# Exploring Simulators for Particle Picking in Cryo-Electron Tomography

**Serena M. Arghittu**[1,2*] **Lars Dingeldein**[2,3] **Geoffrey Woollard**[4] **LingLi Kong**[5]

**Magnus Petersen**[2] **Sonya Hanson**[6], **Roberto Covino**[2,7] **Pilar Cossio**[6,8†]

## Abstract

To understand how proteins function, we need to know the conformations that they adopt and with what they interact in their native cellular environment. Cryo-electron tomography (cryo-ET) offers a powerful tool by enabling *in situ* imaging of proteins. But high noise levels and the need for expertise in particle identification limit its scalability. In this study, we present a machine learning framework for automated recognition and localization of particles in cryo-ET data. We treat particle picking as an object recognition task and employ a U-Net-based architecture for multi-class segmentation. To overcome the scarcity of annotated data, we train our model on synthetic tomograms generated by a simulator that incorporates empirical noise from publicly available cryo-ET datasets. Our results show that training on a mixed dataset containing both synthetic and empirical backgrounds provides the most effective particle-picking performance, enhancing the model's robustness to different background types. Furthermore, we demonstrate that training exclusively on simulated particles enables the model to reliably distinguish particles from background in real tomograms, highlighting the potential of simulation-based training strategies in cryo-ET.

## 1 Introduction

Biomolecules function within cells forming complexes with other biomolecules in different environments characterized by different physio-chemical conditions. To understand their roles, it is essential to identify their spatial localization, structural arrangements, and binding partners *in situ*. Cryo-electron tomography (cryo-ET) images thin slices of frozen biological specimens, offering detailed tomograms and revealing diverse cellular architectures.

Cryo-ET involves the extraction of one or more thin layers of a vitrified specimen that is then imaged by an electron microscope. The sample is tilted within the microscope to obtain multiple 2D projections at different angles. The obtained *tilt-series* of images is then used to reconstruct the tomogram (i.e., 3D volume). This process comes with inherent limitations. To limit radiation damage, only a minimal electron dose can be utilized during imaging, resulting in a low signal-to-noise ratio (SNR). In addition, both the structured background and the lamella thickness reduce particle contrast. Thicker lamellae broaden the spread of the electron beam, while the background exhibits

*Main contributor.

†[1] International Max Planck Research School on Cellular Biophysics, Frankfurt am Main HE 60438, Germany, [2] Frankfurt Institute for Advanced Studies, Frankfurt am Main HE 60438, Germany, [3] Institute of Physics, Goethe University Frankfurt, Frankfurt am Main HE 60438, Germany, [4] University of British Columbia, Vancouver BC V6T 1Z4, Canada [5] UMass Chan Medical School, Worcester, MA 01655, [6] Center for Computational Biology, Flatiron Institute, New York, NY 10010, USA, [7] Institute of Computer Science, Goethe University Frankfurt, Frankfurt am Main HE 60325, Germany, [8] Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA.

Machine Learning for Structural Biology Workshop

scattering strengths comparable to those of the particles. Furthermore, hardware limitations inherent to electron microscopes impede complete angular sampling, leading to the *missing wedge* effect. The missing projections introduce anisotropic resolution that distorts the fidelity of the particles' structural features. Lastly, the crowded and heterogeneous nature of cellular environments further obscures individual macromolecules. As a result, the dense and noisy backgrounds in cryo-ET tomograms present significant challenges for particle localization and segmentation, which are crucial for analyzing macromolecular structures within the cell [1].

Recent advancements in deep learning, particularly U-Net and CNN models, have shown promise for particle picking in cryo-ET by training on highly curated particle datasets [2, 3, 4]. Few-shot learning addresses some of the challenges of costly tomogram annotation [4, 5] but still requires a set of partial annotations for training. On the other hand, training with simulated data alleviates manual annotation burdens and allows the representation of complex biological structures. Simulation-based inference has been successful in extracting mechanistic information from complex experimental data in areas ranging from astrophysics to neuroscience [6] and single-particle cryo-EM [7]. However, developing realistic simulators for *in situ* environments and for the cryo-ET imaging process remains challenging [8, 9, 10]. In this work, we aim to develop a tool that utilizes simulation-based training to identify and locate macromolecules *in situ*, with particular focus on exploring various background models in cryo-ET data simulations.

