THE MICROSCOPE IS THE MASK: PRIVILEGED VIEWS AND LABELS FROM A CRYO-ET FORWARD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we use simulated data to train a model for protein annotation in crowded cryo-electron tomography volumes reconstructed from images collected at limited tilt angles and severely corrupted by the measurement operator. Firstly, we leverage the corruptions imposed by the forward model to generate domain-specific augmented views of the exact same scene for an invariance objective integrated into the LeJEPA self-supervised training framework. Secondly, we use additional information from the simulation pipeline such as the positions and identities of proteins in the simulated volumes to inform the architecture of the model and the loss function, so that semantic information is localised at protein positions in the dense feature volume. The resulting model, CARNIVAL, is evaluated without finetuning on classification and detection tasks on a multi-protein benchmark dataset and two in-situ datasets spanning three processing types. We show that CARNIVAL outperforms a state-of-the-art model trained using a contrastive objective on simulated data with a single labelled protein per input block, rather than dense supervision over multiple proteins in the scene, in both tasks on the benchmark dataset, matches or outperforms it in in-situ classification, and computes feature volumes seven times faster. A nine-arm ablation across the five evaluation settings decomposes the mechanism, showing that scene-level supervision is essential and pointing to how the views interact as the source of further gains.

1 INTRODUCTION

Cryo-electron tomography (cryo-ET) images frozen cells, yielding three-dimensional volumes where individual protein complexes sit in their native cellular context. The reconstructed tomograms are severely degraded by the physics of the measurement: the electron dose must be kept low to limit radiation damage, and the specimen can be tilted only over a restricted angular range, so part of Fourier space is never measured at all (a slice from a tomogram is shown in Figure 1A). Macromolecules are buried in noise and surrounded by dense, unlabelled cellular material. At the same time, the structures themselves are largely known, with over 200,000 experimentally determined entries in the RCSB Protein Data Bank (Burley et al., 2025) and AlphaFold (Bertoni et al., 2026) predicted models covering most of the proteome. The problem we address is connecting the two: given a tomogram, identify which known structures are present and where. Supervised learning on real tomograms is not a route to this, as annotations mostly cover large, abundant protein types and are missing precisely the small and rare proteins that are hard to find in increasingly larger datasets. We therefore turn to simulation as the source of training data at the scale required. Because most components of the forward model of a cryo-ET experiment are well characterised, this is a viable strategy.

In this paper, we propose using the forward model to generate physically-grounded and domain-specific augmented paired views of the same scene and additional information from the simulator to learn representations invariant to nuisance variables. For example, masked modelling (e.g. masked autoencoders (He et al., 2022) or the patch-masking objective in DINOv2 (Oquab et al., 2024)) involves tuning the mask ratio as a hyper-parameter and randomly masking patches. In contrast, the mask here is not a design choice: it is the missing wedge, a specific anisotropic region of Fourier space imposed by the microscope and determined by the acquisition parameters. As proteins have arbitrary orientations, each protein loses a different part of its spectrum due to the missing wedge.

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.

We incorporate these ideas into the LeJEPA self-supervised learning (SSL) framework (Balestriero & LeCun, 2025), yielding CARNIVAL (**C**ryo-ET **A**nnotation from **R**epresentations trained for **N**uisance-**I**nvariance via **V**iews **a**nd **L**abels), an encoder trained only on simulated tomograms that maps tomogram blocks to dense spatial features. We evaluate CARNIVAL without finetuning on three datasets of real tomograms, a multi-protein benchmark and two in-cell datasets, and compare it with TomoTwin (Rice et al., 2023), a state-of-the-art model also trained on simulated data, but with a triplet loss on subvolumes centred on a single labelled protein. CARNIVAL outperforms TomoTwin in both tasks on the benchmark and in classification on *C. reinhardtii*, is on par in classification on *S. pombe* and computes feature volumes about seven times faster, while TomoTwin detects better in five of the seven in-cell tomograms. Ablations over nine training recipes show that the identity loss across scenes is essential and reveal room for improvement in how the different views interact.

2 Related work

Self-supervised learning and augmentations Self-supervised vision models learn representations by requiring two views of the same image to map to nearby points in the latent space, and differ mainly in how collapse is avoided, for example through negatives (SimCLR (Chen et al., 2020)), teacher-student asymmetry (DINOv2 (Oquab et al., 2024)), or an explicit distributional prior on the embeddings (LeJEPA (Balestriero & LeCun, 2025)). The augmentations they use include random crops, flips, colour jitter and blur, chosen because they preserve the semantic content of natural images. Moreover, masked modelling (masked autoencoders (He et al., 2022) and the patch-level objective of DINOv2 (Oquab et al., 2024)) involves hiding part of the input and learning to predict it, with the mask ratio (e.g. 75 % for masked autoencoders) tuned as a hyper-parameter and the masked patches drawn at random. In our approach, the augmentations are given the measurement itself (tomogram corruptions), and the mask is the missing wedge, applied in Fourier space rather than to input patches, determined by the tilt angles used during acquisition (here $\pm 45^\circ$ to $\pm 60^\circ$). As proteins have arbitrary orientations, each loses a different part of its spectrum due to the same wedge.

Latent representations of physical systems Joint-embedding objectives are increasingly applied to physical systems: as surrogates for expensive simulation (He et al., 2019), as representations of spatiotemporal fields (Qu et al., 2026), and as world models whose latents are shown to encode physical structure (Maes et al., 2026). There the paired views are related by time evolution and the latent is judged by whether it supports prediction. A cryo-ET specimen is frozen, so the forward operator takes the place of the transition function. It is the known physics that generates the corrupted views, but it projects the scene through an instrument instead of advancing it in time, and the latent is judged by what identity information survives.

Annotation models for cryo-ET Many deep learning-based annotation methods in cryo-ET are fully supervised and are trained on hand-curated particle lists or segmentations, for example crYOLO (Wagner et al., 2019), DeepFinder (Moebel et al., 2021), DeePiCt (de Teresa-Trueba et al., 2023) and Easymode (So-Last et al., 2026). CARNIVAL belongs to the category of models that learn an embedding space in which particles of the same type are neighbours, and annotate by reference rather than by retraining. TomoTwin (Rice et al., 2023) is the earliest of these, trained with a triplet loss on subvolumes of simulated tomograms. Other recent methods include MiLoPYP (Huang et al., 2024), AffinityVAE (Famili et al., 2025), CryoSiam (Stojanovska et al., 2025), and ProPicker (Wiedemann et al., 2026). None of these methods use the forward model as the corruption that generates the paired views their invariance objective is applied to: where two views are used, they are related by generic image augmentations (intensity jitter, cropping, low and high-pass filtering, voxel dropout and additive noise) rather than by independent draws of the acquisition parameters. Moreover, those trained on simulated data generate it in advance and store it, which bounds how much of the parameter space a fixed dataset can cover. Our simulation runs on GPU inside the training loop, so new clean scenes are generated frequently (every 25 optimiser steps), and corrupted tomograms are generated using independent parameter draws for every batch, giving the pairing that the objective needs for invariance.

Domain randomisation Randomising simulator parameters so that real data is captured by the distribution of the training data was introduced for transferring robotic policies from simulation

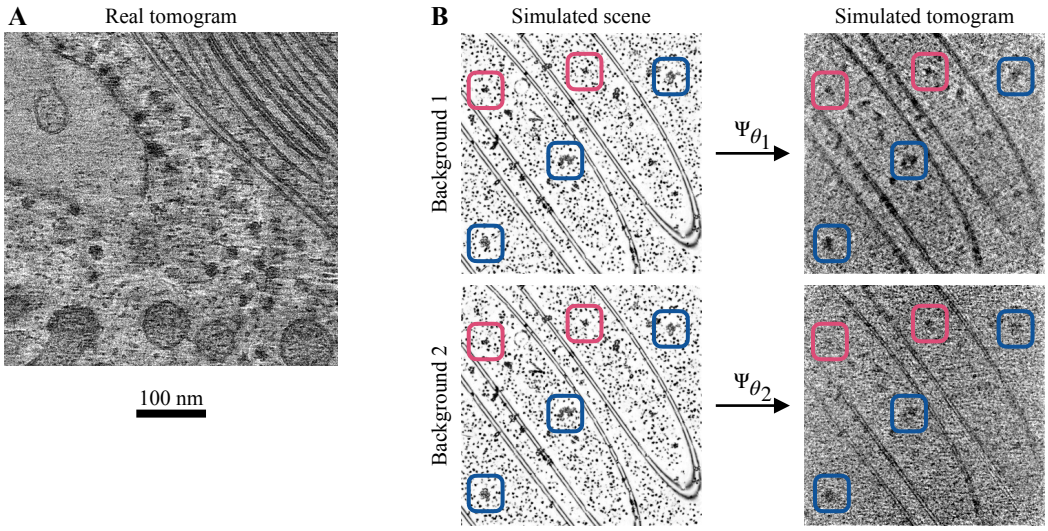

Figure 1: **A** Slice of real tomogram from (Kelley et al., 2026). **B** From one arrangement of proteins and membranes we generate six views (four shown). Two clean views (left) share those structures and differ only in the small-protein background. Each corrupted view (right) is obtained by applying the measurement operator Ψ to the clean view in the same row, at its own parameter draw θ_1 or θ_2 (e.g. tilt range, contrast transfer function, coloured noise). Blue and pink boxes mark two protein types; the same colour marks the same type in all four panels. Scale bar 100 nm is common to all panels.

in (Tobin et al., 2017), and recently applied to brain MRI data in (Billot et al., 2023). In the context of cryo-ET, domain randomisation with simulated data is used in (Harastani et al., 2025), where a crowded scene is generated by inserting foreground proteins and background particles at random positions, tilt images are simulated with the physics-based simulator Parakeet (Parkhurst et al., 2021) and tomograms are reconstructed with IMOD (Kremer et al., 1996), then it is shown that DeepFinder re-trained on this simulated data improves over the baseline model. In the same spirit, we generate clean crowded scenes (Appendix B) which we use to simulate diverse tomograms. We then further use domain randomisation to provide an invariance target: because the clean scene is held fixed while the parameters are redrawn, each pair of views tells the model which variation to discard, and so additional supervision is provided by the pairs. This idea is also employed in (Dey et al., 2025), where two views synthesised from the same randomly generated label map are used to train a nuisance-invariant general-purpose model for 3D biomedical volumes. The views share the geometric augmentations and only differ in appearance. Our views are instead related by the forward model of the microscope, so the corruption is sampled in the space of acquisition parameters rather than of image effects. This also gives every corrupted view a clean counterpart and adds a second, biological axis of variation (different crowded cellular environment and different protein orientations) that appearance randomisation does not provide.

3 Physically-grounded views and privileged labels

To leverage a possibly imperfect simulator of the image formation process, we generate three sources of information that real data alone cannot provide, used jointly during training:

1. **Paired clean and corrupted views of the same specimen: the physics axis.** Corruption due to the measurement process, applied in Fourier space (e.g. missing wedge). Real data has no clean counterpart. This defines the invariance the model learns (rows in Figure 1B).
2. **Paired views sharing structure positions but not orientations or background: the content axis.** The same proteins are embedded in a different background soup and with different orientations, with diversity due to the specimen rather than the instrument: a second view under a nuisance not covered by the measurement model (left column in Figure 1B).

3. **Complete scene labels.** Each protein's position and identity is known, so positives are built between distinct blocks that share identity. This is privileged information in the sense of (Vapnik & Vashist, 2009) (available during training, absent at deployment), and is what lets the objective operate on features read out at protein locations (protein labels in Figure 1B).

