to1,i}}{\sum_{i=1}^{B} \omega_i + \epsilon}. \tag{15}$$

The complete global contrastive objective is

$$\mathcal{L}_{\mathrm{MoCo}} = \mathcal{L}_{1\to2} + \mathcal{L}_{2\to1}. \tag{16}$$

Normalizing by $\sum_i \omega_i$ changes the relative contribution of uncertain pose-neighbor pairs without causing the overall contrastive-loss scale to decrease when a batch contains many low-confidence pairs. The 20° angular threshold enforces viewing-direction locality, the candidate caps bound search and sampling density, uniform sampling introduces variation within the eligible neighborhood, and confidence weighting limits the influence of more distant selected neighbors.

Masked patch-level prediction. Inspired by the patch-level self-distillation principle of iBOT (Zhou et al., 2022), we apply masked patch-level prediction only to SPA micrographs and cryo-ET tilt images. For each augmented view, the online encoder receives a patch-masked input, whereas the momentum encoder receives the same augmented view without masking. The momentum branch consequently provides stop-gradient targets computed from the complete content of that view. The objective predicts momentum-encoder patch features rather than image pixels.

After excluding the CLS and register tokens, let $h_{\theta,i,p}^{(v)}$ and $h_{\bar\theta,i,p}^{(v)}$ denote the online and momentum patch features at spatial position p in view $v \in \{1,2\}$. For the mask $\mathcal{M}_i^{(v)}$ applied to view v, these features are

$$\begin{aligned} h_{\theta,i,p}^{(v)} &= f_\theta\Big(\mathrm{Mask}\Big(x_i^{(v)}, \mathcal{M}_i^{(v)}\Big)\Big)_p, \\ h_{\bar\theta,i,p}^{(v)} &= f_{\bar\theta}\Big(x_i^{(v)}\Big)_p. \end{aligned} \tag{17}$$

Let r_θ denote the patch-feature predictor. For each masked position $p \in \mathcal{M}_i^{(v)}$, we compute

$$\ell_{i,p}^{(v)} = 1 - \cos\Big(r_\theta\Big(h_{\theta,i,p}^{(v)}\Big), \mathrm{sg}\Big[h_{\bar\theta,i,p}^{(v)}\Big]\Big), \tag{18}$$

where sg denotes stop-gradient. The loss is averaged first over masked positions, then over the two views, and finally over eligible samples:

$$\mathcal{L}_{\mathrm{MIM}} = \frac{1}{|\mathcal{B}_M|} \sum_{i\in\mathcal{B}_M} \frac{1}{2} \sum_{v=1}^{2} \frac{1}{|\mathcal{M}_i^{(v)}|} \sum_{p\in\mathcal{M}_i^{(v)}} \ell_{i,p}^{(v)}. \tag{19}$$

Here, $\mathcal{B}_M$ contains wide-field samples with at least one masked patch. This normalization prevents examples with larger masks from receiving greater weight solely because they contain more masked positions. Block masks cover a uniformly sampled fraction between 0.3 and 0.5 of the spatial patch grid. For particle-image batches, we set $\mathcal{L}_{\mathrm{MIM}} = 0$.

Table 4: **Pretraining hyperparameters.**

Setting	Value
Backbone / input channels	ViT-B/14 / 1
Training schedule	60 epochs; 10 warm-up epochs
Global batch size	2560 for particle images; 320 for micrographs and tilt images
Batch-type sampling	65% particle / 35% wide-field; 4:1 SPA-to-cryo-ET ratio within each regime
Per-dataset epoch cap	20,000 images
View generation	rotation (0.8); resized crop (0.45); center crop (0.25); circular mask (0.2)
Optimizer / initial learning rate	AdamW / 2×10^{-5}
Contrastive temperature	0.08
Pose-neighbor angular threshold	20°
Momentum schedule	cosine from 0.99 to 1
Number of register tokens	4
MIM inputs / masking	micrographs and tilt images only / block masks with ratios of 0.3–0.5
Compute / wall-clock time	8 NVIDIA A100 GPUs (80 GB each) / approximately 19 days

KoLeo regularization. Following DINOv2 (Oquab et al., 2024), we apply KoLeo regularization to the normalized online representations $u_i^{(v)}$. For each view, the nearest non-identical representation within the batch is

$$n_v(i) = \underset{j \in \{1,\dots,B\} \setminus \{i\}}{\arg\min} \left\| u_i^{(v)} - u_j^{(v)} \right\|_2 . \tag{20}$$

The KoLeo objective is

$$\mathcal{L}_{\mathrm{KoLeo}} = -\frac{1}{2B} \sum_{v=1}^{2} \sum_{i=1}^{B} \log\left(\left\| u_i^{(v)} - u_{n_v(i)}^{(v)} \right\|_2 + \epsilon \right) . \tag{21}$$

The implementation uses the single nearest non-identical representation defined above. This objective discourages crowding among normalized online representations. Together, $\mathcal{L}_{\mathrm{MoCo}}$ organizes global representations using image-level and pose-aware positive relations, $\mathcal{L}_{\mathrm{MIM}}$ preserves spatially localized information in wide-field observations, and $\mathcal{L}_{\mathrm{KoLeo}}$ promotes a dispersed representation distribution.