## 2 Methods

**Simulator** We simulate small cubic volume crops, e.g., 32 voxels per edge at a magnification of 5-15 Å per voxel edge (Table 1), instead of a full tomogram, to improve memory handling and facilitate parallelization during training, testing, and inference. The simulation pipeline consists of four steps (Fig. 1A). First, we (i) sample the particle types that will populate the volume crop, assuming that each particle adopts a single known structure (i.e., the template). Each volume crop may contain from zero to an arbitrary number of particles. Then, we (ii) rotate and shift the particles, and simulate the volume with the appropriate voxel size. We then (iii) add the noise rescaled according to a given SNR and (iv) remove the missing wedge. Fig. 1B shows an example simulated volume. To provide our model with a diverse and comprehensive training set, we train it using simulations with a broad range of voxel sizes, rotations, shifts, SNRs, crowding per crop, and missing wedge angles (Table 1). Directly simulating the reconstructed volume crops enables us to bypass the simulation of aberration corrections and tilt-series alignment, making the simulator computationally more efficient. To explore the performance of different noise background simulators, we test three noise models: i) pure Gaussian white noise (synthetic), ii) empirical background taken from experimental tomograms (Table 1), and iii) a mixed simulator that simulates batches of volumes, with half containing synthetic noise and the other half containing an empirical background. For the empirical background crops, we rescale the voxel size accordingly. The experimental noise tomograms were obtained from *in vitro* samples and hippocampal tissue, and the samples differed between the training and testing sets.

**Architecture** State-of-the-art computer vision techniques for life science applications utilize U-Nets to perform multiclass segmentation of images and volumes [2, 4, 5, 11]. We build on this methodological framework to carry out particle picking in tomograms. We set up a dynamic U-Net with a number of pooling and upsampling layers, depending on the chosen crops' edge size (Fig. 1C). In addition, we implemented a deep supervision scheme on the U-Net latent representation by training a classifier head. This particle classifier comprises a 3-layer multilayer perceptron (MLP) and serves as a projector head in the context of object recognition tasks. These choices aim to improve the model's generalization performance by enhancing the learning of meaningful latent representations, and are not used for inference [12].

**Training** Training relies on supervision from the U-Net for multiclass semantic segmentation and from the MLP projector head for particle classification. We trained it exclusively on data simulated on the fly. We trained a model for each of the three different background simulators. For the mixed simulator case, we used a curriculum training (CT) schedule: we started with synthetic noise for the first third of training, then switched to the mixed simulator. To improve the model's classification performance in distinguishing between particles and background, we force our model to identify an additional 'background' class along with the other classes, such that the number of class channels becomes $N_{\text{classes}} = N_{\text{templates}} + 1$ [13].

The U-Net produces pixel-level logits for each class. Thus, to supervise its output, we utilize the ground-truth (GT) segmentation maps generated along with the simulated data. Each particle is segmented as a sphere centered on the particle with a fixed radius. These are compared pixel-wise with the predicted segmentation maps using a multiclass focal loss; we refer to Appendix 3.1 for further details.
To supervise the particle classification head, we designed a loss function that accounts for the permutation invariance of particle entries within each crop, since multiple particles in the same crop may belong to the same class. The classification head returns the logits for each class channel, particle-wise, which we supervise using a cross-entropy loss; we refer to Appendix 3.1 for further details. To enforce permutation invariance, we use an implementation of a variant of the Jonker-Volgenant algorithm with no initialization [14] to match the target and predicted particle label strings prior to evaluating the cross-entropy loss.
We use the AdamW optimizer [15] with a learning rate of $3 \times 10^{-4}$ and a weight decay of $1 \times 10^{-5}$. We train each model for 650,000 iterations and use a relative weight between the losses of 1.0. We use a batch size of 64. The ablation studies in Appendix Table 2 support our architectural choices.