Together, these three sources make an imperfect simulator an asset rather than a liability. Two principles govern how we use it. *Invariance, not imitation*: the model is never asked to reproduce a simulated density, only to recognise what corrupted versions of one have in common. For example, the biology is the part of our simulator most obviously wrong (fixed library of structures, arbitrary crowding statistics, no real cellular context), which constrains how we use it: the content axis supplies a nuisance to be discarded, not a target to be matched. Therefore, measurement corruptions supply training signal in the form of invariances to be learned (with respect to such corruptions). Simulator error reaches the representation only insofar as it changes what survives corruption, so it is still important that the simulator captures the nuisance distribution accurately enough. *Coverage, not calibration*: the parameters governing corruption strength are not calibrated per dataset. Ranges are chosen so that plausible experimental configurations fall inside the sampled bounds, with the aim of making a single model usable across acquisition settings without per-dataset calibration dependent on a perfect simulator. This resembles domain randomisation (Tobin et al., 2017), except that the paired views make randomisation over forward model parameters the invariance target, not just input diversity.

3.1 PAIRED VIEWS

Let s be a 3D volume representing the structures of interest in the biological specimen: proteins at known positions $\{p_j\}$ and orientations $\{\omega_j\}$ with identities $\{c_j\}$, membranes, and other large-scale structures. We model the process of obtaining a tomogram x using the forward operator Ψ_θ, $x = \Psi_\theta(s + b)$, where the parameter vector θ encapsulates all image acquisition, reconstruction and processing parameters, and b is an independent instance of the random cellular background environment, a soup of small proteins below the detection limit and filling the intracellular space that generates a noisy and crowded background. We then write $\Psi_\theta = R_{\theta_R} \circ M_{\theta_M}$ with $\theta = [\theta_R, \theta_M]^T$, the composition of a measurement operator M_{θ_M} that represents the physics of the image acquisition process and outputs the collected tilt images, and a reconstruction operator R_{θ_R} that takes the tilt images and outputs the reconstructed tomogram, including all the processing steps applied during and after the reconstruction (e.g. filtering, denoising).

The three clean and three corrupted views v_{clean_i}, $v_{\text{corrupted}_i}$ used to train the model are

$$v_{\text{clean}_i} := s_i + b_i, \qquad v_{\text{corrupted}_i} := \Psi_{\theta_i}(v_{\text{clean}_i}), \qquad \text{for} \qquad i = 1, 2, 3,$$

where $s_1 = s_2 = s$, and s_3 keeps the positions $\{p_j\}$ and identities $\{c_j\}$ of s but redraws the orientations $\{\omega_j\}$. The first two views draw their own parameters and background, $\theta_i \sim p(\theta)$ and $b_i \sim q(b)$ for $i = 1, 2$, while the third reuses those of the first, $\theta_3 = \theta_1$ and $b_3 = b_1$, so that it differs from v_1 in protein orientation alone. Examples of clean and corrupted views for $i = 1, 2$ are shown in Figure 1B.

The measurement operator M_{θ_M} is further decomposed into a sequence of transformations, each with their own parameters that we sample from their distributions during training. For example, to simulate the tilt images, the maximum tilt angle $\theta_{\max}$ is sampled from $\{45, 50, 55, 60\}^\circ$, the tilt increment $\Delta\theta$ from $\{2, 3, 5\}^\circ$, the defocus is sampled from $[1, 5]\mu m$, and the noise colour exponent from $[0, 1.2]$. The reconstruction operator R_{θ_R} is the standard weighted backprojection algorithm with a regularisation parameter applied in Fourier space, whose value is sampled during training. The full list of transformations with their ranges is given in Appendix C, Table 5. Both the clean data generation (Appendix B) and the forward model transformations are applied on-the-fly (implemented using PyTorch and run on GPU), enabling the generation of clean and corrupted views.

3.2 IDENTITY LOSS ACROSS SCENES

CARNIVAL consists of an encoder backbone E, a readout network F, and an optional segmentation head G, trained jointly. E maps a simulated volume x to a dense spatial feature volume $y = E(x)$ at half the input resolution. For a protein at position p in x, the local patch centred at p in the feature volume y is mapped by the readout network to a 512-dimensional vector

$z(x,p) = F\big(\mathrm{crop}_p(E(x))\big)$. To fully leverage the multi-protein blocks with known protein positions and identity made possible by the simulator, the loss $\mathcal{L}_{\text{scene}} = \frac{1}{|\mathcal{P}_{\text{scene}}|}\sum_{\mathcal{P}_{\text{scene}}} \|z - z'\|^2$ is applied to all pairs $(z_i, z'_j) \in \mathcal{P}_{\text{scene}}$, the set of embeddings obtained from crops of proteins with distinct positions but the same identity. The full loss is $\mathcal{L} = \mathcal{L}_{\text{view}} + \lambda_{\text{s}}\mathcal{L}_{\text{scene}} + \lambda_{\text{r}}\mathcal{L}_{\text{SIGReg}} + \lambda_{\text{g}}\mathcal{L}_{\text{seg}}$. LeJEPA (Balestriero & LeCun, 2025) is $\mathcal{L}_{\text{view}}$ (same scene, different view) plus the regularisation $\mathcal{L}_{\text{SIGReg}}$, which is applied on batch subsets with no repeating proteins.

The paired views of Section 3 (physics and content axes) are consumed by $\mathcal{L}_{\text{view}}$, and the scene labels by $\mathcal{L}_{\text{scene}}$ and by $\mathcal{L}_{\text{seg}}$, an auxiliary loss for training a segmentation head using protein and membrane masks from the clean data.

4 Experiments

We evaluate CARNIVAL frozen, with no finetuning, on three real datasets, and compare it with TomoTwin (Rice et al., 2023), which is also trained only on simulated data, but with a triplet loss on subvolumes centred on a single protein. All comparisons with TomoTwin use one CARNIVAL checkpoint, the last of training (Appendix A). Tomograms are used at a voxel spacing of $\sim$10 Å, the spacing of CARNIVAL's training data; datasets deposited at other spacings are resampled in Fourier space. We evaluate two tasks. In **classification**, the particles' positions are given, and each particle's embedding is classified by a leave-one-out k-nearest-neighbour vote ($k = 5$) over all particles in the dataset; we report the F_1 score per protein type and its mean over types (macro F_1). In **detection**, particles are located by their cosine distance to five template particles per type, drawn at random 30 times, and we report F_4, which weights recall 16 times over precision, following the Object Identification challenge (Peck et al., 2026a). Both protocols are detailed in Appendix D.

Phantom dataset The public test set of the CZII CryoET Object Identification challenge (Peck et al., 2025; 2026a) (DS-10445 in the CryoET Data Portal (Ermel et al., 2024)) contains 121 tomograms of a phantom sample with six protein types, each deposited both band-pass filtered and additionally denoised. We evaluate five types (ribosome, apoferritin, PP7 VLP, thyroglobulin and β-galactosidase; 15,226 particles), leaving out β-amylase, which the challenge did not score because of low confidence in its annotations (Peck et al., 2026a). Detection uses two of the tomograms in both processing forms: `TS_5_2`, which contains all five types, and `TS_43_9`, which is dominated by apoferritin.

***S. pombe* datasets** DS-10000 and DS-10001 (de Teresa-Trueba et al., 2023) each contain 10 tomograms of *S. pombe* cells, imaged with defocus contrast and with a Volta phase plate respectively, with annotated ribosomes and fatty acid synthase (FAS) complexes (26,267 and 26,042 particles). We use the raw reconstructions, resampled from 13.48 to 10 Å. The annotations are picks of the convolutional network of the same study, trained on template-matching results, then filtered and completed by hand. Detection uses two tomograms from each dataset.

***C. reinhardtii* dataset** From DS-10301 (Khavnekar et al., 2023) we use three denoised tomograms of *C. reinhardtii* cells, chosen because they have ground-truth annotations of two particle types, ribosomes and nucleosomes (3,825 particles). They are resampled from 7.84 to 10 Å and used for both tasks. The annotations are the particles retained after 3D classification, not every particle in the volume, so detection precision on this dataset is a lower bound.

4.1 Classification

To evaluate the quality of the learned embeddings, we test how well they discriminate between protein types, given the proteins' ground-truth positions in the tomograms. On the Phantom dataset (Figure 2A), CARNIVAL leads TomoTwin on every protein type in both processing forms, with an overall F_1 score (averaged over protein types) of 0.90 vs 0.86 on the filtered tomograms and 0.88 vs 0.80 on the denoised ones; the largest margins are on PP7 VLP and β-galactosidase. Denoising lowers CARNIVAL's overall score by 0.02 and TomoTwin's by 0.06. TomoTwin's workflow prescribes tomograms that are not denoised, so the denoised condition is outside its intended use; we include it because the portal provides both forms. In cells (Table 1), both models identify the abundant class almost perfectly (ribosomes in *S. pombe*, 0.99; nucleosomes in *C. reinhardtii*, 0.97 vs 0.95), so the

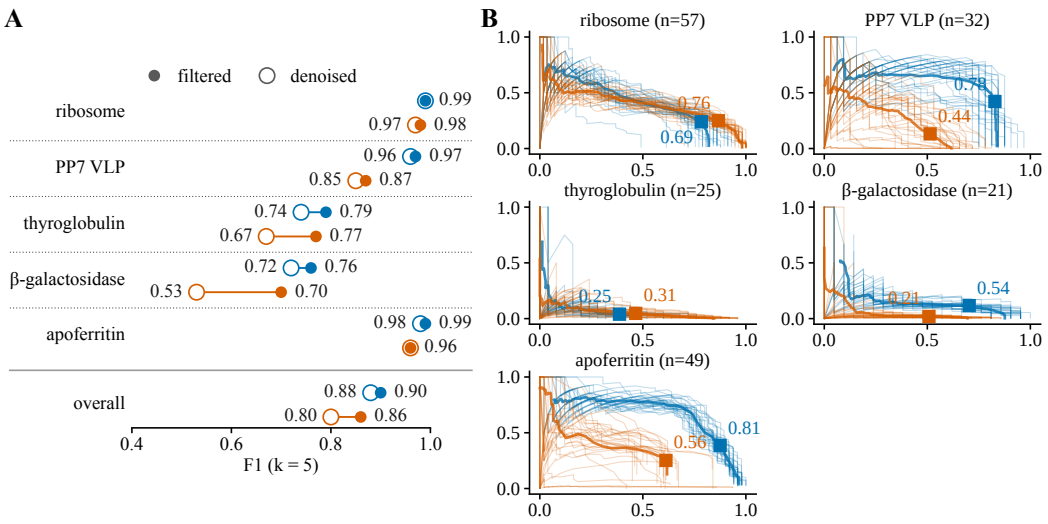

Figure 2: **CARNIVAL** vs **TomoTwin** on the Phantom dataset: **A** Classification F_1 score at $k = 5$, per protein type and overall (averaged over protein types), on filtered (filled) and denoised (open) tomograms. CARNIVAL leads on every type in both processing forms (overall 0.90 vs 0.86 filtered, 0.88 vs 0.80 denoised). **B** Detection precision (vertical) against recall (horizontal) on one filtered tomogram (`TS_5_2`) containing five protein types. Faint curves are individual template draws (5 templates/draw, 30 draws), bold curves their mean. The square marks the best F_4 over the threshold sweep per method.

Table 1: Classification in cells: F_1 per protein type and macro F_1 (the mean over the two types) at $k = 5$ (FAS: fatty acid synthase). Tomograms are raw reconstructions for DS-10000 and DS-10001, and denoised for DS-10301. The standard deviation of **CARNIVAL**'s values over its last ten checkpoints is at most 0.012; **TomoTwin** is one fixed model. On *S. pombe*, the sign of the difference between the two models depends on k (Appendix D.1).

