TOMODETR: POINT-SET PREDICTION WITH CROSS-WINDOW REFINEMENT FOR CRYO-ET PARTICLE LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Particle localization is a critical step in cryo-electron tomography (cryo-ET), providing the coordinates needed for downstream structural analysis. State-of-the-art supervised methods have achieved strong localization performance by first predicting dense particle masks or response maps and then extracting center coordinates. Yet the target itself is a labeled set of particle centers. We therefore develop TomoDETR, a DETR-based framework that formulates particle localization as point-set prediction from point annotations. Large tomograms require overlapping-window inference, causing the same particle to be predicted repeatedly from different window contexts. We find that these repeated observations differ in localization quality, while classification confidence does not reliably identify the most accurate one. Motivated by this finding, TomoDETR incorporates a Cross-Window Query Refiner (CWQR), which learns interactions among nearby candidates from different windows to jointly refine candidate centers and class scores before final selection. On the CZII particle localization benchmark, TomoDETR achieves the highest weighted F_4 among the evaluated methods. Under a shared backbone and matched training setup, direct point-set prediction improves weighted F_4 by 2.30 points over dense prediction with center extraction, while CWQR provides a further 2.85-point improvement with only 2.9% measured inference-time overhead. Additional evaluations on cellular and simulated tomograms further support the effectiveness of cross-window refinement. Our code will be released publicly upon acceptance.

1 INTRODUCTION

Cryo-electron tomography (cryo-ET) reveals macromolecules in their cellular context. Particle localization identifies their centers and classes, providing the coordinates needed to extract subvolumes for structural analysis (Rice et al., 2023). Reliable localization in large, noisy tomograms is therefore a central challenge in cryo-ET analysis.

State-of-the-art supervised particle-localization methods achieve strong performance, but commonly formulate localization as dense voxel-wise prediction rather than direct coordinate prediction. DeepFinder and DeepETPicker, for example, predict dense segmentation or center-derived masks from point annotations and require a separate coordinate-extraction step to recover particle centers (Moebel et al., 2021; Liu et al., 2024). Yet the target itself is a labeled set of particle centers. We instead formulate particle localization directly in this point-set representation, predicting continuous centers and classes from center annotations (Figure 1).

DETR-style bipartite matching provides a natural foundation for set prediction (Carion et al., 2020; Zhu et al., 2021). For large tomograms, however, full-volume 3D inference is impractical under GPU memory limits, so detection is performed over overlapping windows. The same physical particle can therefore appear in several windows under different local contexts, producing multiple window-level predictions. These predictions are generated independently, with no mechanism to compare or combine information across windows before final selection. We find that repeated predictions of the same particle can differ substantially in localization quality, while classification confidence does not

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

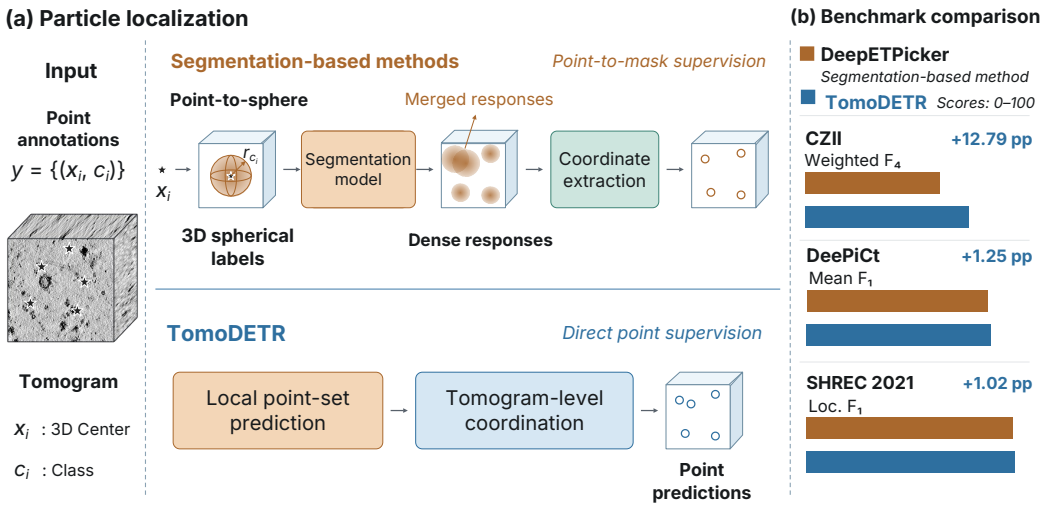

Figure 1: **Point-set prediction for particle localization.** (a) Center-derived mask supervision followed by coordinate extraction, compared with TomoDETR's direct point supervision and tomogram-level coordination. (b) Comparison with DeepETPicker on three benchmarks; bars use the reported scores, averaging repeated runs where available. DeePiCt bars further average the VPP and defocus mean scores. Full comparisons and standard deviations are in Table 1.

reliably identify the most accurate one. Treating them only as duplicates to suppress can therefore discard useful localization information.

We introduce TomoDETR, which combines window-level point-set prediction with the *Cross-Window Query Refiner* (CWQR). The first stage directly predicts particle centers and classes within each 3D window. CWQR then brings candidates into tomogram coordinates, enables nearby predictions from different windows to exchange information, and jointly refines center coordinates and class logits under tomogram-level one-to-one supervision. This gives each candidate access to cross-window context before final selection, without requiring predefined duplicate groups.

Our experiments evaluate point-set prediction, diagnose repeated observations, and test the benefit of learned cross-window refinement. On CZII (Peck et al., 2025; 2026), under a shared backbone, training data, and update budget, direct point-set prediction improves weighted F_4 by 2.30 points over dense Gaussian-heatmap prediction with center extraction. We further find that repeated predictions of the same particle differ substantially in localization quality, while confidence-based selection does not reliably identify the best-localized prediction. CWQR improves the same window-level detector by a further 2.85 points, whereas fixed suppression and fusion rules recover little of this gain. A parameter-matched self-only control further supports the benefit of cross-window communication beyond additional refinement capacity. CWQR also improves the window-level detector on SHREC2021 (Gubins et al., 2021) and both DeePiCt imaging conditions (de Teresa-Trueba et al., 2023).

Our main contributions are summarized as follows:

- A point-set formulation for particle localization in reconstructed tomograms, predicting continuous particle centers and classes directly from center annotations.
- CWQR, a sparse cross-window refiner that enables interaction among nearby candidates from different windows and jointly refines coordinates and class logits with tomogram-level set supervision.
- An empirical analysis showing that repeated cross-window observations vary in localization quality, together with controlled experiments demonstrating gains beyond fixed fusion and parameter-matched self-only refinement.