C.3 Controlled pretraining variants

CryoVERSE (w/o particle observations) excludes SPA particles and tilt-derived cryo-ET particles from pretraining. CryoVERSE (w/o wide-field observations) excludes SPA micrographs and cryo-ET tilt images. CryoVERSE (w/o cryo-ET observations) excludes cryo-ET tilt images and tilt-derived cryo-ET particles. CryoVERSE (w/o pose-aware positives) disables pose-aware cross-particle positive selection while retaining the same pretraining data and remaining objectives.

D Downstream adaptation

D.1 CryoRanker adaptation

CryoRanker augments the pretrained encoder with a scalar prediction head. The pretraining head is replaced by a single layer predictor, and the sigmoid-transformed logit is used as the continuous quality score. Models initialized from the Cryo-IEF and CryoVERSE encoders are denoted CryoRanker-IEF and CryoRanker-VERSE, respectively. They use matched downstream settings and differ only in encoder architecture and pretrained initialization.

For a particle with continuous quality label $y \in [0, 1]$, the fine-tuning objective is

$$\mathcal{L}_{\mathrm{ranker}} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{rank}}. \tag{22}$$

The binary cross-entropy term is computed only for particles with unambiguous labels: particles with $y \leq 0.3$ are treated as negative, whereas those with $y > 0.7$ are treated as positive. Particles with intermediate labels do not contribute to this term. The margin-ranking term uses the continuous labels and is computed only for particle pairs from the same acquisition group that satisfy $0.2 < |y_i - y_j| < 1$, with a margin of 0.2.

Table 5: **CryoRanker fine-tuning hyperparameters.**

Setting	Value
Initialization	CryoVERSE online encoder
Training schedule	prediction head only for 5 epochs; joint encoder–head fine-tuning for 75 epochs
Global batch size	512
Optimizer / weight decay	AdamW / 10^{-4}
Nominal base learning rate	5×10^{-6}, scaled linearly by global batch size divided by 256
Head / backbone learning-rate multiplier	1.0 / 0.3
Learning-rate schedule / minimum ratio	cosine / 0.001
BCE label thresholds	negative if $y \leq 0.3$; positive if $y > 0.7$
Ranking margin / eligible label gap	0.2 / $0.2 < \lvert y_i - y_j \rvert < 1$
Compute / wall-clock time	2 NVIDIA A40 GPUs (48 GB each) / approximately 28.7 hours

D.2 PARTICLE-PICKING ADAPTATION

For particle-picking fine-tuning, we initialize the backbone from the pretrained CryoVERSE online encoder and replace the pretraining head with a randomly initialized dense prediction head. For each 672×672 input, the ViT-B/14 encoder produces a 48×48 grid of 768-dimensional patch tokens. We remove the class token and four register tokens, then reshape the remaining tokens into a spatial feature map. A shared 3×3 convolution maps these features to 256 channels, followed by batch normalization and ReLU. Two prediction branches then produce a 48×48 particle-center heatmap and a 48×48 two-channel sub-grid offset map. Each branch comprises a 3×3 convolution, ReLU, and a 1×1 output layer.

Particle-center annotations are mapped to the 48×48 prediction grid. For a particle center at continuous grid coordinates (x, y), the corresponding center cell and sub-grid offset are

$$(i, j) = (\lfloor x \rfloor, \lfloor y \rfloor), \qquad \boldsymbol{\delta} = (x - i,\ y - j). \tag{23}$$

The center target is an isotropic Gaussian heatmap centered at the continuous particle location. We convert the particle diameter to the heatmap scale and set $\sigma = \mathrm{clip}(D_{\mathrm{hm}}/6, 1, 6)$ grid cells, where D_{hm} is the diameter in heatmap coordinates. Overlapping Gaussian targets are combined by a pointwise maximum, while the offset target is defined only at each center cell. Training minimizes

$$\mathcal{L}_{\mathrm{pick}} = \mathcal{L}_{\mathrm{center}} + \mathcal{L}_{\mathrm{offset}}, \tag{24}$$

where $\mathcal{L}_{\mathrm{center}}$ is a CenterNet-style focal heatmap loss (Zhou et al., 2019). The offset term $\mathcal{L}_{\mathrm{offset}}$ is an ℓ_1 loss evaluated only at particle-center cells. Both terms use unit weight.

During training and inference, each selected micrograph is resized with its aspect ratio preserved so that its shorter side is 3072 pixels. We divide the resized image into overlapping 1536×1536 tiles with a stride of 1024 pixels. An edge-aligned tile is added when needed to cover the image boundary. All resulting tiles are retained, resized bilinearly to 672×672, and normalized using their individual means and standard deviations. No additional data augmentation is applied.

At inference, the sigmoid-transformed tile heatmaps are projected into a common heatmap in the resized-micrograph coordinate system. For overlapping regions, we retain the highest score and the corresponding offset from the same tile. Particle-center candidates are identified as 3×3 local maxima and refined using the predicted sub-grid offsets. The 48×48 prediction grid gives an effective stride of 32 pixels over each 1536×1536 tile. The decoded coordinates are mapped back to the original micrograph coordinate system. We then apply score-ordered, radius-based non-maximum suppression (NMS) to remove duplicate predictions. For each accession, the NMS radius is set to $\lceil D/2 \rceil$ pixels, where D is the metadata-defined particle diameter in the original micrograph coordinate system. This radius is not tuned on test performance.