	S. pombe, defocus DS-10000			*S. pombe*, phase plate DS-10001			*C. reinhardtii* DS-10301		
	ribosome	FAS	macro	ribosome	FAS	macro	ribosome	nucleosome	macro
# particles	25,901	366		25,311	731		362	3,463	
CARNIVAL	0.994	0.253	0.624	0.988	0.345	0.667	0.666	0.972	0.819
TomoTwin	0.993	0.280	0.637	0.987	0.296	0.641	0.321	0.952	0.636

comparison rests on the rarer class. On *C. reinhardtii*, CARNIVAL identifies ribosomes far better (0.67 vs 0.32), a lead of 0.18 in macro F_1. On *S. pombe*, the two models are on par: FAS scores 0.25 vs 0.28 with defocus contrast and 0.35 vs 0.30 with the phase plate, and the sign of the macro F_1 difference depends on k, whereas the leads on the Phantom and *C. reinhardtii* datasets hold at every k we tried (Appendix D.1).

4.2 Detection

We follow TomoTwin's reference-based protocol, replacing its two manual choices — reference selection and distance threshold — with five references randomly drawn per type over 30 draws and a sweep over the threshold to compute precision-recall curves, so the results do not depend on dataset-specific tuning. For each type, we report the best F_4 score over the sweep (an upper bound available to neither method at deployment) and summarise a tomogram by the mean over its evaluated types, weighted as in the challenge on the Phantom dataset (Appendix D.2). On the Phantom dataset (Table 2), CARNIVAL leads on both tomograms in both processing forms. Figure 2B shows the per-type precision-recall curves for the filtered tomogram `TS_5_2`: TomoTwin leads on ribosome (0.757 ± 0.049 vs 0.691 ± 0.073) and thyroglobulin (0.305 ± 0.079 vs 0.255 ± 0.070), while CAR-

Table 2: Detection on the Phantom dataset (DS-10445): weighted mean of the best F_4 over the threshold sweep for each evaluated protein type in that tomogram (five and two respectively). The unweighted mean ranks the two methods the same way in every condition. Per-type values, with their variability across template draws, are given in Table 7.

Tomogram	Filtered		Denoised	
	CARNIVAL	TomoTwin	CARNIVAL	TomoTwin
TS_5_2	**0.555**	0.397	**0.538**	0.374
TS_43_9	**0.803**	0.676	**0.698**	0.682

Table 3: Detection in cells: the mean, over the two evaluated protein types of each tomogram, of the best F_4 over the threshold sweep (ribosome and fatty acid synthase in *S. pombe*, ribosome and nucleosome in *C. reinhardtii*). Tomograms are raw reconstructions in DS-10000 and DS-10001 and denoised in DS-10301; the DS-10301 annotations are a classification-filtered subset of the particles, so precision there is a lower bound. Per-type values, with their variability across template draws, are given in Tables 8–10.

Dataset	Tomogram	CARNIVAL	TomoTwin
S. pombe, defocus (DS-10000)	TM-623	0.253	**0.328**
	TM-625	0.319	**0.407**
S. pombe, phase plate (DS-10001)	TM-2269	0.346	**0.358**
	TM-2272	0.327	**0.399**
C. reinhardtii (DS-10301)	TM-14317	**0.092**	0.073
	TM-14327	0.224	**0.250**
	TM-14331	**0.140**	0.119

NIVAL's leads are much larger (VLP: 0.784 ± 0.037 vs 0.438 ± 0.176, apoferritin: 0.814 ± 0.030 vs 0.565 ± 0.145, β-galactosidase: 0.543 ± 0.075 vs 0.205 ± 0.080; $\pm$ is one standard deviation across the 30 template draws), and on the denoised version of the same tomogram CARNIVAL leads on all five types (Figure 4).

In cells (Table 3), both methods score lower, and TomoTwin leads on five of the seven tomograms. On *S. pombe*, where both methods detect FAS poorly ($F_4 \leq 0.09$, except TomoTwin's 0.18 on TM-2272), the tomogram score is effectively that of the ribosome: TomoTwin leads on it by about 0.15 with defocus contrast, and the two methods tie with the phase plate. On *C. reinhardtii*, CARNIVAL leads on nucleosomes in all three tomograms and TomoTwin on ribosomes in two. Per type, CARNIVAL leads on apoferritin, VLP, β-galactosidase and nucleosomes, and TomoTwin on ribosomes in seven of the eleven volumes. This does not contradict the classification results: detection also requires rejecting the cellular background, which classification at annotated positions does not test. Both models identify *S. pombe* ribosomes at 0.99, so the gap in detecting them is not one of identity; and although CARNIVAL identifies *C. reinhardtii* ribosomes far better (0.67 vs 0.32), TomoTwin detects them better in two of the three tomograms. Per-type curves and F_4 tables for every tomogram are given in Appendix D.2.

Timing On the same six-GPU node, with both methods using all six GPUs and run separately, computing the feature volume of a Phantom tomogram ($184 \times 630 \times 630$ voxels) takes CARNIVAL 4.2 minutes and TomoTwin 30.7 minutes, averaged over six tomograms: TomoTwin takes 7.3 times as long. Each method runs at the setting used for every result above (stride 8 for CARNIVAL, stride 2 for TomoTwin, both in single precision), and both produce feature volumes at half the tomogram sampling.

4.3 Ablations

To see which parts of the recipe these results depend on, we train nine models from scratch for $10{,}000$ steps each, with the same seed, learning-rate schedule and effective batch size, and compare

them on classification (Table 4), which reads the embedding at annotated positions and so isolates the representation from the detection procedure. The reference arm **R** uses CARNIVAL's recipe, with which CARNIVAL itself was trained for about $84{,}000$ steps; the arms therefore sit below CARNIVAL in absolute terms, and only their differences from R carry the argument. Every other arm changes one element of R. **A3** and **A4** remove a loss term, setting $\lambda_{\mathrm{s}} = 0$ ($\mathcal{L}_{\mathrm{scene}}$) and $\lambda_{\mathrm{g}} = 0$ ($\mathcal{L}_{\mathrm{seg}}$) respectively; the terms are still computed, so both arms see the same data as R. Three arms change how the views are built. **A1** draws both views of a pair from the same corrupted volume and separates them by generic 3D augmentations only: the forward model still shapes every input, but no longer generates the difference between the views. **A2** keeps R's views but drops the terms of $\mathcal{L}_{\mathrm{view}}$ that match clean views to corrupted ones (the physics axis), so a corrupted view is no longer matched to its clean counterpart, although it still enters $\mathcal{L}_{\mathrm{scene}}$, $\mathcal{L}_{\mathrm{SIGReg}}$ and $\mathcal{L}_{\mathrm{seg}}$. **A6** replaces the clean member of each pair with a second reconstruction of the same clean volume under an independent parameter draw, as two acquisitions of one specimen would give. **A5** combines A1's views with $\lambda_{\mathrm{s}} = \lambda_{\mathrm{g}} = 0$, and **RP** and **A6P** add A1's generic augmentations to the views of R and A6. The generic augmentations are similar to those used by CryoSiam (Stojanovska et al., 2025). They are applied to each view independently, in a fixed order, each with its own probability and with parameters drawn uniformly from the given ranges: an intensity rescaling by a factor in $[0.2, 1.3]$ plus an offset in $[-0.3, 0.3]$ (probability 0.8; tomograms are normalised to unit variance), a Gaussian low-pass filter with $\sigma \in [0.5, 1.5]$ voxels (0.6), additive white Gaussian noise with a standard deviation of 0.1–1.5 times the block's own (0.8), and unsharp masking, $x + \alpha\,(x - G_\sigma * x)$ with $\sigma \in [1, 3]$ voxels and $\alpha \in [0.2, 1.5]$ (0.4).

The results rest on the identity loss. Removing it costs 0.08–0.16 macro F_1 in four of the five settings (A3), and removing it together with $\mathcal{L}_{\mathrm{seg}}$ from A1 costs 0.06–0.15 in the same four (A5); on *C. reinhardtii*, neither changes the result. The segmentation loss has no consistent effect on classification (A4 differs from R by -0.03 to $+0.03$). The headroom is in the view construction. A1 matches R on the Phantom dataset (-0.01 and -0.00) and improves in-cell classification by 0.08–0.12, which on these datasets means the rare class: FAS F_1 rises from 0.19 to 0.43 with defocus contrast and from 0.37 to 0.62 with the phase plate, and *C. reinhardtii* ribosome F_1 from 0.57 to 0.72. Training R's recipe about eight times longer, as for CARNIVAL, narrows A1's lead on *C. reinhardtii* from 0.08 to 0.03, but on *S. pombe* A1 at $10{,}000$ steps stays ahead of CARNIVAL by 0.09 and 0.14. No other view arm improves consistently: A6 loses 0.04–0.11 on the Phantom and *C. reinhardtii* datasets, and A2's differences are mixed (-0.03 to $+0.07$). Adding generic augmentation to R's views (RP) does not reproduce A1's gain on *S. pombe* (-0.02 and -0.03), though it does on *C. reinhardtii* ($+0.065$ against A1's $+0.080$), and adding it to A6's views (A6P) brings A6 back to within 0.035 of R everywhere. What sets A1 apart is that the two views of a pair share one corruption, the smallest possible physics difference between them, while every input still carries the full range of physics. This suggests, as a hypothesis rather than a result, keeping full-range physics in every input while pairing views that differ less in it.

Each arm is one seed, and CARNIVAL's recipe was fixed before the in-cell ablation results were available. The differences we interpret (0.04 and larger) are several times the checkpoint-to-checkpoint standard deviations in Table 4 (at most 0.011), and every difference from R larger than 0.05 keeps its sign at $k = 1$, 3 and 10 (Appendix D.1). RP and A6P were trained on a different machine from the other arms, at the same effective batch size, so their differences from R include that change.

5 Discussion

CARNIVAL outperforms TomoTwin on the Phantom dataset and in classification on *C. reinhardtii*, and is on par with it in classification on *S. pombe*. The ablations show that the identity loss across scenes, which relies on complete scene labels that only a simulator provides, is essential in four of the five settings, and they reveal room for improvement in how the different views interact: A1, whose two views share one corruption, matches R in classification on the Phantom dataset and outperforms it in cells.

The proposed strategy has two fundamental limitations, neither specific to cryo-ET. Firstly, coverage protects against mis-set ranges, not missing corruption axes. Widening a sampled range guards against a parameter whose real value is not estimated accurately, but if real data varies along a dimen-

Table 4: Ablations: classification macro F_1 at $k = 5$ for arms trained from scratch for 10,000 steps with one seed, each changing one element of the reference recipe R (CARNIVAL's recipe). A1: views separated by generic augmentation of one corrupted volume; A2: no matching of clean to corrupted views; A3: no $\mathcal{L}_{\text{scene}}$; A4: no $\mathcal{L}_{\text{seg}}$; A5: A1 without $\mathcal{L}_{\text{scene}}$ and $\mathcal{L}_{\text{seg}}$; A6: two reconstructions in place of a clean view and its reconstruction; RP and A6P: R and A6 with generic augmentation. Each value is taken at the last step, $\pm$ the standard deviation over the last ten checkpoints (steps 9,100–10,000): how much the value depends on where training stops, not variation across seeds. n is the number of particles. RP and A6P, below the rule, were trained on a different machine at the same effective batch size. Per-type values for every arm, CARNIVAL and TomoTwin are given in Appendix E.