**Testing** To evaluate the model's predictions, we chose the F1 score [16], which provides a measure of the model's inference performance, accounting for both its precision and recall (Appendix 3.2). To evaluate the F1 score on segmentation predictions, we first filter the segmentation maps to keep only voxels with class probabilities above $0.6$. We then post-process these maps using the connected-components-3d algorithm [17] to extract centroids of segmented objects (Fig. 2A, B). Connected components with voxel counts below 0.01 or above the 0.8 percentiles are excluded to avoid spurious centroid predictions arising either from the background class (higher end), or from single- or few-voxel connected components (lower end). We apply this procedure to both the GT segmentation maps and the predictions, then compare the centroids for each class per crop, allowing a tolerance of 5-voxel distance in each direction.

**Training and Testing data** Following the cryo-ET segmentation Kaggle challenge [18], we chose a similar task to test the background simulators. The tomograms for the empirical background model and the templates used during training and testing are found in Table 1. Note that training and testing empirical backgrounds differ.

## Results

**The model can pick particles in synthetic tomogram crops.** We start by visually inspecting the segmentation maps predicted by the model when using different background training strategies. We find that the model can correctly identify the particles when evaluated on test volumes using the same type of background as used for training, namely, either Gaussian noise or empirical background (Fig. 2A, B, respectively). Notably, the empirical background samples used for testing were never seen by the model during training, suggesting that these models can learn to discern the particles from the background irrespective of the tomogram used.

**The U-Net bottleneck's latent space shows separation between the classes.** Next, we examined the model's performance in representation learning. We applied dimensionality reduction to the U-Net bottleneck latent space, which has a dimensionality of 512. The latent space principal component analysis (PCA) is partially structured with the background class being well-differentiated from the others and the VLP class and the $\beta$-galactosidase class diverging in opposite directions (blue and pink, respectively, in Fig.2C top). Moreover, particle classification appears not to be strongly affected by differences in SNR values: values in the experimental range do not necessarily translate into class mixing (Fig. 2C, bottom).

**Training with mixed background makes the model's performance robust to background types.** We evaluated the particle-picking performance of our models trained on different background types (synthetic, real, and mixed; Fig. 1D, and Figs. 3, 4A, B). The model trained with a synthetic background performs satisfactorily only when evaluated on particles in a Gaussian background. Similarly, the model trained using the empirical background scored satisfactorily only when evaluated on particles in empirical background samples. This suggests that training only on one type of background does not allow the model to generalize across background types. Conversely, when