Where applicable, the controlled variants in Figure 4 use the same pretrained checkpoints as the corresponding ablations in Table 2. All variants use the same prediction head, training data, and downstream optimization protocol unless otherwise specified. CryoVERSE (from scratch) initializes the complete model randomly, whereas CryoVERSE (head only) freezes the pretrained encoder and optimizes only the prediction head. CryoVERSE (w/o wide-field observations) uses the corresponding controlled pretraining checkpoint defined above. CryoVERSE (w/o MIM & register tokens) uses the corresponding ablation checkpoint, with the encoder and prediction head fine-tuned

Table 6: **Particle-picking fine-tuning hyperparameters.**

Setting	Value
Initialization	CryoVERSE online encoder
Training schedule	joint encoder–head fine-tuning for 80 epochs
Global batch size	16
Micrograph scaling	aspect-preserving resize to a short side of 3072 pixels
Tiling / network input	1536×1536 tiles with stride 1024 / 672×672 input
Tile sampling / augmentation	all grid tiles / none
Optimizer / weight decay	AdamW / 0.01
Head / backbone learning rate	10^{-5} / 10^{-6}
Learning-rate schedule / warm-up	constant / none
Compute / wall-clock time	1 NVIDIA RTX 3090 / approximately 7.7 hours

jointly. Because MIM and register tokens are removed together, this variant evaluates their combined contribution rather than isolating either component.

E FROZEN-FEATURE EVALUATION PROTOCOLS AND TASK-SPECIFIC RESULTS

E.1 SHARED FROZEN-FEATURE EVALUATION PROTOCOL

All frozen-feature evaluations use the final backbone class token without fine-tuning. Let $\mathbf{z}_i$ denote the ℓ_2-normalized, 768-dimensional feature of query i. For an eligible reference sample j, the similarity and neighbor weight are

$$s_{ij} = \mathbf{z}_i^\top \mathbf{z}_j, \qquad w_{ij} = \exp\left(\frac{s_{ij}}{T}\right), \qquad T = 0.5. \tag{25}$$

Because the features are normalized, s_{ij} equals their cosine similarity. Each query is excluded from its own reference set, while additional eligibility constraints are defined for each task. No learned prediction head or validation-tuned decision boundary is introduced. Metrics are computed within each dataset and summarized by an unweighted macro-average unless stated otherwise.

E.2 SPA STRUCTURE AND POSE

Let $\mathcal{N}_5(i)$ contain the five eligible reference particles with the largest similarity to query particle i. The class scores and predicted label are

$$S_i(c) = \sum_{j \in \mathcal{N}_5(i)} w_{ij} \mathbf{1}\{y_j = c\}, \qquad \hat{y}_i = \arg\max_c S_i(c). \tag{26}$$

Top-1 accuracy is the percentage of queries for which $\hat{y}_i = y_i$.

For pose awareness, reference neighbors are searched separately within each EMD dataset and the labels are discrete ground-truth pose classes. For structure awareness, the reference bank pools particles from all 12 datasets and the label is the structure identity; the query particle is again removed before retrieval. We report accuracy for each dataset and their unweighted macro-average. This leave-one-out protocol evaluates the local organization of a frozen representation and introduces no learned classifier or validation-tuned decision boundary.

For the generic visual baselines in Table 2, we used released MAE ViT-B/16 and MoCo v3 ViT-B/16 checkpoints without cryo-EM adaptation. Grayscale particle images were converted to three-channel inputs and processed using the image size and normalization expected by each released model. We extracted and ℓ_2-normalized the final global representation and applied the same leave-one-out weighted 5-NN evaluation. These comparisons evaluate direct transfer from generic visual pretraining rather than performance after cryo-EM fine-tuning.

Table 7: **Pose and structure classification comparison between Cryo-IEF and CryoVERSE.** Values are weighted 5-NN top-1 accuracies (%). Macro-average values are unweighted means across the 12 simulated datasets. Bold indicates the better result between Cryo-IEF and CryoVERSE for each dataset and task.

	Pose classification		Structure classification	
Dataset	Cryo-IEF	CryoVERSE	Cryo-IEF	CryoVERSE
EMD-0097	87.69	**93.21**	89.99	**94.42**
EMD-4608	64.60	**77.88**	96.08	**99.25**
EMD-9648	77.81	**88.50**	97.42	**98.85**
EMD-10102	98.97	**99.93**	99.32	**99.99**
EMD-10847	98.54	**99.71**	97.50	**99.74**
EMD-16651	96.15	**99.43**	98.75	**99.80**
EMD-24057	97.69	**99.67**	97.23	**99.53**
EMD-24823	93.93	**97.98**	95.09	**98.36**
EMD-26062	96.90	**99.31**	97.04	**99.33**
EMD-27990	83.57	**92.78**	97.39	**98.21**
EMD-29879	98.15	**99.70**	98.87	**99.89**
EMD-41856	99.72	**99.99**	98.59	**99.93**
Macro-average	91.14	**95.67**	96.94	**98.94**