Model n	Phantom filtered 15226	Phantom denoised 15226	DS-10000 26267	DS-10001 26042	DS-10301 3825
R	0.878 ± 0.002	0.875 ± 0.001	0.592 ± 0.004	0.681 ± 0.003	0.765 ± 0.007
A1	0.863 ± 0.001	0.873 ± 0.002	0.710 ± 0.004	0.805 ± 0.002	0.845 ± 0.003
A2	0.858 ± 0.001	0.863 ± 0.003	0.565 ± 0.003	0.751 ± 0.003	0.739 ± 0.004
A3	0.734 ± 0.004	0.795 ± 0.002	0.496 ± 0.002	0.522 ± 0.003	0.756 ± 0.011
A4	0.876 ± 0.001	0.868 ± 0.001	0.589 ± 0.005	0.649 ± 0.003	0.795 ± 0.006
A5	0.799 ± 0.003	0.796 ± 0.004	0.557 ± 0.009	0.696 ± 0.008	0.845 ± 0.002
A6	0.823 ± 0.002	0.836 ± 0.002	0.596 ± 0.007	0.718 ± 0.003	0.658 ± 0.011
RP	0.876 ± 0.002	0.873 ± 0.001	0.575 ± 0.006	0.648 ± 0.005	0.830 ± 0.004
A6P	0.858 ± 0.001	0.865 ± 0.002	0.624 ± 0.008	0.715 ± 0.005	0.761 ± 0.003

sion the forward model does not capture, no width helps and nothing signals the omission. Secondly, the corruption axes are not equally trustworthy. The missing wedge follows exactly from acquisition geometry and the CTF is a well-established physical model, but radiation damage and noise colour are empirical approximations, whose particular forms we chose partly due to their simplicity and runtime, so that views could be generated on-the-fly on GPU during training. An approximate corruption model risks over-invariance, where the model discards variation the measurement still carries. This is one possible explanation for A1's gains: matching a view to its clean counterpart or to an independently corrupted view asks for invariance to every corruption in the sampled range, including those modelled only approximately, whereas matching two views that share one corruption does not. Our ablations change how views are paired, not which corruption axes are sampled, so how best to sample an unreliable axis remains open and will require ablating the axes individually.

In cells, TomoTwin detects better in five of the seven tomograms, although CARNIVAL identifies the annotated particles on par or better. Detection also requires rejecting the cellular background, which in our objective only $\mathcal{L}_{\text{seg}}$ addresses (all other terms are computed on features at protein positions), and our simulated background, a soup of small proteins, lacks real cellular context (Section 3). Conversely, CARNIVAL detects nucleosomes, the smallest complex we evaluate (about 200 kDa), better than TomoTwin in all three tomograms that contain them, and small proteins are precisely those that existing annotations miss (Section 1).

Finally, the annotations of both the Phantom and the in-cell datasets can miss many particles: for example, when annotations come from subtomogram averaging, many initially picked particles are discarded along the way, real ones among them, and detecting those counts as a false positive. Besides further datasets and benchmarks such as POPSICLE (Schwartz et al., 2026), this calls for an evaluation that does not depend on existing annotations, for example by subtomogram averaging of the detected and classified particles.

AI USE STATEMENT

In this work, we used generative AI tools for creating and editing software code, finding additional relevant related work, generating the tables of results from the outputs of the scripts implementing the experiments, for improving the layout of the figures, and for drafting and editing parts of the paper text, including checking that the stated claims accurately reflect the results of the experiments. We have reviewed all AI-assisted work. In particular, all the code has been reviewed by the first author and regularly checked for correctness. The code that generates the training data and computes

the evaluation metrics is unit tested, and every number in the tables was checked against the raw outputs of the experiments. We have read all the suggested articles to ensure they are relevant for this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The details of the model, the training data generation, and the training objective are given in the main text of the paper in Section 3 and in Appendices A–C. The details required for reproducing the experiments are given in Section 4 and Appendix D. All evaluation datasets are public on the CryoET Data Portal, and Appendix D lists the tomograms and annotation depositions used. Upon article acceptance, we will make the model weights and all the code public, including scripts and documentation for running the experiments shown in the paper and for allowing users to run CARNIVAL on their own data.

References

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

Damian Bertoni, Maxim Tsenkov, Paulyna Magana, Sreenath Nair, Ivanna Pidruchna, Marcelo Querino Lima Afonso, Adam Midlik, Urmila Paramval, Dare Lawal, Ahsan Tanweer, Meera Last, Risha Patel, Agata Laydon, Dariusz Lasecki, Nick Dietrich, Hamish Tomlinson, Augustin Žídek, Tim Green, Oleg Kovalevskiy, Andy Lau, Shaun Kandathil, Nicola Bordin, Ian Sillitoe, Milot Mirdita, David Jones, Christine Orengo, Martin Steinegger, Jennifer R Fleming, and Sameer Velankar. AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. *Nucleic Acids Research*, 54(D1):D358–D362, 2026. doi: 10.1093/nar/gkaf1226. URL https://academic.oup.com/nar/article/54/D1/D358/8340156.

Benjamin Billot, Colin Magdamo, You Cheng, Steven E. Arnold, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences*, 120(9):e2216399120, 2023. doi: 10.1073/pnas.2216399120. URL https://www.pnas.org/doi/10.1073/pnas.2216399120.

Stephen K Burley, Rusham Bhatt, Charmi Bhikadiya, Chunxiao Bi, Alison Biester, Pratyoy Biswas, Sebastian Bittrich, Santiago Blaumann, Ronald Brown, Henry Chao, Vivek Reddy Chithari, Paul A Craig, Gregg V Crichlow, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Justin W Flatt, Sutapa Ghosh, David S Goodsell, Rachel Kramer Green, Vladimir Guranovic, Jeremy Henry, Brian P Hudson, Michael Joy, Jason T Kaelber, Igor Khokhriakov, Jhih-Siang Lai, Catherine L Lawson, Yuhe Liang, Douglas Myers-Turnbull, Ezra Peisach, Irina Persikova, Dennis W Piehl, Aditya Pingale, Yana Rose, Jared Sagendorf, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, James Smith, Michael Trumbull, Brinda Vallat, Maria Voigt, Ben Webb, Shamara Whetstone, Amy Wu-Wu, Tongji Xing, Jasmine Y Young, Arthur Zalevsky, and Christine Zardecki. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Research*, 53(D1): D564–D574, 2025. doi: 10.1093/nar/gkae1091. URL https://doi.org/10.1093/nar/gkae1091.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37 th International Conference on Machine Learning*. arXiv, 2020. doi: 10.48550/arXiv.2002.05709. URL http://arxiv.org/abs/2002.05709.

Irene de Teresa-Trueba, Sara K. Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E. Zimmerli, Mauricio Toro-Nahuelpan, Dorothy W. C. Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, Alba Diz-Muñoz, Anna Kreshuk, Julia Mahamid, and Judith B. Zaugg. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, 2023. doi: 10.1038/s41592-022-01746-2. URL https://www.nature.com/articles/s41592-022-01746-2.

Neel Dey, Benjamin Billot, Hallee E. Wong, Clinton Wang, Mengwei Ren, Ellen Grant, Adrian V Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xOmC5LiVuN.

Utz Ermel, Anchi Cheng, Jun Xi Ni, Jessica Gadling, Manasa Venkatakrishnan, Kira Evans, Jeremy Asuncion, Andrew Sweet, Janeece Pourroy, Zun Shi Wang, Kandarp Khandwala, Benjamin Nelson, Dannielle McCarthy, Eric M. Wang, Richa Agarwal, and Bridget Carragher. A data portal for providing standardized annotations for cryo-electron tomography. *Nature Methods*, 21(12): 2200–2202, 2024. doi: 10.1038/s41592-024-02477-2. URL https://www.nature.com/articles/s41592-024-02477-2.

Marjan Famili, Jola Mirecka, Camila Rangel Smith, Anna Kostanska, Nikolai Juraschko, Beatriz Costa-Gomes, Colin M. Palmer, Jeyan Thiyagalingam, Tom Burnley, Mark Basham, and Alan R. Lowe. Affinity-VAE: Incorporating Prior Knowledge in Representation Learning from Scientific Images. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi (eds.), *Computer Vision – ECCV 2024 Workshops*, pp. 189–206. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-91721-9_12.

Timothy Grant and Nikolaus Grigorieff. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife*, 4:e06980, 2015. doi: 10.7554/eLife.06980. URL https://doi.org/10.7554/eLife.06980.

Mohamad Harastani, Gurudatt Patra, Charles Kervrann, and Mikhail Eltsov. Template Learning: Deep learning with domain randomization for particle picking in cryo-electron tomography. *Nature Communications*, 16(1):8833, 2025. doi: 10.1038/s41467-025-63895-0. URL https://www.nature.com/articles/s41467-025-63895-0.

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. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://ieeexplore.ieee.org/document/9879206.

Siyu He, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos. Learning to predict the cosmological structure formation. *Proceedings of the National Academy of Sciences*, 116(28):13825–13832, 2019. doi: 10.1073/pnas.1821458116. URL https://www.pnas.org/doi/10.1073/pnas.1821458116.

Qinwen Huang, Ye Zhou, and Alberto Bartesaghi. MiLoPYP: self-supervised molecular pattern mining and particle localization in situ. *Nature Methods*, pp. 1–10, 2024. doi: 10.1038/s41592-024-02403-6. URL https://www.nature.com/articles/s41592-024-02403-6.

Ron Kelley, Sagar Khavnekar, Ricardo D. Righetto, Jessica Heebner, Martin Obr, Xianjun Zhang, Saikat Chakraborty, Grigory Tagiltsev, Alicia K. Michael, Sofie van Dorst, Florent Waltz, Caitlyn L. McCafferty, Lorenz Lamm, Simon Zufferey, Philippe Van der Stappen, Hugo van den Hoek, Wojciech Wietrzynski, Pavol Harar, William Wan, John A. G. Briggs, Jürgen M. Plitzko, Benjamin D. Engel, and Abhay Kotecha. Toward community-driven visual proteomics with large-scale cryo-electron tomography of Chlamydomonas reinhardtii. *Molecular Cell*, 86(1): 213–230.e7, 2026. doi: 10.1016/j.molcel.2025.11.029. URL https://www.cell.com/molecular-cell/abstract/S1097-2765(25)00970-0.

Sagar Khavnekar, Ron Kelley, Florent Waltz, Wojciech Wietrzynski, Xianjun Zhang, Martin Obr, Grigory Tagiltsev, Florian Beck, William Wan, John Briggs, Ben Engel, Juergen Plitzko, and Abhay Kotecha. Towards the Visual Proteomics of C. reinhardtii using High-throughput Collaborative in situ Cryo-ET. *Microscopy and Microanalysis*, 29(Supplement_1):961–963, 2023. ISSN 1431-9276. doi: 10.1093/micmic/ozad067.480. URL https://doi.org/10.1093/micmic/ozad067.480.

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. URL https://www.sciencedirect.com/science/article/pii/S1047847796900131.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels, 2026. URL http://arxiv.org/abs/2603.19312.

Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D. Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, Wolfgang Baumeister, Tingying Peng, Benjamin D. Engel, and Charles Kervrann. Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms. *Nature Methods*, 18(11):1386–1394, 2021. doi: 10.1038/s41592-021-01275-4. URL https://www.nature.com/articles/s41592-021-01275-4.

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

James M. Parkhurst, Maud Dumoux, Mark Basham, Daniel Clare, C. Alistair Siebert, Trond Varslot, Angus Kirkland, James H. Naismith, and Gwyndaf Evans. Parakeet: a digital twin software pipeline to assess the impact of experimental parameters on tomographic reconstructions for cryo-electron tomography. *Open Biology*, 11(10):210160, 2021. doi: 10.1098/rsob.210160. URL https://doi.org/10.1098/rsob.210160.

Ariana Peck, Yue Yu, Jonathan Schwartz, Anchi Cheng, Utz Heinrich Ermel, Joshua Hutchings, Saugat Kandel, Dari Kimanius, Elizabeth A. Montabana, Daniel Serwas, Hannah Siems, Feng Wang, Zhuowen Zhao, Shawn Zheng, Matthias Haury, David A. Agard, Clinton S. Potter, Bridget Carragher, Kyle Harrington, and Mohammadreza Paraan. A realistic phantom dataset for benchmarking cryo-ET data annotation. *Nature Methods*, pp. 1–5, 2025. doi: 10.1038/s41592-025-02800-5. URL https://www.nature.com/articles/s41592-025-02800-5.