2 Related Work

Particle localization in reconstructed tomograms. Supervised dense-prediction methods achieve state-of-the-art particle-localization performance in reconstructed cryo-ET tomograms. DeepFinder and DeepETPicker learn dense semantic labels or center-derived masks and subsequently recover particle coordinates from the predicted volumes (Moebel et al., 2021; Liu et al., 2024). Other supervised approaches use Gaussian centroid responses or voxel-wise macromolecule labels (Shah et al., 2025; Dhakal et al., 2026). Despite differences in target design and architecture, these approaches share an intermediate dense representation from which particle centers must be extracted. Alternative paradigms include promptable segmentation, embedding-based picking, self-supervised pattern mining, and structural template matching (Wiedemann et al., 2026; Rice et al., 2023; Huang et al., 2024; Cruz-León et al., 2024).

Set prediction. DETR introduced bipartite matching for one-to-one set prediction, with Deformable DETR extending this framework to efficient multi-scale attention (Carion et al., 2020; Zhu et al., 2021). Related work has applied detection to other cryo-EM representations: CryoTransformer operates on 2D micrographs, while FullTilt predicts macromolecule locations directly from aligned tilt series (Dhakal et al., 2024; Ho & Bartesaghi, 2026). To our knowledge, TomoDETR is the first to formulate supervised particle localization in reconstructed cryo-ET tomograms as point-set prediction from center annotations.

Cross-window coordination. Overlapping-window inference requires predictions from different windows to be reconciled during post-processing. DeepETPicker, for example, uses fixed mean-pooling non-maximum suppression together with overlap-tile processing to recover particle centers from predicted voxel labels (Liu et al., 2024). Prior object-detection work has learned either suppression decisions or relations among detection candidates (Hosang et al., 2017; Hu et al., 2018). CWQR enables learned information exchange among spatially nearby candidates from different windows, jointly updating particle centers and class scores before final selection rather than only deciding which prediction to retain. It uses tomogram-level one-to-one supervision without predefined duplicate groups.

3 Method

3.1 Problem Formulation and Overview

Given a tomogram $V \in \mathbb{R}^{D\times H\times W}$, our goal is to predict a particle-center set $\hat{\mathcal{Y}} = \{(\hat{\mathbf{x}}_n, \hat{c}_n)\}_{n=1}^{\hat{N}}$, where $\hat{\mathbf{x}}_n \in \mathbb{R}^3$ is a continuous center coordinate in the tomogram and $\hat{c}_n \in \{1, \ldots, C\}$ is its particle class. TomoDETR first predicts a particle set within each overlapping 3D window; we

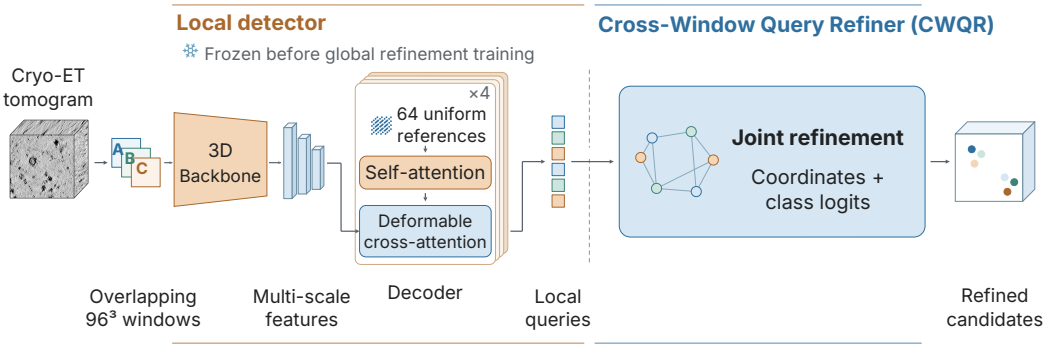

Figure 2: **TomoDETR architecture.** Multi-scale volumetric features support query-based local set prediction. After local training, the detector is frozen. CWQR pools its candidates in tomogram coordinates, exchanges information over a sparse cross-window graph, and refines class logits and coordinates.

refer to these window-level predictions as local predictions. Overlapping windows observe the same physical particle from different window contexts, producing repeated predictions. Cross-Window Query Refiner (CWQR) enables nearby candidates from different windows to interact and jointly refines their class logits and center coordinates before final selection (Figure 2).

3.2 Local Spatial Set Prediction

Spatial queries and deformable decoding. To provide spatial coverage throughout each 3D window, we place $Q = 64$ references $\{\mathbf{r}_i\}_{i=1}^{Q}$ at the cell centers of a uniform $4 \times 4 \times 4$ grid. The 64-query budget covers the observed window counts in our audit of experimentally acquired CZII and DeePiCt data; denser synthetic SHREC2021 windows are reported separately in Appendix C. For each window W_m, a 3D residual U-Net produces three feature scales, and each reference initializes a query by combining its positional encoding with the local visual feature:

$$\mathbf{q}_{mi}^{0} = \phi(\mathbf{r}_i) + \mathbf{f}_m(\mathbf{r}_i), \tag{1}$$

where ϕ encodes the reference position and $\mathbf{f}_m(\mathbf{r}_i)$ is obtained by trilinear interpolation from the highest-resolution feature map. Four decoder layers then iteratively update the queries using self-attention and 3D deformable cross-attention over all three feature scales (Carion et al., 2020; Zhu et al., 2021). Each layer predicts class logits and center updates; the final layer outputs local candidates consisting of continuous centers, class logits, and query features, which are passed to CWQR.

Window-level set supervision. Let $\mathcal{Y}_m = \{(\mathbf{u}_{mj}, c_{mj})\}_{j=1}^{N_m}$ denote the particle annotations in window W_m. At each decoder layer ℓ, we compute a one-to-one Hungarian assignment using the center-and-class matching cost

$$\mathcal{C}_m^{\ell}(i,j) = \lambda_{\text{center}} \left\| \mathbf{r}_{mi}^{\ell} - \mathbf{u}_{mj} \right\|_1 - \lambda_{\text{class}} \log p_{mi}^{\ell}(c_{mj}), \tag{2}$$

where $\mathbf{r}_{mi}^{\ell}$ and $\mathbf{u}_{mj}$ are normalized predicted and target centers, and p_{mi}^{ℓ} denotes the predicted class probabilities. Matched queries receive L_1 center regression and cross-entropy classification supervision, while unmatched queries are assigned to no-object; auxiliary supervision is applied at all decoder layers. Because matching is performed independently within each window, predictions of the same physical particle from different overlapping windows remain uncoordinated.

3.3 Cross-Window Query Refiner

Candidate pooling and cross-window graph. Before NMS, we map local candidates from all windows to tomogram coordinates and retain those whose maximum foreground probability exceeds a low threshold τ_0, together with their query features, class logits, centers, and source-window identities. CWQR connects spatially nearby candidates originating from different windows (Figure 3). For candidate i with center $\mathbf{x}_i$ and source window w_i, we define

$$\begin{aligned} \mathcal{A}(i) &= \{j : w_j \neq w_i,\ \|\mathbf{x}_j - \mathbf{x}_i\|_2 \leq r_g\}, \\ \mathcal{N}(i) &= \{i\} \cup \text{Nearest}_{K-1}(\mathcal{A}(i)). \end{aligned} \tag{3}$$

Here r_g is the interaction radius. Neighborhoods are defined by source-window identity and spatial proximity, without predefined duplicate groups. The graph is constructed once from the local predictions and contains at most K neighbors per candidate, avoiding tomogram-wide all-to-all attention.