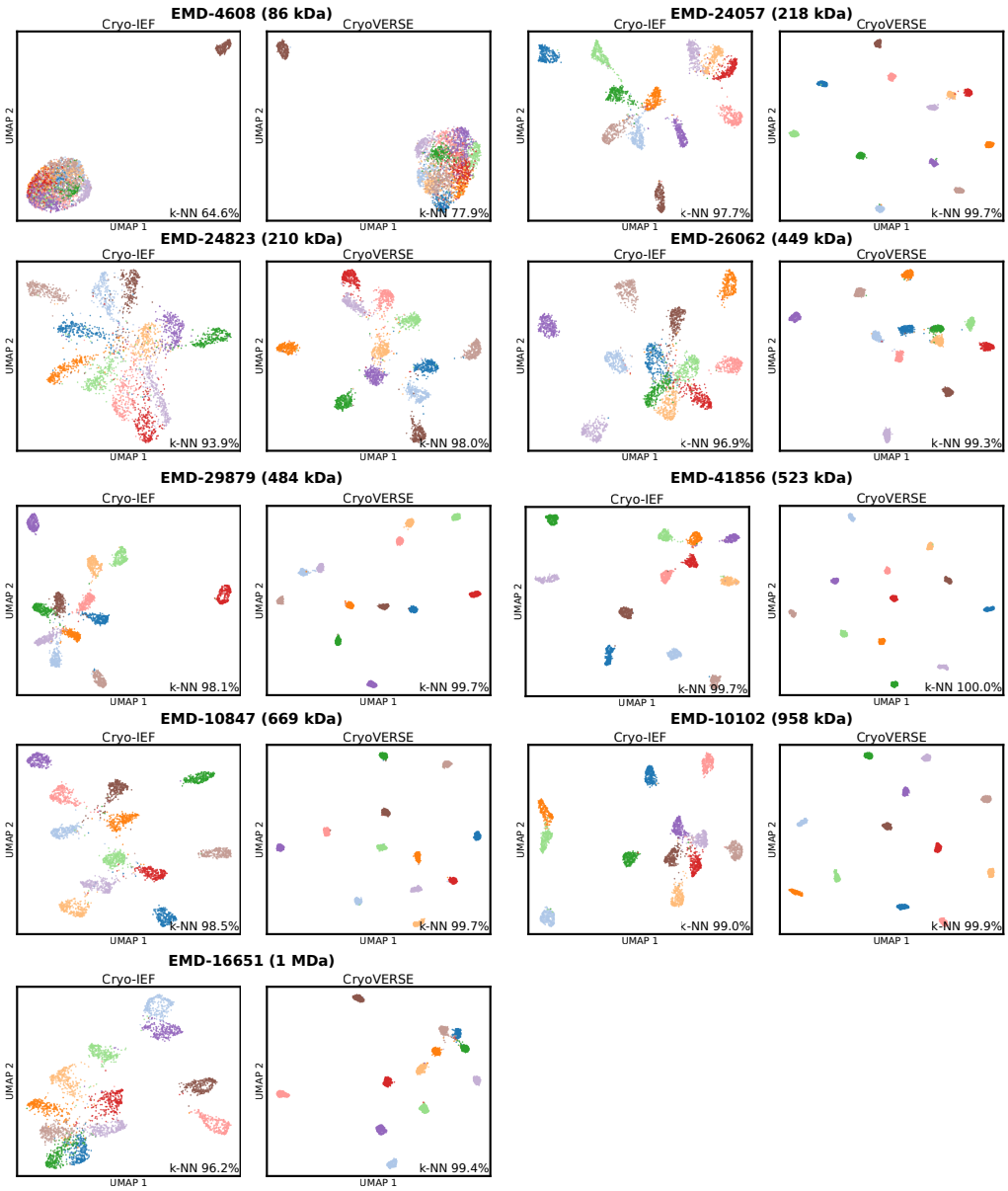

Figure 6: **CryoVERSE improves pose organization of SPA particles over Cryo-IEF.** UMAP visualizations compare pose organization under Cryo-IEF and CryoVERSE for nine representative datasets, with points colored by ground-truth pose classes. Precise dataset-level pose classification accuracies and the complete evaluation protocol are provided in Appendix Table 7 and Appendix E.2.

E.3 PARTICLE QUALITY AND RECONSTRUCTION

E.3.1 FROZEN-FEATURE PARTICLE-QUALITY EVALUATION

For frozen particle-quality evaluation, neighbors are searched separately within each of the four validation datasets. For query particle i, let $\mathcal{N}_5(i)$ contain the five particles with highest cosine similarity. The similarity-weighted proxy score is

$$\hat{y}_i = \frac{\sum_{j \in \mathcal{N}_5(i)} w_{ij} y_j}{\sum_{j \in \mathcal{N}_5(i)} w_{ij}}, \tag{27}$$

where $y_j \in [0, 1]$ is the curated continuous quality label. This leave-one-out regression uses validation labels from neighboring particles and is therefore a measure of frozen-feature readability, not a standalone unsupervised predictor.

For AUROC and AP, we binarize the continuous ground-truth quality labels as $b_i = \mathbf{1}\{y_i > 0.5\}$ while retaining $\hat{y}_i$ as a continuous ranking score; neither metric uses a fixed threshold on the predicted scores. Spearman and Pearson correlations instead compare $\hat{y}_i$ directly with the continuous label y_i. Each metric is computed independently within each dataset and then averaged without weighting across the four datasets. For Appendix Figure 7a, we selected 300 particles per dataset nearest to each target quality level (0, 0.5, and 1), yielding 3,600 particles, and applied UMAP to their normalized features only for qualitative visualization.