Ariana Peck, Joshua Hutchings, Jonathan Schwartz, Yue Yu, Utz H. Ermel, Saugat Kandel, Dari Kimanius, Zhuowen Zhao, Shawn Zheng, Brendan Artley, David List, Sergio A. Silva, Walter Reade, Jeremy Asuncion, Kira Evans, Jessica Gadling, Kandarp Khandwala, Suzette McCanny, Dannielle G. McCarthy, Jun Xi Ni, Janeece Pourroy, Manasa Venkatakrishnan, Zun Shi Wang, David A. Agard, Clinton S. Potter, Bridget Carragher, Kyle I. S. Harrington, and Mohammadreza Paraan. Lessons learned from a Kaggle challenge for particle picking in cryo-electron tomography. *Nature Methods*, 2026a. doi: 10.1038/s41592-026-03198-4. URL https://www.nature.com/articles/s41592-026-03198-4.

Ariana Peck, Yue Yu, Mohammadreza Paraan, Dari Kimanius, Utz H. Ermel, Joshua Hutchings, Jonathan Schwartz, Daniel Serwas, Hannah Siems, Norbert S. Hill, Mallak Ali, Julia Peukes, Garrett A. Greenan, Shu-Hsien Sheu, Elizabeth A. Montabana, Bridget Carragher, Clinton S. Potter, David A. Agard, and Shawn Zheng. AreTomoLive: automated reconstruction of comprehensively corrected and denoised cryo-electron tomograms in real time and at high throughput. *Nature Methods*, 23(6):1121–1125, 2026b. doi: 10.1038/s41592-026-03093-y. URL https://www.nature.com/articles/s41592-026-03093-y.

Helen Qu, Rudy Morel, Michael McCabe, Alberto Bietti, François Lanusse, Shirley Ho, and Yann LeCun. Representation Learning for Spatiotemporal Physical Systems. In *ICLR 2026 Workshop on AI & PDE*, 2026. doi: 10.48550/arXiv.2603.13227. URL http://arxiv.org/abs/2603.13227.

Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining. *Nature Methods*, 20(6):871–880, 2023. doi: 10.1038/s41592-023-01878-z. URL https://www.nature.com/articles/s41592-023-01878-z.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://ieeexplore.ieee.org/document/9878449.

Jonathan Schwartz, Utz Heinrich Ermel, C. Braxton Owens, Zhuowen Zhao, Ariana Peck, Gus L. W. Hart, Grant J. Jensen, Bridget Carragher, and Dari Kimanius. POPSICLE: Benchmark Datasets for Segmentation and Localization in CryoET, 2026. URL http://arxiv.org/abs/2606.10255.

Mart G. F. So-Last, Alister Burt, Thomas Hale, and Matteo Allegretti. Easymode: general pretrained networks for cellular cryo-ET enable flexible approaches to subtomogram averaging, 2026. URL https://www.biorxiv.org/content/10.64898/2026.05.19.726344v1.

Frosina Stojanovska, Ricardo M. Sanchez, Rasmus K. Jensen, Julia Mahamid, Anna Kreshuk, and Judith B. Zaugg. CryoSiam: self-supervised representation learning for automated analysis of cryo-electron tomograms, 2025. URL https://www.biorxiv.org/content/10.1101/2025.11.11.687379v2.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133. URL https://ieeexplore.ieee.org/abstract/document/8202133.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608009001130.

Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, Dennis Quentin, Daniel Roderer, Sebastian Tacke, Birte Siebolds, Evelyn Schubert, Tanvir R. Shaikh, Pascal Lill, Christos Gatsogiannis, and Stefan Raunser. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Communications Biology*, 2(1):1–13, 2019. doi: 10.1038/s42003-019-0437-z. URL https://www.nature.com/articles/s42003-019-0437-z.

Simon Wiedemann, Zalan Fabian, Mahdi Soltanolkotabi, and Reinhard Heckel. ProPicker: Promptable Segmentation for Particle Picking in Cryogenic Electron Tomography. *Journal of Structural Biology*, 218:108298, 2026. doi: 10.1101/2025.02.27.640512. URL http://biorxiv.org/lookup/doi/10.1101/2025.02.27.640512.

A Model architecture and training

The encoder E is based on the latent diffusion encoder (Rombach et al., 2022) adapted to 3D with two downsampling steps, and consists of ResNet blocks and an attention block at the final level before the output, followed by a 2x upsampling block consisting of interpolation, convolution and ResNet blocks and a skip connection from the layer before the second downsampling, resulting in spatial features that are downsampled by a factor of 2 with respect to the input volume resolution. E maps 64^3 input volumes x extracted from a tomogram to 64×32^3 dense spatial feature volumes $y = E(x)$ with 64 channels. The readout network F consists of a convolutional block and an MLP block. F is only used in training, and it maps a 64×16^3 volume, cropped from the spatial feature volume and centred at protein position, to an embedding $z \in \mathbb{R}^{512}$ that the loss function in Section 3.2 consumes. An optional segmentation head, consisting of a upsampling, ResNet blocks and a skip connection from the layer before the first downsampling in the encoder, maps the spatial feature volume $y = E(x)$ to a segmentation mask 2×64^3 at the original input volume resolution, with two channels: one for proteins and one for membranes. The segmentation loss $\mathcal{L}_{\text{seg}}$ uses ground-truth masks obtained by thresholding the simulated proteins and membranes. The full model

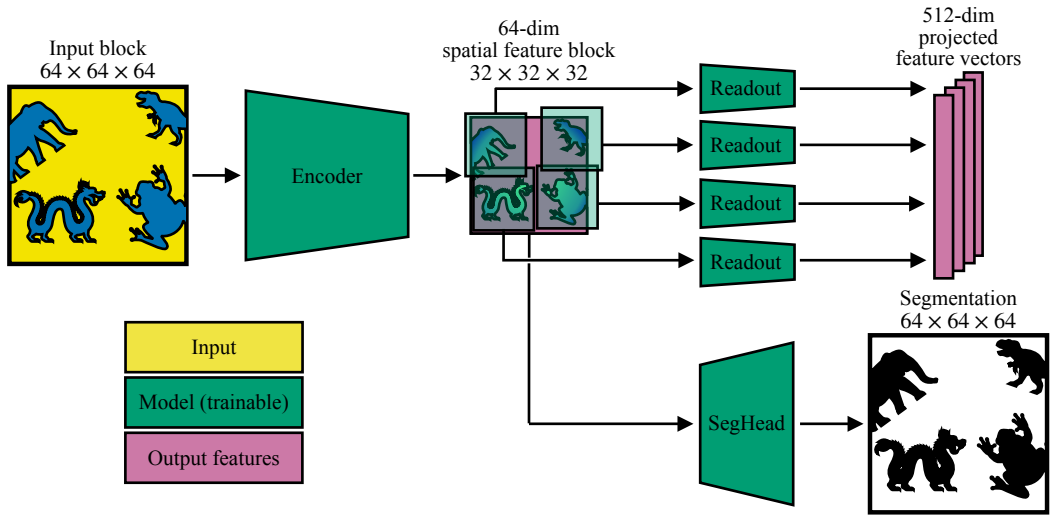

Figure 3: Model architecture.

has 9,708,706 parameters and all the components are trained jointly. A schematic of the architecture is shown in Figure 3.

The model was trained from scratch, with the full set of transformations in Table 5 throughout, using the `AdamW` optimiser in PyTorch with learning rate of `5e-5` and `weight_decay=5e-4`, `betas = (0.9, 0.999)` and `eps = 1e-8`, with 1000 warmup steps and cosine annealing to a minimum learning rate of `5e-8` on a schedule set for 100,000 steps. The run was stopped at step 80,250 and followed by an annealing run of 4,000 steps from that checkpoint, in which the learning rate is warmed up over 1,000 steps to 4.8×10^{-6}, the value the main schedule had reached, and then decays to `5e-8` along a cosine. The final checkpoint of the annealing run, after 84,000 steps in total, is used for all results in this paper.

The model sees six different versions of each batch sample (a 64^3 volume): first, one clean scene containing proteins and membranes ("big structures", Appendix B) is sampled every 50 batches (25 optimiser steps), which is used to generate two distinct clean volumes with the same big structures but different crowded background. A third clean volume is generated with the same protein identities and positions but in different orientations. For each batch, the first two clean volumes are corrupted with independent parameter draws from Table 5 and the third with the parameters of the first (see Section 3.1 and Appendix B), and the batch elements are extracted from the same positions in each of the resulting six volumes (with a small jitter of up to ± 9 voxels for each view): three clean views and three corrupted views of the same scene. Each GPU processes 8 different positions per batch, and gradients are accumulated over two batches before each optimiser step; with six GPUs, the effective batch size is 96, with six views per batch element. Training was done on nodes with $6\times$ NVIDIA GeForce RTX 5090 GPUs with 32 GB of memory each, and took about 25 days for the first 80,250 steps and under two days for the annealing run.

The ablation arms of Section 4.3 use the same recipe, trained from scratch for 10,000 steps on a cosine schedule set for 10,000 steps. RP and A6P were trained on the same type of node as CARNIVAL. The other arms were trained on a national HPC cluster, on nodes with $4\times$ NVIDIA GH200 Grace Hopper Superchips with 96 GB of GPU memory each, with 12 positions per GPU and two accumulation steps, so that the effective batch size is also 96.

B Training data

While a future goal is to train CARNIVAL using most proteins in the RCSB Protein Data Bank (Burley et al., 2025) and the AlphaFold Protein Structure Database (Bertoni et al., 2026), the model described in this paper has been trained on the 112 PDB IDs from TomoTwin's training set larger than 100kDa (Rice et al., 2023), and downloaded from the Protein Data Bank.

First, for each protein in the list above, the atomic model was downloaded from the Protein Data Bank, which was then used to compute a volume containing the protein's density by placing Gaussians at the atom coordinates in a 256^3 box with 2Å/voxel, low-pass filtered to 20Å resolution and rescaled to a 64^3 box with 10Å/voxel. We refer to these as the *foreground proteins*. An additional set of 1000 32^3 volumes containing small proteins of up to 30kDa, whose atomic models were randomly selected from the PDB, are computed using the same steps. These are used to generate the random crowded environment of the tomograms, so we will refer to these as the *background proteins*.

Every 50 batches (25 optimiser steps), foreground proteins sampled from the volumes above are inserted in a slab of random thickness between 120 and 220 voxels, within a box of dimensions $350 \times 350 \times 350$, at random positions sampled using Poisson disk sampling, together with a simulated membrane. The membrane is obtained by tiling a simulated membrane patch along a curved surface with parameters randomly drawn. Using this configuration of foreground structures, three clean views that correspond to the same scene are created: in the first two, the foreground proteins are in the same (random) orientations, but are surrounded by different random background protein configurations, also sampled using Poisson disk sampling, using a smaller radius and taking into account the foreground proteins and the membrane to avoid overlaps. The third clean view contains the same foreground proteins in different orientations, and the same background as the first clean view. For each batch, two corrupted tomograms are computed using the first two clean views after sampling the parameters for each transformation in Table 5, and a third corrupted tomogram is obtained from the third clean view using the transformation parameters of the first view. This results in six total views at 10Å/voxel: three clean and three corrupted. Finally, `batch_size` volumes are extracted from each of the six views from the same positions across views, with a small random jitter in each view. Each GPU performs this independently using disjoint sets of `PDB_IDs`. We ensure that each volume contains at least one protein, and also return a list of protein coordinates and identities present in each volume and the corresponding foreground protein and membrane masks.