Geometry-aware cross-window interaction. Each candidate combines its local decoder feature with tomogram- and window-level geometry:

$$\mathbf{h}_i^{0} = E_q(\mathbf{q}_i) + \phi_g(\mathbf{x}_i) + \phi_w(\mathbf{g}_i^{\text{win}}), \tag{4}$$

where $\mathbf{q}_i$ is the decoder query feature, ϕ_g encodes its tomogram position, and ϕ_w encodes its position and boundary proximity within the source window. Attention within $\mathcal{N}(i)$ further incorporates relative geometry:

$$\begin{aligned} a_{ij} &= \frac{(\mathbf{W}_Q \mathbf{h}_i)^{\top} (\mathbf{W}_K \mathbf{h}_j)}{\sqrt{d_h}} + \psi\left(\frac{\mathbf{x}_j - \mathbf{x}_i}{r_g}, \frac{\|\mathbf{x}_j - \mathbf{x}_i\|_2}{r_g} \right), \\ \alpha_{ij} &= \text{softmax}_{j \in \mathcal{N}(i)}(a_{ij}). \end{aligned} \tag{5}$$

Here d_h is the head dimension and ψ is a learned head-specific bias from relative displacement and distance. Two sparse Transformer blocks update the candidate representations through these neighborhoods.

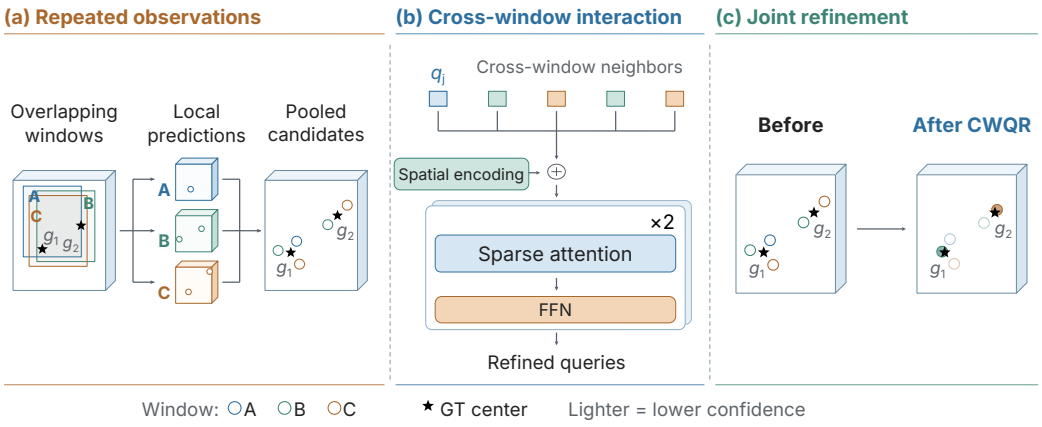

Figure 3: **Cross-window query refinement.** Local predictions from overlapping windows are pooled in tomogram coordinates, connected through sparse cross-window neighborhoods, and jointly updated using query features and relative geometry before final selection. Colors denote source windows; lighter markers indicate lower confidence. The refinement effect is schematic.

Joint refinement and tomogram-level supervision. From the final cross-window representation $\mathbf{h}_i$, two zero-initialized heads predict residual updates to the local class logits and centers:

$$\begin{aligned} \mathbf{z}'_i &= \mathbf{z}_i + g_z(\mathbf{h}_i), \\ \mathbf{x}'_i &= \mathrm{clip}_\Omega[\mathbf{x}_i + \Delta_{\max} \tanh(g_x(\mathbf{h}_i))] \,. \end{aligned} \tag{6}$$

Here $\Delta_{\max}$ bounds coordinate displacement along each axis and clip_Ω keeps predictions inside the tomogram. We then jointly match all refined candidates to the complete tomogram annotation using a single Hungarian assignment with the same center-and-class matching cost as Eq. (2). Each particle is assigned to at most one candidate; unmatched candidates receive the no-object label. CWQR is optimized with cross-entropy classification and SmoothL1 center regression, with center errors normalized by the class-specific localization tolerance.

We first train and freeze the local detector and then train CWQR on its cached complete-tomogram candidate sets. At inference, CWQR updates the pooled candidates before NMS and score thresholding (class-aware NMS for CZII and DeePiCt, and class-agnostic NMS for SHREC2021); additional implementation and loss details are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate particle localization in three complementary cryo-ET settings. **CZII** contains experimentally acquired realistic-phantom tomograms (Peck et al., 2025; 2026). We use seven tomograms for training, 40 separate tomograms for model and threshold selection, and reserve 364 tomograms for testing. **SHREC2021** contains simulated tomograms (Gubins et al., 2021). We develop on tomograms 0–7, select the training schedule and inference settings on tomogram 8, then retrain on tomograms 0–8 with these settings fixed and evaluate on tomogram 9. **DeePiCt** contains cellular *S. pombe* tomograms acquired with Volta phase plate (VPP) or defocus imaging (de Teresa-Trueba et al., 2023). For each imaging condition independently, we use three fixed splits containing six training, two calibration, and two test tomograms.

Evaluation. For CZII, we use the official class-weighted F_4 score. For each class, $F_4 = 17PR/(16P + R)$, where P and R denote precision and recall. The choice of $\beta = 4$ prioritizes recall over precision, penalizing missed particles more heavily than false positives. Class weights further emphasize the more challenging β-galactosidase and thyroglobulin classes, each assigned

twice the weight of the other scored classes. Class weights and the matching tolerance of half the class radius follow the official benchmark protocol. For SHREC2021, we follow the official localization F_1 protocol, which scores unique particle detections independently of class labels (Gubins et al., 2021). For DeePiCt, predictions are matched one-to-one to annotations of the same class within 10 voxels; unmatched and duplicate predictions count as false positives. We report pooled F_1, computed by pooling TP, FP, and FN counts across ribosomes and fatty acid synthase (FAS). Full metric definitions are provided in Appendix A.2.

Comparisons and training. We retrain DeepETPicker (Liu et al., 2024), DeepFinder (Moebel et al., 2021), and OCTOPI (Schwartz et al., 2026) on the corresponding training data. We evaluate TomoDETR without CWQR, which predicts particle centers within individual windows, and the complete system with CWQR.

For CZII, TomoDETR with and without CWQR and the same-backbone Gaussian-heatmap baseline select class-specific thresholds on the 40 validation tomograms and fix them before testing. The reproduced external CZII baselines use 121 public tomograms for selection, a superset of our 40 validation tomograms. Additional architecture, training, and selection details are given in Appendix A.