E.3.2 FINE-TUNED CRYORANKER EVALUATION

For direct CryoRanker evaluation, the model output is used without retrieving labels from other validation particles. Within each dataset, positive and negative samples, defined by the ground-truth label threshold $y_i > 0.5$, are balanced using a fixed random seed before AUROC, AP, Spearman, and Pearson are calculated. The continuous model outputs are supplied directly to AUROC and AP without thresholding. We report the unweighted mean of each metric over datasets. For the retention analysis, particles are sorted by predicted quality within each dataset; at each retained fraction, we compute the mean continuous quality label of the highest-ranked subset and then report the dataset macro-average and full across-dataset range.

E.3.3 RECONSTRUCTION EVALUATION

For the reconstruction-based evaluation, CryoRanker-VERSE scores were used to rank all particles independently within each of the four reconstruction datasets. Each ranked stack was partitioned into five equal-sized, non-overlapping quintiles, from the highest-scoring 20% to the lowest-scoring 20%. Each quintile was processed independently using the same cryoSPARC Non-Uniform Refinement protocol as in the reconstruction experiment reported for Cryo-IEF (Yan et al., 2026a), with reconstruction settings held fixed across the five quintiles of a dataset. We report the nominal resolution returned for each independently refined map. For the direct model comparison, the CryoRanker-IEF top-20% resolutions for EMPIAR-10250, EMPIAR-10405, and EMPIAR-10556 were taken from the published matched experiment; the corresponding EMPIAR-10560 result was reproduced using CryoRanker-IEF and the same protocol because this dataset was absent from the original evaluation. This experiment tests whether the predicted ranking enriches particles that support higher-resolution reconstruction; it does not assume or require a strictly monotonic resolution change across all five quintiles.

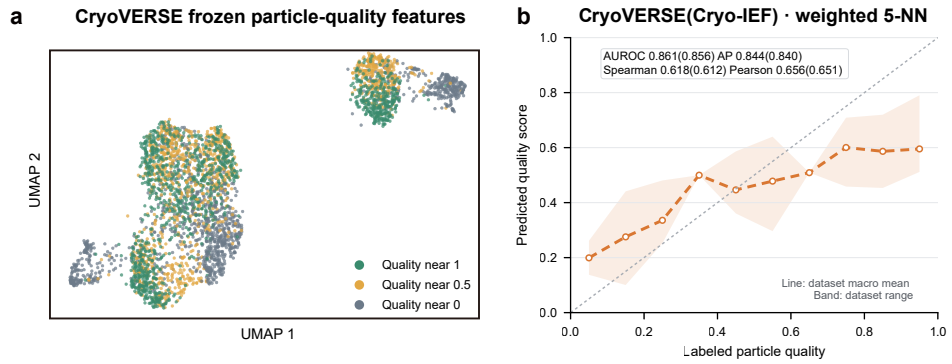

Figure 7: **Evaluation of particle quality assessment by CryoVERSE. a**, UMAP visualization of four held-out SPA datasets colored by curated quality labels. **b**, Weighted 5-NN estimates of continuous quality. The line and shading show the dataset macro-average and range scores, respectively.

Table 8: **Reconstruction resolutions from the top-ranked 20% of particles for four test datasets.** Values are nominal resolutions in Å; lower values are better. Bold indicates the best result for each dataset; ties are both bolded.

Dataset	CryoRanker-IEF (Å)	CryoRanker-VERSE (Å)
EMPIAR-10250	3.38	**3.13**
EMPIAR-10405	**3.29**	**3.29**
EMPIAR-10556	2.77	**2.72**
EMPIAR-10560	2.95	**2.93**

E.4 Micrograph particle-picking evaluation

We evaluated CryoVERSE on the CryoPPP accession-level splits defined in Appendix A.3. The generic visual baselines were MAE ViT-B/16 (He et al., 2022) and MoCo v3 ViT-B/16 (Chen et al., 2021). The cryo-EM baselines were Topaz (Bepler et al., 2019), DRACO (Shen et al., 2024), Cryo-EMMAE (Zamanos et al., 2025), and CryoTransformer (Dhakal et al., 2024).

For each generic visual encoder, we applied the dense prediction head used by CryoVERSE to spatial features from the released pretrained model. We converted grayscale inputs to three channels, applied the model-specific normalization, and jointly fine-tuned the encoder and prediction head.

The cryo-EM baselines retained their native task formulations and inference procedures. Topaz 0.3.18 used the official pretrained ResNet16-64 model (`resnet16_u64`), whereas DRACO used its official encoder and Detectron2-based Faster R-CNN architecture. Both received the same particle-coordinate annotations as CryoVERSE. Cryo-EMMAE instead used the same training micrographs without coordinate supervision and followed its native self-supervised reconstruction and latent-clustering pipeline. We therefore treat Cryo-EMMAE as a label-free reference, not an annotation-matched supervised baseline. All six baselines were trained for 80 epochs, and checkpoints were selected using only the validation accessions. Appendix D.2 details the adaptation of CryoVERSE and its controlled variants.

We evaluated all scored coordinates using a shared matching and metric protocol implemented in Topaz 0.3.18. For each accession, the matching radius was $\lceil D/2 \rceil$ pixels in the original micrograph, where D denotes the particle diameter in the CryoPPP metadata. The resulting radii were 112, 87, 86, 75, and 82 pixels for EMPIAR-10028, EMPIAR-10532, EMPIAR-10093, EMPIAR-10345, and EMPIAR-11056, respectively. These metadata-derived radii were fixed across methods and were not tuned on test results.