C Full list of transformations in the measurement operator

All transformations used to generate the physically-grounded tomogram augmentations for training are given in Table 5. These are used as described in Section 3.1, Appendix A and Appendix B.

D Experiments: further details

All datasets were downloaded from the CryoET Data Portal (Ermel et al., 2024), and every tomogram is used at a voxel spacing of ~ 10 Å, the spacing at which the model was trained. No normalisation or finetuning is applied, and the same CARNIVAL checkpoint (Appendix A) is used for the classification and detection experiments.

Phantom dataset We use the public test set from the CZII CryoET Object Identification challenge (Peck et al., 2026a) (DS-10445, with the challenge ground truth from deposition CZCDP-10310), which provides each tomogram in two processing forms that we use as deposited: a band-pass filtered reconstruction, and one that has additionally been denoised with DenoisET (Peck et al., 2026b). The tomograms are stored on the portal at 10.012 Å and are used without resampling. As in the challenge, β-amylase is not evaluated: the organisers did not score it because of low confidence in the quality of its annotations (Peck et al., 2026a).

***S. pombe* datasets** DS-10000 (defocus contrast) and DS-10001 (Volta phase plate) (de Teresa-Trueba et al., 2023) contain 10 tomograms each, with ribosome and fatty acid synthase annotations from deposition CZCDP-10000. The annotations were made with the convolutional network of de Teresa-Trueba et al. (2023), trained on template-matching results, followed by visual filtering, distance constraints and manual addition; the portal reports no precision or recall for them. The raw reconstructions, stored on the portal at 13.48 Å, are resampled to 10 Å by zero-padding in Fourier space, and the annotation coordinates are scaled accordingly.

***C. reinhardtii* dataset** From DS-10301 (Khavnekar et al., 2023) we use three tomograms (TM-14317, TM-14327 and TM-14331), chosen because they have ground-truth annotations of two different particle types, ribosomes and nucleosomes (deposition CZCDP-10300). The annotations were

Table 5: The sampled operators in the forward model $\Psi_\theta = R_{\theta_R} \circ M_{\theta_M}$, grouped by the two blocks of the decomposition. Every simulated tomogram draws its own values (one draw per tomogram, shared by all `batch_size` blocks extracted from it) and each range is chosen to bracket experimental practice rather than to match any one dataset. Held fixed: 10 Å/voxel, 300 kV, $C_s = 2.7$ mm, amplitude contrast 0.1, and radiation damage, which is physics rather than a processing choice and so is never switched off. The CTF correction-state randomises a processing choice rather than a physical one, as different reconstruction algorithms differ in their CTF treatment. The notation used in each row is independent from the rest of the paper.

Transformation	Form	Sampled from
Measurement operator M_{θ_M}		
Bulk medium (ice / cytoplasm)	$v \mapsto v + \phi\,(v_{\max} - v_{\min})\,d(v)$ on the volume before projection, with $d(v) = \mathrm{clamp}(1 - v/v_{\max},\, 0,\, 1)$ so structure displaces medium; fill level $\phi = \alpha$	$\alpha \in [0, 0.45]$
Large-scale medium variation	ϕ generalised to a smooth field $\phi(x) = \alpha\,(1 + r\,z(x))$ of the same mean, z zero-mean unit-std low-passed by $e^{-2\pi^2\sigma^2 f^2}$, $\sigma = \ell$/pixel size	$r \in [0, 0.5]$ $\ell \in [400, 3200]$ Å
Tilt scheme (missing wedge)	projections at $\theta \in \{-\theta_{\max}, \ldots, \theta_{\max}\}$ in steps of $\Delta\theta$; the wedge follows from $\theta_{\max}$	$\theta_{\max} \in \{45, 50, 55, 60\}^\circ$ $\Delta\theta \in \{2, 3, 5\}^\circ$
Defocus (CTF)	$\hat{I} \mapsto C(k; \Delta f_\theta)\,\hat{I}$ per tilt, with Δf_θ varying with depth along the beam	$\Delta f \in [1, 5]\,\mu$m
CTF envelope	$E(k) = \exp(-Bk^2/4)$ from partial coherence, tilt-independent, multiplied into the CTF	$B \in [0, 300]$ Å^2
Per-tilt contrast decay	tilt θ scaled by $\exp\big(-C\,(\sec\theta - 1)\big)$, the Beer–Lambert fan-out; after CTF and before noise, so the noise floor stays tilt-independent	$C \in [0, 1.2]$
Radiation damage	i-th tilt image scaled by $g_i(f) = \exp\big(-D_i/(2N_c(f))\big)$ before noise, with $N_c(f) = 0.245 f^{-1.665} + 2.81$ and D_i accumulated dose by tilt i under dose-symmetric ordering (Grant & Grigorieff, 2015)	$D_{\max} \in [80, 120]$ e/Å^2
Noise level	Gaussian noise added to every tilt with standard deviation $\sigma\,(I_0^{\max} - I_0^{\min})$, referenced to the contrast range of the zero-tilt frame	$\sigma \in [0, 0.25]$
Noise colour	that noise weighted by $w(f) \propto (f + \varepsilon)^{-\gamma/2}$ renormalised to $\langle w^2 \rangle = 1$ so total variance is unchanged	$\gamma \in [0, 1.2]$
Reconstruction operator R_{θ_R}		
CTF correction state	transfer applied $\in \{1,\ CE,\ |C|E\}$: amplitude-restored, uncorrected, phase-flipped	uniform over the three
Dose weighting	the same $g_i(f)$ re-applied inside backprojection, now to signal + noise, as the matched filter that merges the frames; otherwise plain WBP	applied with probability 0.5
Backprojection regularisation	Fourier-space division $1/(W(k) + \lambda)$, with W the accumulated tilt weights	$\lambda \in \{1, 2, 5, 10\}$

made by template matching followed by manual filtering and 3D classification, so they contain the particles retained by the classification rather than every particle in the volume. The denoised tomograms, stored on the portal at 7.84 Å, are resampled to 10 Å by cropping in Fourier space, and the annotation coordinates are scaled accordingly.

Metrics Both experiments report an F_β score:

$$F_\beta = (1+\beta^2)\,\frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}},$$

computed per protein type and then averaged over types. Precision and Recall are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives (i.e. missed particles).

The classification experiment evaluates embeddings at ground-truth positions, so every particle receives exactly one predicted label. In this case, precision and recall are symmetric and we use the standard F_1, their harmonic mean. Detection has no such symmetry. The operating point is a distance threshold, and lowering it trades false positives (which a later classification or averaging step can still reject) against particles that are missed outright and cannot be recovered downstream. We therefore follow the Object Identification challenge (Peck et al., 2026a) and report F_4, which weights recall $\beta^2 = 16$ times more heavily than precision, both for that reason and so that our numbers are comparable to the challenge's.

Baseline and hardware For the comparison with TomoTwin (Rice et al., 2023), we used the TomoTwin model version `09.2023` made available by the authors on Zenodo at the URL `https://zenodo.org/records/8358240`, under a Creative Commons Attribution 4.0 International licence.

All evaluations were run on two identical internal cluster nodes, each with 6× NVIDIA GeForce RTX 5090 GPUs with 32 GB of memory each, an AMD EPYC 9354 32-core processor and 512 GB of RAM; training is described in Appendix A. On one such node, computing a feature volume takes CARNIVAL about 4 minutes for a Phantom tomogram and 36–52 minutes for an *S. pombe* sub-volume, and TomoTwin about 31 minutes and 5 hours respectively.

D.1 Classification

For a protein at position p in the input tomogram, a 64^3 block around p is encoded, and the embedding is the concatenation of the feature vectors of $E(x)$ within a sphere of radius 4 feature voxels (8 tomogram voxels) around p. The concatenated vectors of all particles in a dataset are L2-normalised and reduced to 64 dimensions by PCA, fitted without labels on those same particles, and cosine distance is the metric in this space. For TomoTwin, the embedding is its 32-dimensional output for the 37^3 subvolume centred on p, also compared by cosine distance. This test asks only whether the embedding space separates protein types. The positions are given, so detection is assumed and what is measured is the geometry of the embeddings themselves. Proteins of the same type should fall in the same region of the embedding space regardless of which tomogram they came from, their orientation, or how much surrounding material is in their neighbourhood.

Within each dataset, and separately for each processing type of the Phantom dataset, the embeddings of all particles are pooled across tomograms and classified by leave-one-out k-nearest-neighbour vote with $k = 5$ and uniform weights, each particle being classified by its neighbours across the whole pooled set with itself excluded. Per-type F_1 scores and their macro average are reported in Table 6 for the Phantom dataset, the same numbers plotted in Figure 2A, and in Table 1 for the in-cell datasets.

We also ran the classification with $k = 1$, 3 and 10. On the Phantom and *C. reinhardtii* datasets, CARNIVAL leads at every k, by 0.03–0.08 and 0.16–0.20 in macro F_1 respectively. On *S. pombe*, the two models stay within 0.04 of each other and the sign of the difference changes with k: CARNIVAL leads on DS-10000 at $k = 1$ and 10, and on DS-10001 at $k = 1$, 3 and 5. There, macro

Table 6: Full classification results on the Phantom dataset with $k = 5$ nearest neighbours. The F_1 score is shown for each protein type and the overall score is averaged over protein types. The same numbers shown in Figure 2A. The standard deviation of CARNIVAL's values over its last ten checkpoints is at most 0.004; TomoTwin is one fixed model.

	Filtered F_1		Denoised F_1		# particles
	CARNIVAL	**TomoTwin**	**CARNIVAL**	**TomoTwin**	
ribosome	**0.991**	0.979	**0.991**	0.968	6,478
VLP	**0.971**	0.871	**0.961**	0.852	726
thyroglobulin	**0.793**	0.769	**0.740**	0.671	1,501
β-galactosidase	**0.756**	0.696	**0.722**	0.531	827
apoferritin	**0.988**	0.963	**0.977**	0.963	5,694
overall	**0.900**	0.856	**0.878**	0.797	15,226

Table 7: Per-type detection results in the phantom dataset. Each entry is the maximum F_4 over the threshold sweep, taken on the precision-recall curve averaged over the 30 template draws, $\pm$ one standard deviation of the individual draws' F_4 at that same threshold. These are the operating points marked in Figures 4–5. Bold marks the higher of the two values within a processing condition. The weighted-mean rows reproduce the corresponding entries of Table 2.

	Filtered		Denoised	
Protein type	**CARNIVAL**	**TomoTwin**	**CARNIVAL**	**TomoTwin**
`TS_5_2`				
ribosome	0.691 ± 0.073	$\mathbf{0.757 \pm 0.049}$	$\mathbf{0.682 \pm 0.064}$	0.567 ± 0.086
VLP	$\mathbf{0.784 \pm 0.037}$	0.438 ± 0.176	$\mathbf{0.768 \pm 0.041}$	0.395 ± 0.140
thyroglobulin	0.255 ± 0.070	$\mathbf{0.305 \pm 0.079}$	$\mathbf{0.250 \pm 0.069}$	0.240 ± 0.082
β-galactosidase	$\mathbf{0.543 \pm 0.075}$	0.205 ± 0.080	$\mathbf{0.527 \pm 0.083}$	0.225 ± 0.089
apoferritin	$\mathbf{0.814 \pm 0.030}$	0.565 ± 0.145	$\mathbf{0.760 \pm 0.052}$	0.728 ± 0.169
Weighted mean	**0.555**	0.397	**0.538**	0.374
`TS_43_9`				
ribosome	0.728 ± 0.037	$\mathbf{0.804 \pm 0.049}$	$\mathbf{0.646 \pm 0.062}$	0.615 ± 0.060
apoferritin	$\mathbf{0.879 \pm 0.021}$	0.548 ± 0.096	$\mathbf{0.750 \pm 0.065}$	0.749 ± 0.101
Weighted mean	**0.803**	0.676	**0.698**	0.682

F_1 is set by the rare FAS class (1.4% and 2.8% of the particles), and both models' macro F_1 drops at $k = 10$, consistent with a larger k outvoting the rare class. Every difference between an ablation arm and the reference that exceeds 0.05 in macro F_1 at $k = 5$ keeps its sign at $k = 1$, 3 and 10.