Reporting. Scores are percentages and differences are percentage points, computed before rounding. Unless stated otherwise, $\pm$ denotes sample standard deviation over three training seeds for CZII and SHREC2021 and over three fixed data splits for DeePiCt. External CZII baselines are single reproduced runs. In Table 1(a), bold scores identify the best complete systems, and light shading groups the two TomoDETR variants.

4.2 Point-Set Prediction for Particle Localization

Our first question is whether particle localization can be effectively formulated as point-set prediction. We evaluate this at two levels: first as a complete localization system across benchmarks, and then in a controlled same-backbone comparison against dense Gaussian-heatmap prediction.

Cross-benchmark localization. Table 1(a) compares the complete TomoDETR system with locally reproduced dense-prediction pipelines and includes the without-CWQR ablation. On CZII, TomoDETR reaches 72.92 ± 0.45 weighted F_4, compared with 66.61 for OCTOPI, the strongest reproduced external baseline in this experiment. On SHREC2021, TomoDETR exceeds the strongest baseline, DeepETPicker, achieving localization F_1 scores of 93.08 and 92.06, respectively.

On cellular DeePiCt VPP data, TomoDETR achieves 82.34 pooled F_1, above DeepETPicker's 81.47. On defocus, TomoDETR obtains 81.67 pooled F_1, compared with 80.04 for DeepETPicker. These results show that point-set prediction provides competitive localization across simulated, realistic-phantom, and cellular cryo-ET settings, with the strongest relative performance on CZII.

Controlled formulation comparison. We compare TomoDETR without CWQR against a Gaussian-heatmap pipeline on CZII using the same backbone, training data, and training budget. Table 1(b) summarizes the matched settings.

Direct point-set prediction improves weighted F_4 from 67.77 ± 0.54 to 70.07 ± 0.40, a paired gain of 2.30 ± 0.15 points across the three seeds (Table 1(b)). Classwise results, including the apo-ferritin case where Dense performs better, are reported in Appendix A.4.

4.3 Repeated Observations Vary in Localization Quality

Point-set prediction is performed on overlapping windows, so one physical particle can produce several local predictions under different window contexts. If all repeated predictions were equally informative, ordinary confidence-based selection or duplicate suppression would be sufficient. We instead test whether the repeated observations differ in localization quality and whether confidence identifies the most accurate one.

Table 1: **Main localization results and controlled formulation comparison.** (a) Scores (%): CZII (Peck et al., 2025; 2026), official weighted F_4; SHREC2021 (Gubins et al., 2021), official localization F_1; DeePiCt (de Teresa-Trueba et al., 2023), pooled F_1. Boldface marks the better score within each comparison; shading groups the two TomoDETR variants. Δ denotes paired gains in percentage points. (b) CZII comparison with matched backbone, training data, and update budget; details in Appendix A.3. Reporting conventions follow Sec. 4.

(a) Cross-benchmark localization

Method	CZII $F_4 \uparrow$	SHREC $F_1 \uparrow$	VPP $F_1 \uparrow$	Defocus $F_1 \uparrow$
DeepETPicker	60.13	92.06 ± 0.71	81.47 ± 5.73	80.04 ± 1.40
DeepFinder	51.65	79.59 ± 2.52	77.80 ± 8.24	77.25 ± 1.85
OCTOPI	66.61	85.40 ± 0.62	72.60 ± 17.75	72.33 ± 4.01
TomoDETR w/o CWQR	70.07 ± 0.40	86.99 ± 2.31	80.16 ± 5.25	79.08 ± 0.90
TomoDETR	**72.92 ± 0.45**	**93.08 ± 0.16**	**82.34 ± 5.91**	**81.67 ± 0.56**
CWQR gain (Δ)	+2.85 ± 0.18	+6.09 ± 2.18	+2.18 ± 0.73	+2.58 ± 0.45

(b) Controlled formulation comparison on CZII

Prediction formulation	Weighted $F_4 \uparrow$
Dense Gaussian heatmap	67.77 ± 0.54
Direct point-set prediction (without CWQR)	**70.07 ± 0.40**
Paired Direct − Dense	+2.30 ± 0.15

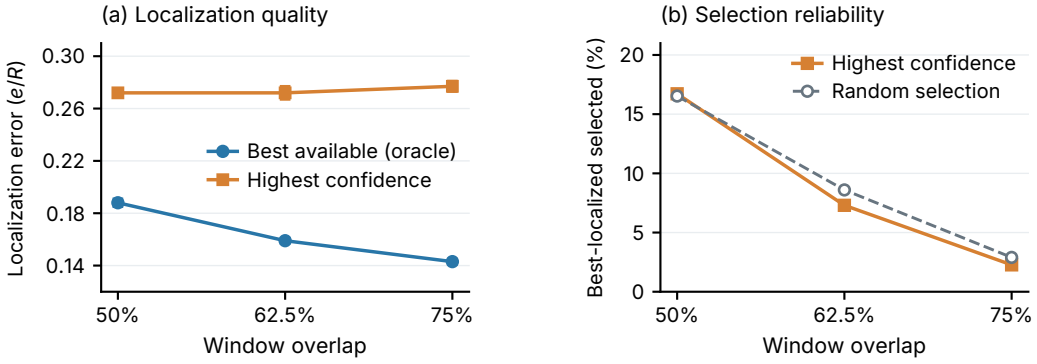

Figure 4: **Repeated observations provide unequal localization evidence.** CZII particles with at least two correct-class observations at every overlap. (a) Localization error of the best available observation (oracle) and the highest-confidence observation, normalized by class radius R; error bars show standard deviations over three seeds. (b) Frequency with which the highest-confidence candidate or a random candidate drawn from the same observation set is the best localized.

Repeated observations contain unequal localization evidence. Figure 4 analyzes correct-class observations associated with the same particle on CZII. At the standard 50% overlap, the best available observation has normalized localization error $e_{\text{best}} = 0.188 \pm 0.003$, whereas selecting the highest-scoring observation gives $e_{\text{score}} = 0.272 \pm 0.002$. Nevertheless, the highest-scoring observation is also the best-localized one for only 16.72% of particles, essentially the same as the 16.53% success rate of a random draw from the same observation set.

To test whether additional observations provide better center estimates, we increase window overlap so that each particle is observed in more windows. From 50% to 75% overlap, the best available error decreases from 0.188 to 0.143 e/R, while the error of the highest-scoring observation remains approximately unchanged (0.272 to 0.277). At the same time, the probability that the highest-scoring

Table 2: **Refinement controls on CZII.** Weighted F_4 (%), mean $\pm$ SD. (a) Selection rules use the same frozen predictions and are calibrated separately. (b) Variants use the same frozen detector for each of the three seeds; full CWQR results are reused from Table 1. Self-only and CWQR have identical parameter counts.