AUPRC was computed from the complete precision–recall curve for each test accession. For each model, we selected a single operating threshold by maximizing the unweighted macro-average F1-score across the five validation accessions. This threshold was then fixed and applied unchanged

to all five held-out test accessions when computing precision, recall, and F1-score. We applied the same validation-based threshold-selection procedure to every method and computed unweighted macro-averages across test accessions. The corresponding macro-averaged results are visualized in Figure 4.

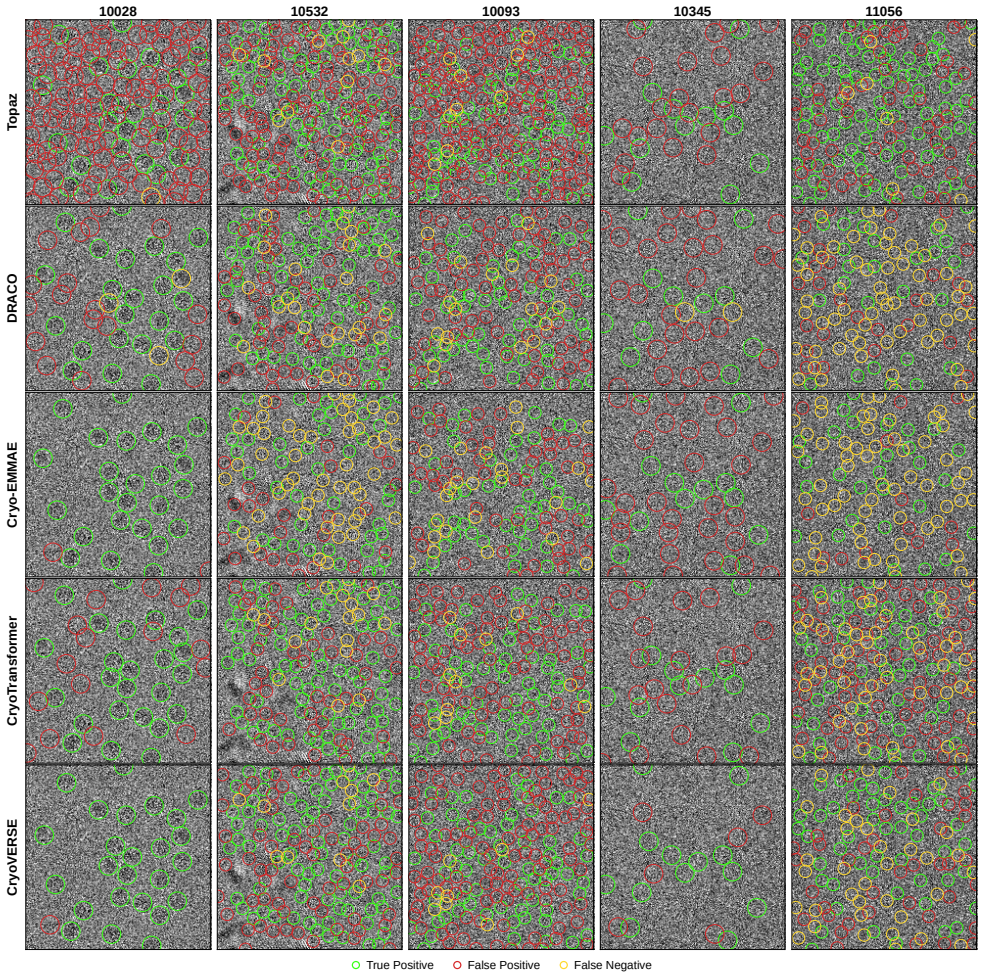

Figure 8: **Qualitative comparison of particle-picking predictions on five held-out CryoPPP test datasets.** Columns show representative micrograph regions from EMPIAR-10028, EMPIAR-10532, EMPIAR-10093, EMPIAR-10345, and EMPIAR-11056. Rows compare Topaz, DRACO, Cryo-EMMAE, CryoTransformer, and CryoVERSE. Green, red, and yellow circles denote true positives, false positives, and false negatives, respectively.

E.5 CRYO-ET REPRESENTATION EVALUATION

We evaluated frozen representations at two complementary spatial levels. Tilt images are complete projections acquired at individual angles of a cryo-ET tilt series, whereas particle views are annotation-linked crops of physical particles at particular tilts. Following the shared protocol above, tilt-image regression probes acquisition-angle information and particle-view classification probes structural identity. Both evaluations use the similarity and neighbor weights defined in Equation 25. UMAP is used only for qualitative visualization.

E.5.1 PARTICLE-VIEW SAMPLING

Particle-view datasets contain repeated two-dimensional observations of the same physical particle across multiple tilt angles. To preserve this multi-view structure while limiting particle-type imbalance, sampling was performed at the level of physical particles rather than individual views. A physical particle is one annotated three-dimensional instance whose projected locations yield multiple two-dimensional views with a shared particle identifier. For each evaluated dataset, we sampled 2,000 physical particles and retained all valid views associated with each selected particle.

Particle-type quotas were assigned using balanced water filling. An approximately equal quota was first allocated across available particle types. If a type contained fewer physical particles than its quota, all particles of that type were retained and the unused quota was redistributed. This procedure continued until 2,000 particles were selected or all available particles were exhausted. Within each type, physical particles were selected round-robin across acquisition groups, with particle order randomized using a fixed seed. This reduces domination by particle-rich groups while preserving deterministic sampling.