D.2 Detection

A feature volume is computed for each tomogram by applying the encoder E in overlapping 64^3 blocks with stride 8, keeping the central 48^3 of each block and blending the overlaps with a smoothly decaying (Hann) window, which avoids the seams a hard crop leaves behind. TomoTwin (Rice et al., 2023) was run with stride 2, as recommended in TomoTwin's documentation at `https://tomotwin-cryoet.readthedocs.io/en/stable/tutorials/tutorials_overview.html#embed-your-tomogram`. Both methods produce feature volumes at half the tomogram sampling, so the candidate grids coincide and the two are scored on identical positions. For the *S. pombe* tomograms, both methods are run on a slab in z that contains every annotated particle, with a margin of about 50 slices on either side, and the full extent in x and y: slices 94–576 of the 674 in TM-623, 132–496 of the 674 in TM-625, 496–836 of the 1,348 in TM-2269 and 442–902 of the 1,348 in TM-2272.

We evaluate detection using a workflow similar to TomoTwin's reference-based workflow (Rice et al., 2023), as follows. For each of 30 independent draws, five ground-truth particles per type are

Table 8: Per-type detection results in DS-10000 (*S. pombe*, defocus contrast, raw reconstructions). Each entry is the maximum F_4 over the threshold sweep, taken on the precision-recall curve averaged over the 30 template draws, $\pm$ one standard deviation of the individual draws' F_4 at that same threshold. Bold marks the higher of the two values. FAS: fatty acid synthase. The mean rows reproduce the corresponding entries of Table 3.

Protein type	CARNIVAL	TomoTwin
TM-623		
ribosome	0.475 ± 0.106	$\mathbf{0.637 \pm 0.089}$
FAS	$\mathbf{0.032 \pm 0.021}$	0.020 ± 0.015
Mean	0.253	**0.328**
TM-625		
ribosome	0.573 ± 0.108	$\mathbf{0.726 \pm 0.055}$
FAS	0.065 ± 0.032	$\mathbf{0.087 \pm 0.040}$
Mean	0.319	**0.407**

Table 9: Per-type detection results in DS-10001 (*S. pombe*, phase-plate contrast, raw reconstructions), as in Table 8.

Protein type	CARNIVAL	TomoTwin
TM-2269		
ribosome	$\mathbf{0.635 \pm 0.084}$	0.634 ± 0.087
FAS	0.057 ± 0.034	$\mathbf{0.081 \pm 0.037}$
Mean	0.346	**0.358**
TM-2272		
ribosome	0.603 ± 0.088	$\mathbf{0.616 \pm 0.059}$
FAS	0.050 ± 0.028	$\mathbf{0.183 \pm 0.026}$
Mean	0.327	**0.399**

Table 10: Per-type detection results in DS-10301 (*C. reinhardtii*, denoised reconstructions), as in Table 8. The annotations are a classification-filtered subset of the particles, so precision is a lower bound.

Protein type	CARNIVAL	TomoTwin
TM-14317		
ribosome	$\mathbf{0.070 \pm 0.030}$	0.045 ± 0.023
nucleosome	$\mathbf{0.114 \pm 0.017}$	0.102 ± 0.025
Mean	**0.092**	0.073
TM-14327		
ribosome	0.165 ± 0.087	$\mathbf{0.267 \pm 0.142}$
nucleosome	$\mathbf{0.283 \pm 0.029}$	0.233 ± 0.075
Mean	0.224	**0.250**
TM-14331		
ribosome	0.066 ± 0.036	$\mathbf{0.091 \pm 0.044}$
nucleosome	$\mathbf{0.214 \pm 0.015}$	0.148 ± 0.037
Mean	**0.140**	0.119

sampled as templates. For a given type, every voxel of the feature volume is scored by its cosine distance to the nearest of that type's template embeddings, and the local minima of the resulting distance volume are the detection candidates. Any ground-truth particle within 8 feature voxels of a template is removed from the test set for that draw, so neither the templates themselves nor their immediate neighbours can be scored as detections. The distance at each candidate is the threshold variable: sweeping it over 100 thresholds between the 0.001st and the 10th percentile of the distance volume traces one precision-recall curve per type per draw. The thresholds are placed at the candidates' own distances, pooled over the draws and log-spaced in rank, so each admits a roughly constant fraction more detections than the last, whatever the spread of a method's distances. A candidate counts as a true positive when it falls within 4 feature voxels (8 tomogram voxels) of a ground-truth particle of that type; matching is one-to-one and greedy, taking candidates in order of increasing distance to their nearest unmatched ground-truth particle, so the best-localised detections are matched first. Protein types with fewer than 10 instances left after template exclusion are dropped from the figures and from the weighted mean calculation in Table 2. A type below that bar still contributes its templates to the nearest-type assignment, so a scarce type cannot inflate another type's precision by going unrepresented.

For each type we report the best F_4 over the sweep (an upper bound available to neither method at deployment, but computed identically for both) and summarise a tomogram by the mean over its evaluated types. On the Phantom dataset the mean weights the hard types (β-galactosidase, thyroglobulin) by 2 and the rest by 1, the weights used to score the Object Identification challenge (Peck et al., 2026a); as in the challenge, β-amylase is not evaluated. On the in-cell datasets every type has weight 1.

Per-type precision-recall curves are shown in Figures 4–6. Figure 4A repeats the tomogram of Figure 2B beside its denoised counterpart (panel B); Figure 5 shows a tomogram dominated by one type (600 apoferritin instances against 42 ribosomes; its 5 thyroglobulin and 8 β-galactosidase instances are too few to evaluate); and Figure 6 shows the in-cell tomograms. Tables 7–10 give the maximum F_4 for each protein type, taken on the draw-averaged precision-recall curve (the values marked in Figures 4–6), each with the standard deviation across the 30 template draws at that threshold.

The Phantom detection tomograms are deposited in the CryoET Data Portal under IDs TM-19098 and TM-19099 (denoised and filtered `TS_43_9`) and TM-19158 and TM-19159 (`TS_5_2`), all in dataset DS-10445, the same public test set used for classification. The in-cell detection tomograms are TM-623 and TM-625 (runs `TS_028` and `TS_030` of DS-10000), TM-2269 and TM-2272 (`TS_0001` and `TS_0004` of DS-10001), and the three DS-10301 tomograms. Slices of the two Phantom detection tomograms are shown in Figure 7.

E Ablations: full classification results

Tables 11–16 give the classification results at $k = 5$ of CARNIVAL, TomoTwin and every ablation arm of Section 4.3, computed as in Appendix D.1: macro F_1 in each evaluation setting, then F_1 per protein type in each setting.

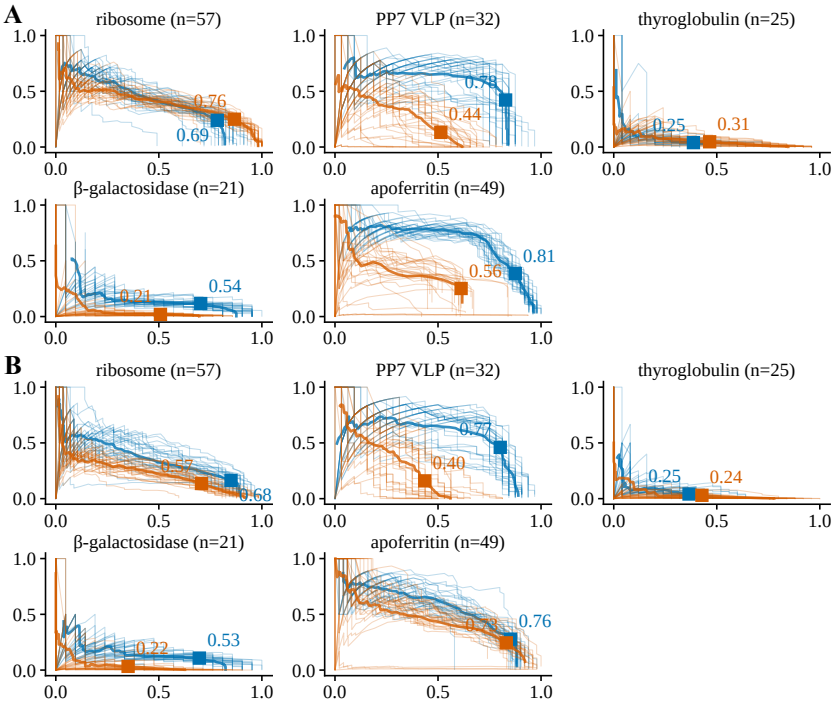

Figure 4: **CARNIVAL** vs **TomoTwin**: Detection results, tomogram `TS_5_2`. **A** Filtered. This panel is identical to Figure 2B, reproduced here for easy visual comparison between the filtered and denoised tomograms. **B** Denoised.

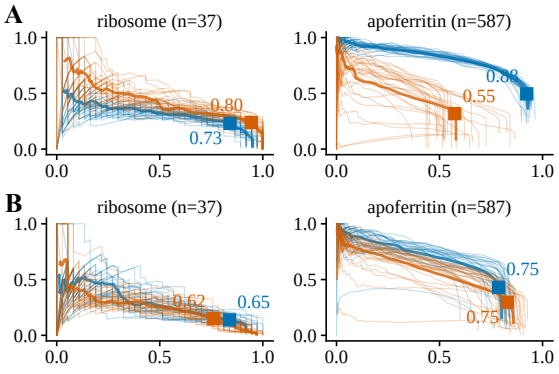

Figure 5: **CARNIVAL** vs **TomoTwin**: Detection results, tomogram `TS_43_9`. **A** Filtered. **B** Denoised.

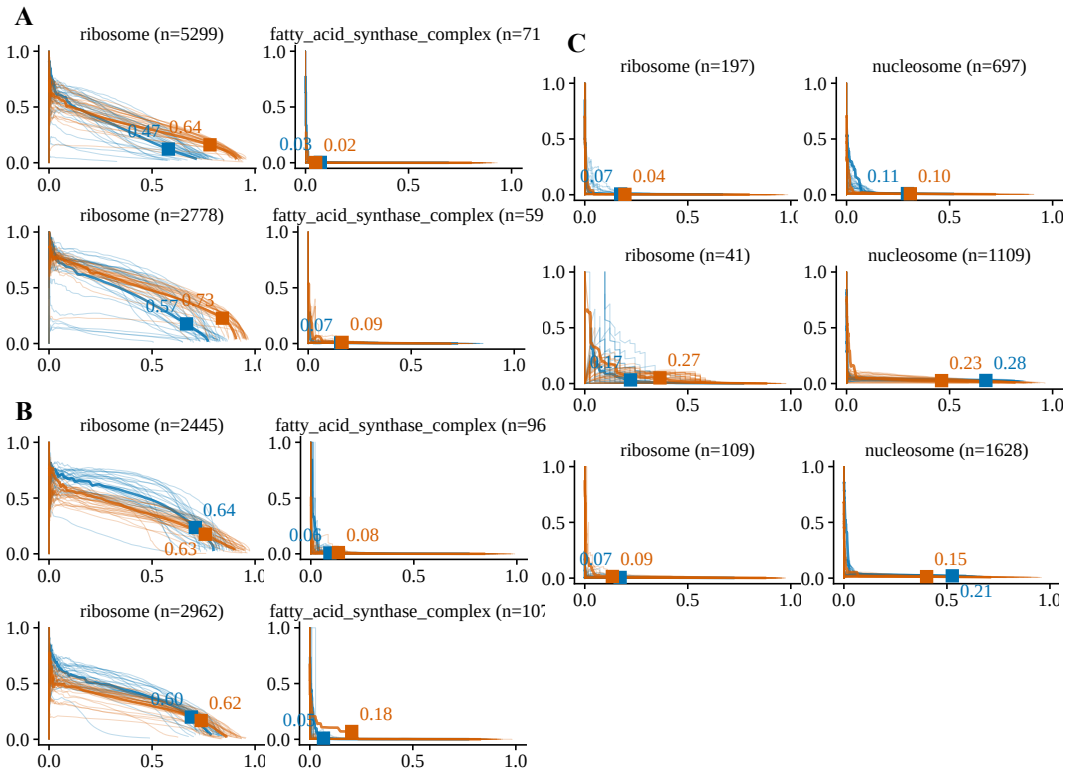

Figure 6: **CARNIVAL** vs **TomoTwin**: Detection results, in-cell tomograms. Each row in each panel is a tomogram. **A** DS-10000 (*S. pombe*, defocus): TM-623 (top) and TM-625 (bottom). **B** DS-10001 (*S. pombe*, Volta phase plate): TM-2269 (top) and TM-2272 (bottom). **C** DS-10301 (*C. reinhardtii*): TM-14317, TM-14327 and TM-14331, from top to bottom.