Configuration	Parameters	Cross-window	Weighted $F_4 \uparrow$
(a) Fixed fusion versus learned refinement			
Hard NMS			70.07 ± 0.40
Distance-based Soft-NMS			70.19 ± 0.37
Uniform averaging			70.31 ± 0.58
Score-weighted averaging			70.34 ± 0.51
CWQR			$\mathbf{72.92 \pm 0.45}$
(b) Cross-window communication			
Pointwise	333K	$\times$	69.36 ± 0.12
Self-only	469K	$\times$	71.09 ± 0.16
CWQR	469K	✓	$\mathbf{72.92 \pm 0.45}$

observation is the best-localized one falls to 2.27%. These results show that overlapping windows provide multiple observations with meaningfully different localization quality, while classification confidence is a poor mechanism for identifying which observation contains the best center estimate.

4.4 Cross-Window Refinement Beyond Fixed Fusion

The repeated-observation analysis motivates learning from information available across overlapping windows. We next test whether CWQR improves localization beyond fixed fusion and whether cross-window communication contributes beyond additional per-candidate refinement capacity.

Cross-benchmark refinement gains. The without-CWQR and full-model rows in Table 1(a) isolate the effect of refinement, with paired gains summarized in the Δ row. On CZII, CWQR increases weighted F_4 from 70.07 ± 0.40 to 72.92 ± 0.45, a gain of 2.85 ± 0.18 points. On SHREC2021, localization F_1 increases from 86.99 ± 2.31 to 93.08 ± 0.16, corresponding to $+6.09 \pm 2.18$ points. On DeePiCt, pooled F_1 improves by 2.18 ± 0.73 points on VPP and 2.58 ± 0.45 points on defocus.

Beyond fixed fusion. Using the same frozen CZII predictions, we compare hard NMS with distance-based Soft-NMS (Bodla et al., 2017), uniform coordinate averaging, and score-weighted coordinate averaging. After separately calibrating each rule, the best fixed alternative reaches 70.34 ± 0.51 weighted F_4, only 0.27 points above hard NMS. In contrast, CWQR reaches 72.92 ± 0.45, a 2.85-point improvement over the same hard-NMS local endpoint (Table 2(a)). Thus, the observed gain is not explained by generic duplicate suppression or by the tested fixed coordinate-fusion rules. Appendix B adds suppression-radius and window-selection controls, FP decomposition, and paired localization analyses.

Benefit of cross-window communication. We next isolate the contribution of cross-window communication from additional refinement capacity. *Pointwise* refines each candidate independently, while *Self-only* uses the same architecture and parameter count as CWQR but restricts attention to each candidate itself. Thus, Self-only provides a parameter-matched control without cross-window information. After final selection, Self-only reaches 71.09 ± 0.16 weighted F_4, whereas CWQR reaches 72.92 ± 0.45, a mean-score improvement of 1.83 points (Table 2(b)). This matched comparison shows that cross-window communication contributes beyond an equally sized per-candidate refiner.

These experiments provide two complementary findings: repeated observations differ in localization quality, and confidence does not reliably select the best-localized one; learned communication among nearby candidates from different windows improves localization beyond the tested fixed fusion rules and an equally sized self-only refiner. Together, these findings motivate and support learned cross-window refinement.

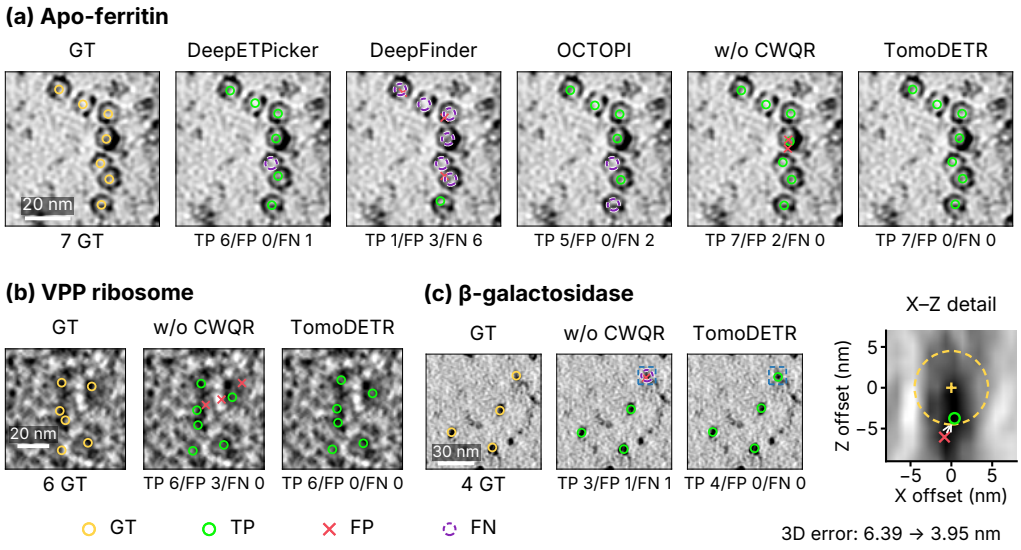

Figure 5: **Qualitative localization comparison.** (a) Particle localization in a crowded apo-ferritin region. (b) False-positive suppression in VPP ribosomes. (c) Improved β-galactosidase localization. Yellow circles denote GT centers, green circles TP, red crosses FP, and dashed purple circles FN. In (c), blue boxes locate the X–Z detail: the yellow cross marks GT, the yellow circle indicates the 4.5 nm matching tolerance, and the arrow compares the selected positions without and with CWQR; both predictions have the correct class. Images are mean projections through slabs of 15.0, 59.3, and 26.0 nm in (a–c), and 6.0 nm along Y for the detail. Matching uses 3D coordinates.

Qualitative comparison. The examples in Figure 5 illustrate how window-level detection and cross-window refinement contribute to localization. TomoDETR separates nearby particles in crowded regions, while CWQR suppresses spurious predictions and can bring a correctly classified prediction within the localization tolerance.

5 Conclusion

In this work, we show that direct point-set prediction is a competitive alternative to supervised dense-prediction approaches for cryo-ET particle localization. We find that repeated predictions from overlapping windows differ in localization quality, while confidence-based selection does not reliably identify the most accurate observation. Our controlled experiments further show that learned communication among nearby candidates from different windows improves localization beyond fixed fusion and parameter-matched self-only refinement.

Limitations and Future Work. TomoDETR remains a supervised particle-localization framework. Models are trained and calibrated separately for their target datasets or imaging conditions, and we do not evaluate transfer to unseen datasets, acquisition conditions, or particle types without retraining. Our evaluation focuses on particle localization against reference center annotations and does not assess downstream subtomogram averaging (STA).

An important future direction is generalizable particle localization across particle types and imaging conditions with little or no task-specific retraining. Point-set prediction provides a simple output representation for this goal, but whether it improves generalization remains open.