E.5.2 TILT-IMAGE ANGLE ORGANIZATION

For tilt-image evaluation, each sample consists of one two-dimensional tilt image together with its acquisition angle ϑ_i. We first visualized the frozen tilt-image representations using UMAP, with each point colored by its labeled tilt angle, to qualitatively inspect whether acquisition-dependent variation remains organized in the representation space. We quantified acquisition-angle organization using similarity-weighted 5-nearest-neighbor (5-NN) regression. For query tilt image i, $\mathcal{N}_5(i)$ contains its five nearest eligible tilt images, excluding the query itself. The acquisition angle is estimated as

$$\hat{\vartheta}_i = \frac{\sum_{j \in \mathcal{N}_5(i)} w_{ij} \vartheta_j}{\sum_{j \in \mathcal{N}_5(i)} w_{ij}}, \tag{28}$$

Here, ϑ_j denotes the recorded acquisition angle of neighbor j.

The evaluation included 662, 1,178, 2,945, and 1,860 valid tilt images from CZI-10301, CZI-10453, CZI-10468, and CZI-10469, respectively. Continuous angle recovery was summarized using mean absolute error (MAE, in degrees) and Spearman rank correlation. We additionally evaluated high-tilt discrimination using AUROC and average precision (AP), treating views with $|\vartheta| > 30^\circ$ as positive and using $|\hat{\vartheta}|$ as the continuous ranking score. Because the labeled angles of neighboring samples are used to construct $\hat{\vartheta}_i$, these measurements quantify the local readability of acquisition-angle information rather than standalone supervised prediction performance. Results are reported in Table 9.

E.5.3 PARTICLE-VIEW STRUCTURAL ORGANIZATION

Each particle view inherits the particle-type annotation of its source physical particle. For a query view i, neighbors are retrieved from the frozen feature bank using cosine distance. All views of the same physical particle as the query are excluded before retrieval, preventing repeated observations of one object from creating trivial correspondences.

Particle type is predicted using the similarity-weighted 5-NN classification rule in Equation 26. Here, $\mathcal{N}_5(i)$ contains the five nearest eligible particle views after applying the physical-particle exclusion above.

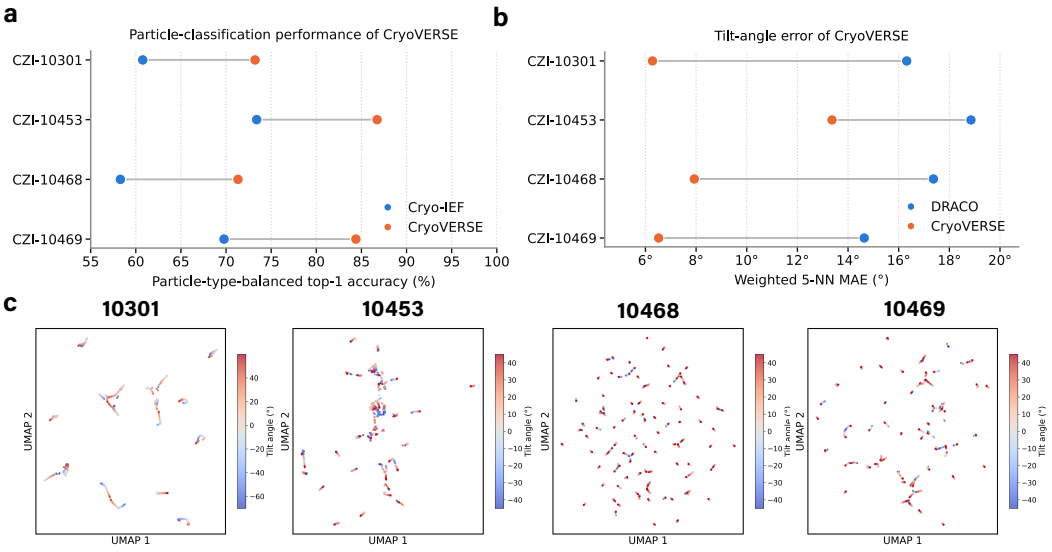

Figure 9: **Evaluation of CryoVERSE on cryo-ET tasks. a**, Cryo-ET particle-type classification performance. Particle-level type-macro accuracy of CryoVERSE and Cryo-IEF across four CZI datasets, evaluated from frozen particle-view representations. **b**, Cryo-ET tilt-angle estimation error. Mean absolute error (MAE) from similarity-weighted 5-NN angle regression on frozen tilt-image representations, comparing CryoVERSE with DRACO across four datasets. Lower values indicate more accurate recovery of acquisition angles. **c**, Cryo-ET tilt-angle representation organization. UMAP visualizations of frozen CryoVERSE tilt-image representations on four CZI datasets, with individual tilt views colored by acquisition angle.