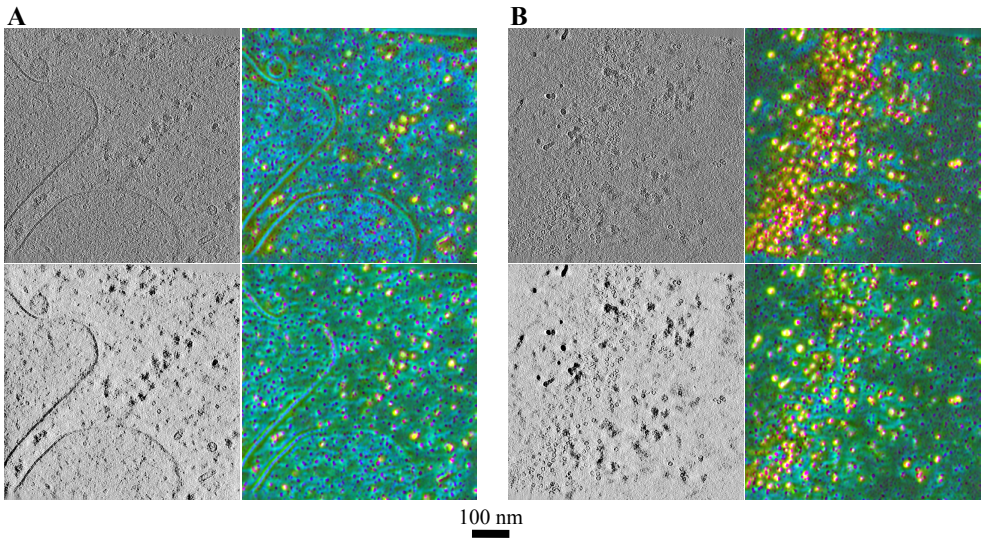

Figure 7: Slices of tomograms and spatial feature volumes from CARNIVAL, with the filtered version on the top row and the denoised version on the bottom row in each panel. As the spatial feature blocks have 64 channels, we show the 3 largest principal components of each feature vector, mapped to RGB, where the principal components were calculated using all feature volumes at the same time. Scale bar 100 nm is common to all panels. **A** Tomogram `TS_5_2`. **B** Tomogram `TS_43_9`.

Table 11: Classification macro F_1 at $k = 5$ for CARNIVAL, TomoTwin and the ablation arms of Table 4. Each value is taken at the last checkpoint, $\pm$ the standard deviation over the last ten checkpoints (steps 9,100–10,000 for the arms and 3,100–4,000 of the annealing run for CARNIVAL): how much the value depends on where training stops, not variation across seeds. TomoTwin is one fixed model. n is the number of particles.

Model n	Phantom filtered 15226	Phantom denoised 15226	DS-10000 26267	DS-10001 26042	DS-10301 3825
CARNIVAL	0.900 ± 0.001	0.878 ± 0.001	0.624 ± 0.006	0.667 ± 0.003	0.819 ± 0.004
TomoTwin	0.856	0.797	0.637	0.641	0.636
R	0.878 ± 0.002	0.875 ± 0.001	0.592 ± 0.004	0.681 ± 0.003	0.765 ± 0.007
A1	0.863 ± 0.001	0.873 ± 0.002	0.710 ± 0.004	0.805 ± 0.002	0.845 ± 0.003
A2	0.858 ± 0.001	0.863 ± 0.003	0.565 ± 0.003	0.751 ± 0.003	0.739 ± 0.004
A3	0.734 ± 0.004	0.795 ± 0.002	0.496 ± 0.002	0.522 ± 0.003	0.756 ± 0.011
A4	0.876 ± 0.001	0.868 ± 0.001	0.589 ± 0.005	0.649 ± 0.003	0.795 ± 0.006
A5	0.799 ± 0.003	0.796 ± 0.004	0.557 ± 0.009	0.696 ± 0.008	0.845 ± 0.002
A6	0.823 ± 0.002	0.836 ± 0.002	0.596 ± 0.007	0.718 ± 0.003	0.658 ± 0.011
RP	0.876 ± 0.002	0.873 ± 0.001	0.575 ± 0.006	0.648 ± 0.005	0.830 ± 0.004
A6P	0.858 ± 0.001	0.865 ± 0.002	0.624 ± 0.008	0.715 ± 0.005	0.761 ± 0.003

Table 12: Classification F_1 per protein type at $k = 5$ on the filtered tomograms of the Phantom dataset, with values and $\pm$ as in Table 11.

Model n	ribosome 6478	PP7 VLP 726	thyroglobulin 1501	β-galactosidase 827	apoferritin 5694
CARNIVAL	0.991 ± 0.000	0.971 ± 0.001	0.793 ± 0.002	0.756 ± 0.002	0.988 ± 0.000
TomoTwin	0.979	0.871	0.769	0.696	0.963
R	0.985 ± 0.001	0.959 ± 0.002	0.725 ± 0.004	0.752 ± 0.004	0.968 ± 0.001
A1	0.980 ± 0.001	0.952 ± 0.002	0.719 ± 0.003	0.694 ± 0.003	0.971 ± 0.000
A2	0.978 ± 0.001	0.943 ± 0.002	0.694 ± 0.003	0.715 ± 0.002	0.958 ± 0.001
A3	0.910 ± 0.002	0.898 ± 0.006	0.456 ± 0.008	0.549 ± 0.009	0.857 ± 0.003
A4	0.982 ± 0.000	0.955 ± 0.001	0.725 ± 0.004	0.751 ± 0.003	0.969 ± 0.001
A5	0.973 ± 0.002	0.932 ± 0.004	0.587 ± 0.007	0.566 ± 0.008	0.935 ± 0.003
A6	0.972 ± 0.001	0.933 ± 0.002	0.627 ± 0.004	0.619 ± 0.005	0.965 ± 0.001
RP	0.983 ± 0.000	0.958 ± 0.001	0.727 ± 0.006	0.747 ± 0.002	0.966 ± 0.001
A6P	0.981 ± 0.000	0.944 ± 0.001	0.695 ± 0.005	0.697 ± 0.004	0.971 ± 0.001

Table 13: Classification F_1 per protein type at $k = 5$ on the denoised tomograms of the Phantom dataset, as in Table 12.

Model n	ribosome 6478	PP7 VLP 726	thyroglobulin 1501	β-galactosidase 827	apoferritin 5694
CARNIVAL	0.991 ± 0.000	0.961 ± 0.000	0.740 ± 0.002	0.722 ± 0.004	0.977 ± 0.000
TomoTwin	0.968	0.852	0.671	0.531	0.963
R	0.988 ± 0.000	0.960 ± 0.002	0.733 ± 0.002	0.723 ± 0.003	0.969 ± 0.001
A1	0.984 ± 0.000	0.946 ± 0.002	0.760 ± 0.003	0.705 ± 0.004	0.972 ± 0.000
A2	0.987 ± 0.000	0.948 ± 0.003	0.719 ± 0.005	0.707 ± 0.007	0.955 ± 0.000
A3	0.963 ± 0.001	0.928 ± 0.003	0.594 ± 0.006	0.567 ± 0.003	0.921 ± 0.001
A4	0.984 ± 0.001	0.953 ± 0.002	0.716 ± 0.001	0.718 ± 0.002	0.970 ± 0.000
A5	0.969 ± 0.001	0.926 ± 0.001	0.610 ± 0.008	0.545 ± 0.010	0.928 ± 0.003
A6	0.982 ± 0.001	0.955 ± 0.001	0.653 ± 0.006	0.627 ± 0.005	0.964 ± 0.001
RP	0.984 ± 0.000	0.954 ± 0.002	0.725 ± 0.003	0.736 ± 0.003	0.966 ± 0.000
A6P	0.987 ± 0.000	0.948 ± 0.001	0.728 ± 0.004	0.693 ± 0.004	0.971 ± 0.000

Table 14: Classification F_1 per protein type at $k = 5$ on DS-10000 (*S. pombe*, defocus contrast, raw reconstructions), as in Table 12 (FAS: fatty acid synthase).

Model	ribosome	FAS
n	25901	366
CARNIVAL	0.994 ± 0.000	0.253 ± 0.012
TomoTwin	0.993	0.280
R	0.993 ± 0.000	0.191 ± 0.007
A1	0.994 ± 0.000	0.426 ± 0.009
A2	0.993 ± 0.000	0.137 ± 0.006
A3	0.993 ± 0.000	0.000 ± 0.004
A4	0.993 ± 0.000	0.185 ± 0.010
A5	0.993 ± 0.000	0.121 ± 0.018
A6	0.993 ± 0.000	0.198 ± 0.014
RP	0.993 ± 0.000	0.158 ± 0.013
A6P	0.994 ± 0.000	0.255 ± 0.016

Table 15: Classification F_1 per protein type at $k = 5$ on DS-10001 (*S. pombe*, phase-plate contrast, raw reconstructions), as in Table 12.

Model	ribosome	FAS
n	25311	731
CARNIVAL	0.988 ± 0.000	0.345 ± 0.005
TomoTwin	0.987	0.296
R	0.988 ± 0.000	0.375 ± 0.007
A1	0.991 ± 0.000	0.619 ± 0.005
A2	0.989 ± 0.000	0.513 ± 0.005
A3	0.985 ± 0.000	0.058 ± 0.007
A4	0.988 ± 0.000	0.311 ± 0.005
A5	0.988 ± 0.000	0.404 ± 0.015
A6	0.988 ± 0.000	0.448 ± 0.006
RP	0.988 ± 0.000	0.308 ± 0.010
A6P	0.988 ± 0.000	0.443 ± 0.010

Table 16: Classification F_1 per protein type at $k = 5$ on DS-10301 (*C. reinhardtii*, denoised reconstructions), as in Table 12.

Model	ribosome	nucleosome
n	362	3463
CARNIVAL	0.666 ± 0.007	0.972 ± 0.001
TomoTwin	0.321	0.952
R	0.567 ± 0.013	0.963 ± 0.001
A1	0.718 ± 0.005	0.971 ± 0.000
A2	0.518 ± 0.007	0.960 ± 0.001
A3	0.548 ± 0.020	0.964 ± 0.001
A4	0.625 ± 0.011	0.965 ± 0.001
A5	0.717 ± 0.004	0.973 ± 0.000
A6	0.360 ± 0.021	0.956 ± 0.001
RP	0.690 ± 0.007	0.970 ± 0.001
A6P	0.557 ± 0.006	0.965 ± 0.000