AI Use Statement

Generative AI tools were used as supporting tools for language refinement, coding assistance, and general research discussion. All experiments, analyses, scientific claims, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5561–5569, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Bodla_Soft-NMS_--_Improving_ICCV_2017_paper.html.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Sergio Cruz-León, Tomáš Majtner, Patrick C. Hoffmann, Jan Philipp Kreysing, Sebastian Kehl, Maarten W. Tuijtel, Stefan L. Schaefer, Katharina Geißler, Martin Beck, Beata Turoňová, and Gerhard Hummer. High-confidence 3D template matching for cryo-electron tomography. *Nature Communications*, 15(1):3992, 2024. doi: 10.1038/s41467-024-47839-8.

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.

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

Ashwin Dhakal, Rajan Gyawali, and Jianlin Cheng. TomoSwin3D: a Swin3D Transformer for the Identification and Classification of Macromolecules in 3D Cryo-ET Tomograms. *bioRxiv*, 2026. doi: 10.64898/2026.04.17.719219. URL https://www.biorxiv.org/content/10.64898/2026.04.17.719219v1. Preprint.

Ilja Gubins, Marten L. Chaillet, Gijs van der Schot, M. Cristina Trueba, Remco C. Veltkamp, Friedrich Förster, Xiao Wang, Daisuke Kihara, Emmanuel Moebel, Nguyen P. Nguyen, Tommi White, Filiz Bunyak, Giorgos Papoulias, Stavros Gerolymatos, Evangelia I. Zacharaki, Konstantinos Moustakas, Xiangrui Zeng, Sinuo Liu, Min Xu, Yaoyu Wang, Cheng Chen, Xuefeng Cui, and Fa Zhang. SHREC 2021: Classification in Cryo-electron Tomograms. In *Eurographics Workshop on 3D Object Retrieval*, pp. 5–17. The Eurographics Association, 2021. doi: 10.2312/3dor.20211307.

Ming-Yang Ho and Alberto Bartesaghi. At FullTilt: Real-Time Open-Set 3D Macromolecule Detection Directly from Tilted 2D Projections. arXiv preprint arXiv:2604.10766, 2026. URL https://arxiv.org/abs/2604.10766.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning Non-Maximum Suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4507–4515, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Hosang_Learning_Non-Maximum_Suppression_CVPR_2017_paper.html.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Relation_Networks_for_CVPR_2018_paper.html.

Qinwen Huang, Ye Zhou, and Alberto Bartesaghi. MiLoPYP: self-supervised molecular pattern mining and particle localization in situ. *Nature Methods*, 21(10):1863–1872, 2024. doi: 10.1038/s41592-024-02403-6.

Kaggle Competition Metrics. CZI_CryoET_ 84969. Kaggle Competition Metric Notebook, n.d. URL https://www.kaggle.com/code/metric/czi-cryoet-84969. Version 22; accessed 2026-09-13.

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. doi: 10.1038/s41467-024-46041-0.

E. Moebel, A. Martinez-Sanchez, L. Lamm, R. D. Righetto, W. Wietrzynski, S. Albert, D. Larivière, E. Fourmentin, S. Pfeffer, J. Ortiz, W. Baumeister, T. Peng, B. D. Engel, and C. 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.

A. Peck, J. Hutchings, J. Schwartz, Y. Yu, U. H. Ermel, S. Kandel, D. Kimanius, Z. Zhao, S. Zheng, B. Artley, D. List, Sergio A. Silva, Jr, W. Reade, J. Asuncion, K. Evans, J. Gadling, K. Khandwala, S. McCanny, D. G. McCarthy, J. X. Ni, J. Pourroy, M. Venkatakrishnan, Z. S. Wang, D. A. Agard, C. S. Potter, B. Carragher, K. I. S. Harrington, and M. Paraan. Lessons learned from a Kaggle challenge for particle picking in cryo-electron tomography. *Nature Methods*, 23(9):1908–1916, 2026. doi: 10.1038/s41592-026-03198-4.

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*, 22(9):1819–1823, 2025. doi: 10.1038/s41592-025-02800-5.

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.

J. Schwartz, U. H. Ermel, D. Ji, and Z. Zhao. Octopi: v1.4. Zenodo, 2026. URL `https://doi.org/10.5281/zenodo.18511916`. Software release.

Pranav N. M. Shah, Ruben Sanchez-Garcia, and David I. Stuart. TomoCPT: a generalizable model for 3D particle detection and localization in cryo-electron tomograms. *Acta Crystallographica Section D Structural Biology*, 81(2):63–76, 2025. doi: 10.1107/S2059798325000865.

Simon Wiedemann, Zalan Fabian, Mahdi Soltanolkotabi, and Reinhard Heckel. ProPicker: Promptable segmentation for particle picking in cryogenic electron tomography. *Journal of Structural Biology*, 218(2):108298, 2026. doi: 10.1016/j.jsb.2026.108298.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=gZ9hCDWe6ke`.

A Implementation and Experimental Details

A.1 Implementation and Training

Table 3: **Architecture and optimization settings.** Distances are in detector-input voxels. Dataset-specific training schedules and inference settings are described in the text.

Component	Setting	Value
Detector	Input window / overlap	96^3 / 50%
	Spatial queries / feature scales / decoder layers	64 / 3 / 4
	Optimizer / initial learning rate / weight decay	AdamW / 10^{-4} / 10^{-4}
	Effective batch size / gradient clipping	8 / 0.1
	Warmup / final learning rate	10 epochs / 10^{-6}
	Center / class loss weights	5 / 1
	No-object class weight	0.1
CWQR	Blocks / hidden dimension / attention heads	2 / 128 / 8
	FFN dimension / dropout	512 / 0.1
	Candidate floor τ_0 / maximum neighbors K	0.01 / 32 (including self)
	Graph radius r_g: CZII / SHREC and DeePiCt	9.4614 / 10
	Maximum displacement $\Delta_{\max}$ per axis	4
	Optimizer / learning rate / weight decay	AdamW / 10^{-4} / 10^{-4}
	Batch size / gradient clipping	1 tomogram / 0.1

Training schedule. We first train the window-level detector, then freeze it and train CWQR on cached complete-tomogram candidates. Detector optimization uses a 10-epoch warmup followed by cosine learning-rate decay. CZII and DeePiCt detectors are trained for 120 epochs; the SHREC2021 schedule selected on tomogram 8 is refit for 100 epochs. Augmentations comprise random right-angle rotations around z, axis flips, intensity scaling and shifting, gamma perturbation, and additive Gaussian noise.

On all three datasets, joint coordinate-and-classification training of CWQR uses a budget of 40 epochs and coordinate-loss weight $\lambda_{\text{ref}} = 0.1$. Other architecture and optimization settings appear in Table 3.