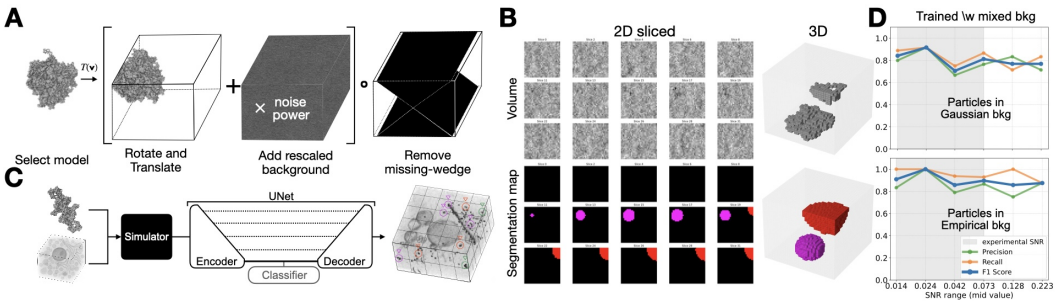


Figure 1: (**A**) The simulator pipeline. From left to right, we select the particle structures to populate the volumes, apply rotation and translation to each, rescale the background according to the SNR, and remove the missing wedge in Fourier space. (**B**) An example of a simulated volume with Gaussian background noise and its ground truth (GT) segmentation map (left), shown as 2D slices along the imaging axis. 3D representations of the volume (background not shown) and its GT segmentation (right). Red indicates a ribosome and magenta $\beta$-galactosidase. (**C**) The model's architecture. We input templates (maps or PDBs) and experimental or Gaussian background crops into the simulator (left), which combines them and passes them to the UNet (center), which finally outputs the segmentation maps of the tomogram crops (right). (**D**) Model performance when training with mixed background simulators. The line plots show F1 score, precision, and recall at different SNR levels (Appendix 3.2). We report the range mid-value.

trained using the simulator with mixed backgrounds, the model achieved an excellent performance in both tasks with an F1 score up to 1.0, demonstrating generalization across different backgrounds.

**Training on synthetic data is sufficient to distinguish particles from background in real tomograms.** To assess whether the simulator can reproduce cryo-ET-like data for training, we passed experimental crops through our best-performing model (trained with mixed backgrounds) and projected their latent embedding onto the PCA space spanned by the embedded simulated crops. We observed that the embeddings overlap (Fig. 5A, B), thereby substantiating our hypothesis. Next, we investigated the representation learning performance of our best model on real tomogram crops. To do this, we performed a PCA directly on the embeddings of the empirical crops. Our observations revealed that, in the latent space, crops with no particle (only background) are positioned separately from those that contain particles (Fig. 5C). This indicates that training the model solely on simulated data enables discrimination between crops containing particles and background-only crops in real tomograms. However, when we visualized the predicted segmentations, we noticed that the model struggles to accurately segment the crop (Fig. 5D).

## Conclusions and Discussion

In this study, we sought to identify optimal training strategies for background simulators in cryo-ET. Using a U-Net–based architecture, we show that a mixed background simulator is key to achieving robustness, compelling the model to distinguish particles from diverse backgrounds rather than memorizing background-specific features. While this strategy enabled distinguishing particles-containing from background-only crops in real tomograms, the model still struggles with classification. Misclassification of real particles cannot be attributed to low SNR, as the model performs well in equally noisy simulations (Fig. 5C, D, 2C). Moreover, the overlap between the simulated and experimental latent embeddings indicates that the simulator captures most experimental conditions (Fig. 5A, B). Thus, we believe the main pitfall lies in the templates. Although templates seem a straightforward choice for particle picking, they poorly represent molecular heterogeneity [19]. Using a single template for a particle creates bias in our training dataset. Also, the simulator does not account for particle distortions caused by tilt-series misalignment or aberration corrections, which could affect performance. In conclusion, while challenges related to particle distortions and heterogeneity remain, this work highlights the importance of mixed background simulators for advancing particle-picking annotation in cryo-ET.

## References


[1] Thorsten Wagner and Stefan Raunser. Cryo-electron tomography: Challenges and computational strategies for particle picking. *Current Opinion in Structural Biology*, 93:103113, 2025.

[2] Jessica E Heebner, Carson Purnell, Ryan K Hylton, Mike Marsh, Michael A Grillo, and Matthew T Swulius. Deep learning-based segmentation of cryo-electron tomograms. *J Vis Exp*, 189:e64435, 2022.

[3] Irene de Teresa-Trueba, Sara K Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E Zimmerli, Mauricio Toro-Nahuelpan, Dorothy WC Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, et al. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, 2023.

[4] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. Deepetpicker: Fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15(1):2090, 2024.

[5] Gokul Adethya, Bhanu Pratyush Mantha, Tianyang Wang, Xingjian Li, and Min Xu. Sasi: A self-augmented and self-interpreted deep learning approach for few-shot cryo-et particle detection. *arXiv preprint arXiv:2505.19948*, 2025.

[6] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[7] Lars Dingeldein, Pilar Cossio, and Roberto Covino. Simulation-based inference of single-molecule experiments. *Current Opinion in Structural Biology*, 91:102988, 2025.

[8] 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.