We report particle-type classification at three complementary levels. At the view level, top-1 accuracy and macro-F1 are computed directly over individual particle views. Because each physical particle contributes multiple views, we additionally aggregate classification correctness within each physical particle r:

$$\mathrm{Acc}_r = \frac{1}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} \mathbf{1}\{\hat{y}_i = y_i\}, \tag{29}$$

where $\mathcal{I}_r$ denotes the retained views of particle r. To reduce sensitivity to particle-type imbalance, particle-level accuracies are then averaged within each structural type c:

$$\mathrm{Acc}_c = \frac{1}{|\mathcal{R}_c|} \sum_{r \in \mathcal{R}_c} \mathrm{Acc}_r, \tag{30}$$

where $\mathcal{R}_c$ is the set of sampled physical particles of type c. The primary dataset-level metric is the macro-average across structural types:

$$\mathrm{Acc}_{\text{type-macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{Acc}_c. \tag{31}$$

Thus, the evaluation distinguishes view-level classification quality from consistency across repeated views of the same physical particle, while the primary type-macro accuracy gives equal weight to each structural category within a dataset. Cross-dataset performance is summarized using the unweighted mean of the four dataset-level type-macro accuracies. Dataset-level results and particle-type-specific results are reported in Tables 10 and 11, respectively.

Table 9: **Tilt-angle estimation of CryoVERSE on four cryo-ET datasets.** CryoVERSE and DRACO are compared using weighted 5-NN regression for tilt-angle estimation. MAE measures angular error in degrees. Spearman's ρ measures agreement between estimated and recorded angle rankings. AUROC and AP measure discrimination of high-tilt images ($|\vartheta| > 30°$) using the absolute estimated angle. Bold indicates the better result for each dataset and metric.

Dataset	Model	MAE (°) ↓	Spearman ρ ↑	AUROC ↑	AP ↑
CZI-10301	DRACO	16.3169	0.6932	0.7779	0.8161
	CryoVERSE	**6.2776**	**0.9415**	**0.9359**	**0.9527**
CZI-10453	DRACO	18.8522	0.4250	0.7673	0.6778
	CryoVERSE	**13.3666**	**0.7324**	**0.7956**	**0.7076**
CZI-10468	DRACO	17.3731	0.5095	0.7776	0.6820
	CryoVERSE	**7.9283**	**0.9125**	**0.8741**	**0.8490**
CZI-10469	DRACO	14.6391	0.6552	0.7559	0.6734
	CryoVERSE	**6.5223**	**0.9336**	**0.8977**	**0.8791**
Macro-average	DRACO	16.7953	0.5707	0.7697	0.7123
	CryoVERSE	**8.5237**	**0.8800**	**0.8758**	**0.8471**

Table 10: **Particle-type classification of CryoVERSE on four cryo-ET datasets.** Cryo-IEF and CryoVERSE are compared using weighted 5-NN on the cryo-ET particle-type classification task. Type-macro accuracy, the primary metric, gives equal weight to each structure; mean particle accuracy gives equal weight to each physical particle. Both metrics first average classification correctness across the views of each particle. View macro-F1 is computed across structures at the view level. Macro-average values are unweighted means across the four datasets. Bold indicates the better result for each dataset and metric.

Dataset	Model	Type-macro acc. ↑	Mean particle acc. ↑	View macro-F1 ↑
CZI-10301	Cryo-IEF	0.6077	0.6316	0.6138
	CryoVERSE	**0.7324**	**0.7498**	**0.7411**
CZI-10453	Cryo-IEF	0.7340	0.7501	0.7362
	CryoVERSE	**0.8679**	**0.8877**	**0.8720**
CZI-10468	Cryo-IEF	0.5829	0.7948	0.5945
	CryoVERSE	**0.7133**	**0.8407**	**0.7329**
CZI-10469	Cryo-IEF	0.6978	0.8242	0.7180
	CryoVERSE	**0.8438**	**0.8950**	**0.8475**
Macro-average	Cryo-IEF	0.6556	0.7502	0.6656
	CryoVERSE	**0.7894**	**0.8433**	**0.7984**

Table 11: **Particle-type classification results at the physical-particle level.** For each particle species, we report the mean view-level classification accuracy over physical particles. **Bold** and underlined values indicate the best and second-best results, respectively. Macro-average is the unweighted mean over species.

Dataset	Particle type	N_{particle}	Cryo-IEF	CryoVERSE	CryoVERSE (w/o cryo-ET observations)	CryoVERSE (w/o pose-aware positives)
CZI-10301	F_1–F_0 complex	666	0.7508	**0.8367**	0.7721	0.7480
	RubisCO complex	333	0.3059	**0.3963**	0.3168	0.3568
	cytosolic ribosome	334	0.5631	**0.9064**	0.5732	0.6138
	microtubule	334	0.9306	**0.9908**	0.9666	0.9554
	nucleosome	333	0.4880	**0.5306**	0.4518	0.4846
	Macro-average	—	0.6077	**0.7324**	0.6161	0.6317
CZI-10453	clathrin coat	333	0.5873	0.6739	0.6443	**0.6743**
	clathrin complex	592	0.7825	**0.8885**	0.8604	0.8275
	V-type ATPase complex	549	0.7726	**0.9500**	0.8153	0.7704
	V-type ATPase V_0	526	0.7934	**0.9570**	0.8052	0.8248
	Macro-average	—	0.7340	**0.8679**	0.7813	0.7743
CZI-10468	cytosolic ribosome	1586	**0.9445**	0.9308	0.9196	0.9333
	virus-like capsid	414	0.2214	**0.4956**	0.3744	0.4802
	Macro-average	—	0.5829	**0.7133**	0.6470	0.7067
CZI-10469	cytosolic ribosome	1551	0.9273	0.9366	0.9263	**0.9448**
	virus-like capsid	449	0.4682	**0.7512**	0.5883	0.7437
	Macro-average	—	0.6978	0.8438	0.7573	**0.8443**