Losses. An independent Hungarian assignment is computed at each detector layer. The local objective averages the same loss over all four layers:

$$\mathcal{L}_{\text{local}} = \frac{1}{4} \sum_{\ell=1}^{4} \left(5\mathcal{L}_{\text{center}}^{\ell} + \mathcal{L}_{\text{class}}^{\ell} \right). \tag{7}$$

The center loss averages absolute error over matched coordinate elements, then over windows containing matched particles. Classification uses weighted cross-entropy over all queries, with weight 0.1 for no-object. Windows without matched targets contribute only to classification. The same center and class weights are used in the matching cost in Eq. (2).

Under the tomogram-level assignment, CWQR minimizes

$$\mathcal{L}_{\text{CWQR}} = \mathcal{L}_{\text{class}}^{\text{global}} + \lambda_{\text{ref}} \mathcal{L}_{\text{center}}^{\text{global}}. \tag{8}$$

The classification term is weighted cross-entropy over all pooled candidates, normalized by the sum of target weights, with foreground weight 1 and no-object weight 0.1. The center term averages SmoothL1 over matched coordinate elements after dividing coordinate errors by the class-specific localization tolerance; it is zero when no candidate is matched.

Candidate geometry. For a window of width P and origin $\mathbf{o}_m$, a normalized local prediction $\mathbf{u}_i$ is mapped to $\mathbf{x}_i = \mathbf{o}_m + (P-1)\mathbf{u}_i$. CWQR encodes tomogram position as $2\mathbf{x}_i/(\mathbf{S}-\mathbf{1})-1$, where $\mathbf{S}$ is the tomogram size, and window position as $2\mathbf{x}_i^{\text{loc}}/(P-1)-1$. Boundary proximity is the distance to the nearest window face divided by $P/2$, clipped to $[0,1]$. The graph is built once from detector coordinates and reused by both blocks; same-window non-self candidates are excluded.

Final prediction selection. Class scores are computed from the predicted class logits, using the CWQR-updated logits and coordinates for the complete model and the original detector outputs for the without-CWQR variant. Final predictions are obtained through NMS and score thresholding, using class-aware NMS for CZII and DeePiCt and class-agnostic NMS for SHREC2021. Thresholds are selected on the designated validation or calibration data and fixed before test evaluation: Val40 for CZII, tomogram 8 for SHREC2021, and the two calibration tomograms of each condition and split for DeePiCt.

A.2 Datasets and Evaluation

Dataset partitions. CZII uses seven DS-10440 tomograms for training, 40 DS-10445 tomograms for model and threshold selection, and 364 DS-10446 tomograms for testing (Peck et al., 2025; 2026). External CZII baselines use 121 public tomograms for selection, a superset of our 40 validation volumes. SHREC2021 development uses tomograms 0–7 and selects settings on tomogram 8; models are then retrained on tomograms 0–8 before evaluation on tomogram 9 (Gubins et al., 2021). DeePiCt contains ten VPP and ten defocus tomograms at $13.48\,\text{Å}$ voxel spacing (de Teresa-Trueba et al., 2023). Within each condition, each split uses six training, two calibration, and two test tomograms (Table 4); their roles are mutually disjoint within a split.

Table 4: **DeePiCt training, calibration, and test partitions.** IDs omit the common `TS_` prefix. Each imaging condition is trained and evaluated independently; seeds for splits 0, 1, and 2 are 1729, 1730, and 1731.

Condition	Split	Training	Calibration	Test
VPP	0	0001, 0003, 0004, 0005, 0009, 0010	0006, 0007	0002, 0008
VPP	1	0001, 0002, 0003, 0005, 0008, 0010	0009, 0004	0006, 0007
VPP	2	0002, 0003, 0005, 0006, 0007, 0008	0010, 0001	0009, 0004
Defocus	0	026, 028, 029, 030, 043, 045	034, 037	027, 041
Defocus	1	026, 027, 028, 030, 041, 045	043, 029	034, 037
Defocus	2	027, 028, 030, 034, 037, 041	045, 026	043, 029

Evaluation metrics. CZII uses the official scorer with matching tolerance $0.5R_c$ (Kaggle Competition Metrics, n.d.). Within each class, TP, FP, and FN are pooled across test tomograms:

$$F_4^{\text{weighted}} = \frac{\sum_c w_c F_{4,c}}{\sum_c w_c}, \qquad F_{4,c} = \frac{17\text{TP}_c}{17\text{TP}_c + 16\text{FN}_c + \text{FP}_c}. \tag{9}$$

Weights are $(1, 0, 2, 1, 2, 1)$ for apo-ferritin, beta-amylase, beta-galactosidase, ribosome, thyroglobulin, and virus-like particle. Beta-amylase is retained in classwise reporting but has zero aggregate weight. All methods use the same evaluation annotations with exact duplicate annotation rows removed.

SHREC2021 follows the official class-independent localization protocol; the evaluator configuration excludes 4V94 and submitted vesicles while retaining fiducials (Gubins et al., 2021). DeePiCt uses one-to-one matching within class at a 10-voxel tolerance, maximizing the number of valid pairs. Unmatched and duplicate predictions are false positives. Pooled $F_1 = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$ combines counts across ribosome and FAS and both test tomograms in a split. Score thresholds are selected on calibration data and fixed before testing. DeePiCt error bars summarize variation across the three splits.

Baseline pipelines. DeepETPicker learns center-derived masks and extracts centers with mean-pooling non-maximum suppression (Liu et al., 2024); its SHREC2021 wrapper also performs cross-class deduplication. DeepFinder predicts voxel labels and recovers centers with mean-shift clustering (Moebel et al., 2021). OCTOPI uses a 3D U-Net with center-derived targets and watershed localization (Schwartz et al., 2026).

Qualitative visualization. Figure 5 uses identical crops and grayscale mappings within each comparison. TP/FP/FN labels are assigned in full-tomogram 3D coordinates before cropping and projection, using the CZII class-specific tolerance or the DeePiCt 10-voxel tolerance. Panels (a)–(c) use 15-, 44-, and 26-slice mean projections.

A.3 Controlled Comparisons and Diagnostic Protocols

Shared-backbone formulation comparison. Dense and Direct use identical CZII partitions, center annotations, 96^3 windows, a 3D residual U-Net with three feature scales, matched backbone initialization, augmentations, and training budget. Both train for 120 epochs with 1,024 patches per epoch, effective batch size 8, and 15,360 optimizer updates, and use the epoch-120 checkpoint. Sampling selects class-balanced particle-centered windows with ± 24-voxel jitter for 80% of patches and random windows for the remaining 20%.

The dense head projects each scale to 128 channels and fuses scales by top-down trilinear upsampling and summation, with a $3 \times 3 \times 3$ convolution, group normalization, and GELU at each fusion stage. Six-channel logits are upsampled to the input grid. Class-c targets are Gaussians with $\sigma_c = R_c/3$ and support radius $3\sigma_c$; centers are rounded to integer voxels and overlapping targets are combined by a voxel-wise maximum. Training uses modified focal loss with $\alpha = 2$, $\beta = 4$, normalized by the number of positive centers, without an offset head or auxiliary loss.