[9] Timothy Grant, Alexis Rohou, and Nikolaus Grigorieff. cis tem, user-friendly software for single-particle image processing. *elife*, 7:e35383, 2018.

[10] Alister Burt, Lorenzo Gaifas, Tom Dendooven, and Irina Gutsche. A flexible framework for multi-particle refinement in cryo-electron tomography. *PLoS biology*, 19(8):e3001319, 2021.

[11] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al. Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. *Nature methods*, 18(11):1386–1394, 2021.

[12] Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.

[17] William Silversmith. cc3d: Connected components on multilabel 3d & 2d images. *Zenodo*, 2021.

[18] Ariana Peck, Yue Yu, Jonathan Schwartz, Anchi Cheng, Utz Heinrich Ermel, Saugat Kandel, Dari Kimanius, Elizabeth Montabana, Daniel Serwas, Hannah Siems, et al. Annotating cryoet volumes: a machine learning challenge. *bioRxiv*, pages 2024–11, 2024.

[19] Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS letters*, 594(20):3243–3261, 2020.

[20] Utz Ermel, Anchi Cheng, Jun Xi Ni, Jessica Gadling, Manasa Venkatakrishnan, Kira Evans, Jeremy Asuncion, Andrew Sweet, Janeece Pourroy, Zun Shi Wang, et al. A data portal for providing standardized annotations for cryo-electron tomography. *Nature Methods*, 21(12):2200–2202, 2024.

# 3 Appendix

## 3.1 Losses: Focal and Cross-entropy Loss

Focal loss:

$$FL = \sum_{c} \sum_{v} \alpha_c (1 - p_{v,c})^{\gamma} \log p_{v,c} \ , \tag{1}$$

where $p_{v,c}$ is the predicted probability per voxel $v$ and class $c$. With this choice, we reweight each class by its frequency of appearance ($\alpha_c$), which is dynamically computed from the GT segmentation maps at each training step. The hyperparameter $\gamma$ controls the model's focus area. A smaller $\gamma$ will force the model to focus on areas that are harder to classify, while a larger $\gamma$ will push the model to focus on higher-confidence areas.

Cross-entropy loss:

$$CE = -\mathbb{E}_P(\log Q) \ , \tag{2}$$

where $P$ is the target distribution and Q is the prediction distribution, both parametrised as an $N_{\text{classes}}$-dimensional simplex.

## 3.2 Metrics: Precision, Recall and F1 Score

The precision is defined as the fraction of true positives (TP) over all the predicted samples predicted as positive:

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

where $FP$ is the number of false positives. The recall is defined as the fraction of true positives over all the target positive samples

$$\text{Recall} = \frac{TP}{TP + FN} \ .$$

The F1 score is defined as

$$\text{F1} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + 0.5(FP + FN)} \ .$$

### 3.3 The Model can Classify and Identify Particles in the Simulated Tomogram Crops

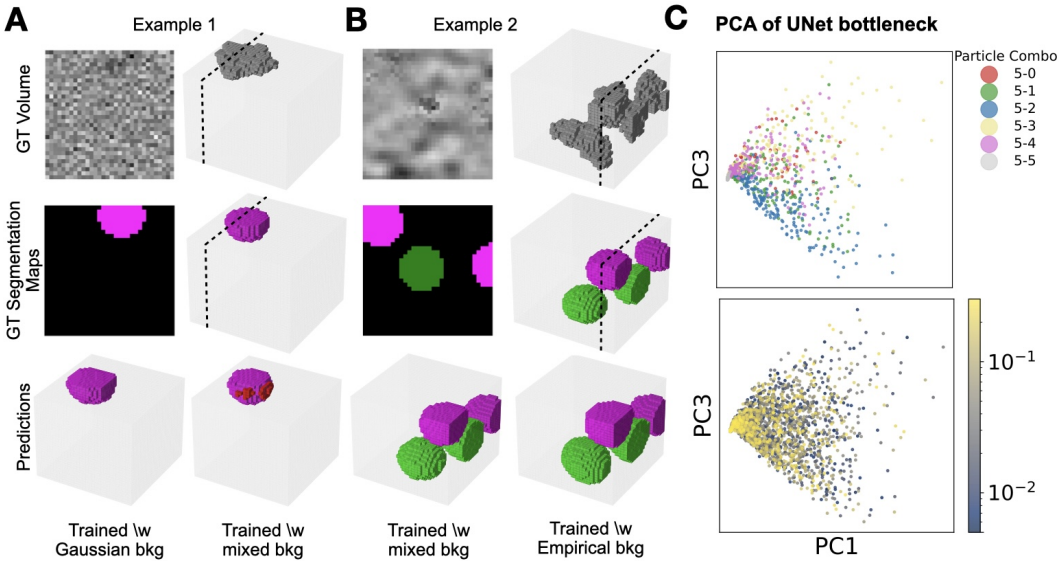


Figure 2: The model classifies and identifies particles in the simulated tomogram crops with Gaussian noise (**A**) and empirical background (**B**). Examples of the GT volumes and segmentation are compared to the predictions; the 2D slices are indicated by dashed lines on the 3D volumes. The bottom row shows the predictions when training with a Gaussian background (bkg) or with the mixed bkg. (**C**) PCA of the UNet bottleneck latent space, evaluated on Gaussian bkg volumes, coloured according to the particle classes (top) and SNR values (bottom) for the model trained with Gaussian bkg.

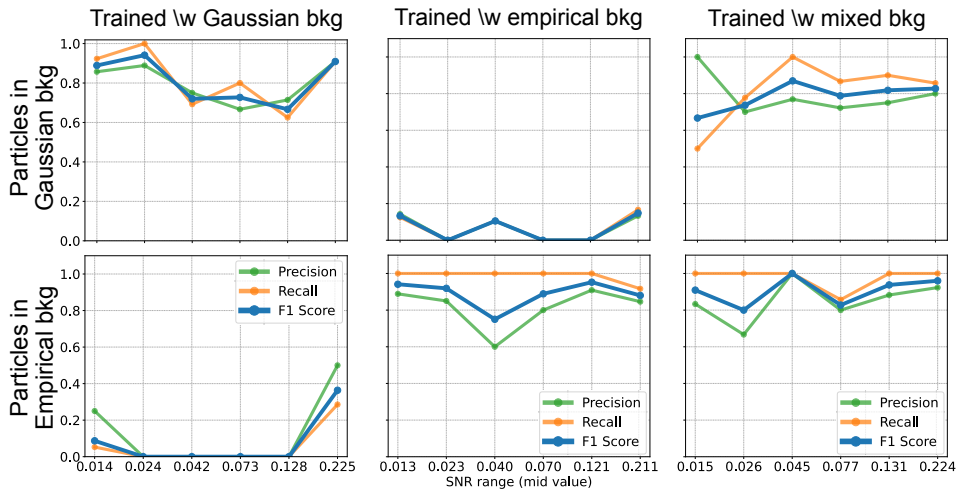


Figure 3: Comparison of model performance between training with synthetic, empirical, and mixed background simulators. The line plots show F1 score, precision, and recall at different SNR levels (Appendix 3.2). We report the range mid-value.

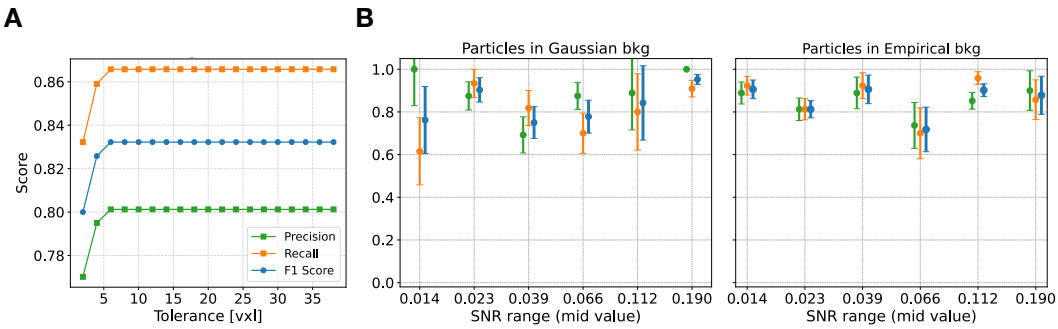


Figure 4: Statistical analysis of the model trained with mixed backgrounds. (**A**) Metrics of the models depending on the allowed distance (in voxels) between the predicted centroids and the ground truth. (**B**) average model metrics across different SNR intervals. The error bars correspond to the standard deviation of the mean.

### 3.4 Training on Synthetic Particles is Sufficient to Distinguish Particles from Background in Real Tomograms

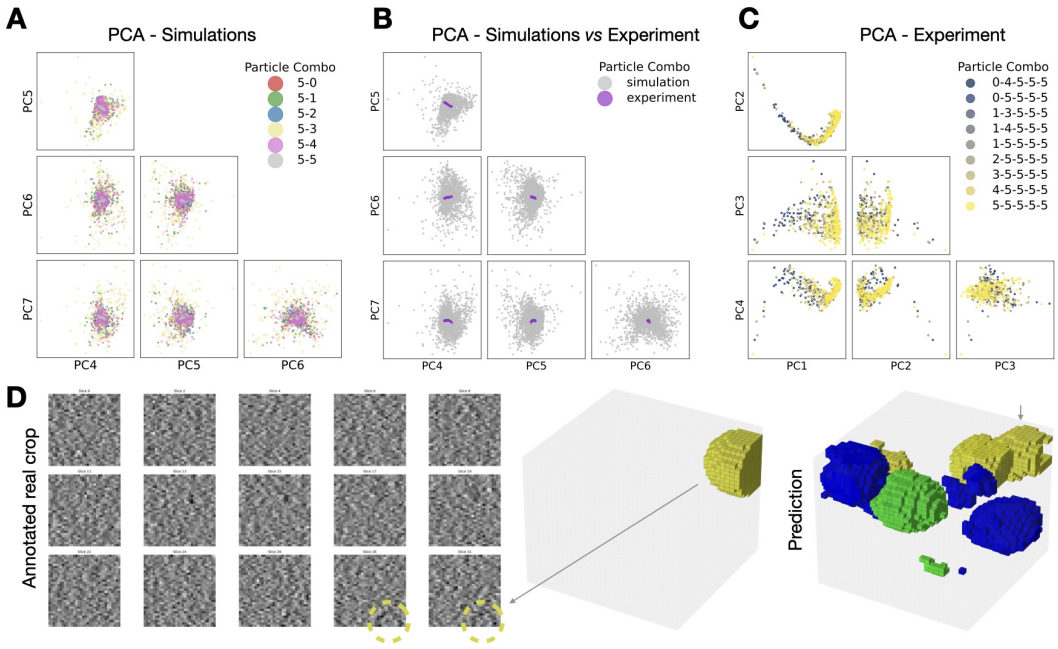


Figure 5: The model's latent representation distinguishes pure background from particles in real tomograms. (**A**) PC4 to 7 of the UNet bottleneck latent space representation (original dimensionality 512) of mixed backgrounds tomograms. (**B**) PC4 to PC7 plots showing the reduced representation of the latent space of simulated and empirical data. After passing the empirical data through the model, we projected them onto the same PCA space as the simulated data. (**C**) First four PCs of the UNet bottleneck latent-space representation of real tomogram crops. The datapoints are coloured according to the combination of class objects in the crops. Note that class 5 is the background class. (**D**) Example of a real tomogram crop (left and center) and its predicted segmentation map (right).

### 3.5 Input Configuration Parameters of the Simulator used during Training and Testing

The tomograms used can be found on the Cryo-ET Data Portal at `https://cryoetdataportal.czscience.com/browse-data/datasets` [20]. The maps used as templates can be found on EMPIAR at `https://www.ebi.ac.uk/empiar`. With an edge length of 32 voxels, the U-Net has a

Table 1: Simulator configuration parameters

| Parameter | Value |
|---|---|
| Batch size | 64 |
| Crop edge | 32 (vxl) |
| Obj per crop | 5 |
| SNR | [0.0005-0.3] |
| Shift | [0-20] (vxl) |
| Voxel size | [5.0-15.0] (Å) |
| Sphere radius | 12 (vxl) |
| Tomograms used as bkg (Train) | TS_103_5.mrc |
| | TS_102_2.mrc |
| | TS_100_3.mrc |
| | TS_73_6.mrc |
| | TS_5_4.mrc |
| | 24nov01d_Position_38.mrc |
| | 24nov01d_Position_42.mrc |
| | 24nov01d_Position_43.mrc |
| Tomograms used as bkg (Test) | TS_6_6.mrc |
| Templates | emd_3883.map (Ribosome, class 0, red) |
| | emd_24181.map (Thyroglobulin, class 1, green) |
| | emd_41917.map (VLP, class 2, blue) |
| | emd_41923.map (Apoferritin, class 3, yellow) |
| | emd_0153.map ($\beta$-galactosidase, class 4, magenta) |

depth of 5 layers, comprising 5 pooling layers (for downsampling) and 5 transposed convolutional layers (for upsampling). Each up- or down-sampling operation is followed by a double convolution. Before the first downsampling layer, we perform a double convolution. Each convolution includes a 3D batch normalization operation followed by a ReLU activation. Each run used 1 A100 Nvidia GPU, 4 CPUs per node, and 16 GB of memory per CPU.

### 3.6 Ablation Studies

We conducted ablation studies to evaluate the projector head's impact on centroid prediction accuracy. Additionally, we examined the impact of using spheres with a fixed radius for each class, rather than a different radius for each class, when creating the GT segmentation maps for segmentation supervision. We performed ablation studies based on our most effective training strategy, which used a mixed background simulator. For testing on real backgrounds, we used different tomograms than those used during training. We report the average precision, recall, and F1 score, with a tolerance of 5 voxels for centroid comparison.

Table 2: Ablation studies. The "Loss weight" field indicates the relative weight applied to the projector loss. P, R and F1 stand for Precision, Recall and F1 score, respectively

| Ablation | Loss weight | Gaussian bkg | | | Empirical bkg | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| Same sized spheres | 1.0 | 0.7 | 0.8 | 0.8 | 0.8 | 0.9 | 0.9 |
| Same sized spheres | 0.1 | 0.7 | 0.7 | 0.8 | 0.7 | 0.7 | 0.8 |
| Same sized spheres | 0.0 | 0.7 | 0.8 | 0.9 | 0.7 | 0.8 | 0.8 |
| Different sized spheres | 1.0 | 0.5 | 0.5 | 0.5 | 0.4 | 0.6 | 0.5 |

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract discusses the exploration of various background models in cryo-ET simulators for particle picking, which is the primary focus of our experiments.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In the conclusion, we explicitly discuss the pitfalls of our approach.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: We do not explore any theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide all the necessary information concerning the architecture and training strategies to reproduce our experiments in the methods section and Appendix.

   Guidelines:

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

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

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

Answer: [No]

Justification: We provide detailed instructions to reproduce the experimental results shown.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide the data IDs and source URLs of the databases we used in Appendix

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide the results obtained on a large simulated data sample to ensure statistical significance (10000 simulated crops per experiment)

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We comment on the resources needed for training in the Appendix

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We do follow NeurIPS code of ethics and preserve anonymity.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We briefly comment on the importance of cryo-ET technology.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our work does not pose such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper cites all the relevant literature and assets.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: We are not releasing new assets as the work reported is part of an ongoing project.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

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

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

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

    Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The use of LLMs is not scientifically relevant in this work.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.