At inference, dense probabilities are averaged across overlapping windows, and centers are extracted as $3 \times 3 \times 3$ local maxima without top-K truncation, followed by class-aware NMS at $0.5R_c$. Both formulations use 50% overlap, a candidate score floor of 0.01, and the same coarse-to-fine class-specific threshold search on Val40. Direct uses the detector checkpoints underlying the without-CWQR row of Table 1.

Repeated-observation diagnostic. Figure 4 uses the same within-seed particle cohort at every tested overlap, requiring at least two correct-class observations for each included particle at every overlap. Errors are normalized by class radius R. Within each particle's observation set, the oracle error is the minimum localization error and the confidence-selected error is that of the highest-confidence candidate. Random selection draws uniformly from the same observation set. The selection frequency is the fraction of particles for which a rule selects a best-localized candidate. These quantities describe candidate localization before final selection.

Confidence and localization quality. At 50% overlap, the within-particle rank correlation between classification confidence and localization error is -0.023 ± 0.012 across three seeds. On the same particle cohort, the paired difference in best-observation selection frequency between the highest-confidence rule and a uniform random draw is 0.20 percentage points, with a 95% tomogram-bootstrap interval of $[-0.07, 0.47]$ from 10,000 resamples. The interval includes zero, consistent with confidence providing little advantage for identifying the best-localized observation. The bootstrap summarizes tomogram sampling with the models fixed; the reported standard deviation summarizes variation across training seeds.

Fixed suppression and fusion. All fixed controls use the same frozen CZII candidate predictions. Starting from the highest-scoring candidate, same-class candidates within $0.5R_c$ form a group. Hard NMS and distance-based Soft-NMS retain the anchor coordinate; uniform and score-weighted fusion replace it with the unweighted or confidence-weighted coordinate mean. Soft-NMS applies score decay based on Euclidean point distance (Bodla et al., 2017). Each rule selects its own class-specific thresholds on Val40 before evaluation on Private364.

Learned interaction controls. Table 2(b) uses the three detector checkpoints underlying Table 1. For each seed, Pointwise, Self-only, and full CWQR use the same frozen detector and candidate predictions under matched evaluation settings. Full CWQR results are reused from Table 1. Pointwise processes each candidate independently and has 332,554 parameters. Self-only preserves CWQR's architecture and 468,506 parameters but restricts attention to self. Full CWQR uses the same parameter count and enables cross-window neighbors. Results are evaluated after final selection.

A.4 Detailed Localization Results

Table 5 reports classwise CZII performance, including beta-amylase. The Dense-versus-Direct comparison shows variation across classes: for example, the Gaussian-heatmap baseline performs better on apo-ferritin than the window-level Direct detector. Table 6 separates ribosome and FAS performance under the within-condition 6/2/2 protocol. Table 7 provides the individual SHREC2021 training runs.

Table 5: **Classwise CZII results on 364 test tomograms.** Official F_4 (%); same-backbone Dense and both TomoDETR rows report mean ± sample SD over three training seeds, and external baselines are fixed runs. Beta-amylase has zero aggregate weight. Boldface marks the best score in each column.

Method	Apo.	β-Amy.	β-Gal.	Ribo.	Thyro.	VLP	Weighted F_4
DeepETPicker	80.12	29.81	41.58	78.81	43.73	91.37	60.13
DeepFinder	26.69	21.55	42.49	71.68	46.85	84.50	51.65
OCTOPI	66.44	6.11	60.70	73.71	57.05	90.63	66.61
Dense Gaussian heatmap	82.25 ± 1.08	28.42 ± 1.36	52.33 ± 1.22	85.41 ± 0.27	54.47 ± 1.04	93.12 ± 1.01	67.77 ± 0.54
TomoDETR w/o CWQR	70.15 ± 0.86	31.64 ± 0.90	59.18 ± 0.50	87.32 ± 0.35	60.73 ± 0.48	$\mathbf{93.18 \pm 0.60}$	70.07 ± 0.40
TomoDETR	$\mathbf{83.66 \pm 0.71}$	$\mathbf{40.04 \pm 1.18}$	$\mathbf{61.02 \pm 0.73}$	$\mathbf{87.36 \pm 0.48}$	$\mathbf{62.15 \pm 0.82}$	93.08 ± 0.70	$\mathbf{72.92 \pm 0.45}$

Table 6: **DeePiCt classwise F_1 (%).** Mean and sample SD over three fixed 6/2/2 splits per imaging condition.

Condition	Method	Ribosome $F_1 \uparrow$	FAS $F_1 \uparrow$
VPP	DeepETPicker	82.63 ± 5.31	38.97 ± 9.18
VPP	DeepFinder	78.85 ± 7.90	36.35 ± 10.28
VPP	OCTOPI	74.66 ± 18.57	23.47 ± 10.73
VPP	TomoDETR without CWQR	81.50 ± 4.70	35.15 ± 15.16
VPP	TomoDETR	83.76 ± 5.36	37.43 ± 9.77
DEFOCUS	DeepETPicker	80.50 ± 1.28	35.34 ± 17.56
DEFOCUS	DeepFinder	77.72 ± 1.69	35.33 ± 15.72
DEFOCUS	OCTOPI	74.63 ± 5.67	12.10 ± 1.71
DEFOCUS	TomoDETR without CWQR	79.95 ± 0.68	15.69 ± 1.56
DEFOCUS	TomoDETR	82.86 ± 0.81	13.88 ± 2.56

Table 7: **SHREC2021 repeated training.** Official model-9 localization F_1 (%) across three independently trained systems per method. Boldface marks the best score in each column.

Method	Seed 17	Seed 23	Seed 42	Mean ± SD
DeepFinder	81.56	80.47	76.75	79.59 ± 2.52
DeepETPicker	91.29	92.19	92.70	92.06 ± 0.71
OCTOPI	85.98	85.49	84.74	85.40 ± 0.62
TomoDETR	**93.16**	**93.20**	**92.90**	$\mathbf{93.08 \pm 0.16}$

A.5 Computational Cost

Table 8 compares training and inference on a single RTX 5090 with serial execution. Training the window-level detector takes approximately 63–69 minutes (about 65 minutes centrally), and training CWQR takes 4.69 minutes. The combined training budget is therefore approximately 68–74 minutes, summarized as 1.16 GPU·h in the table. At 50% window overlap, the detector pipeline takes 5.185 seconds per tomogram and CWQR adds 0.148 seconds, giving 5.333 seconds in total and approximately 32.35 minutes for all 364 test tomograms.

Despite having more parameters, the complete system has approximately $2.5\times$, $15.9\times$, and $38.1\times$ faster recorded inference than DeepETPicker, OCTOPI, and DeepFinder, respectively. For one train-

Table 8: **Training and inference cost on CZII.** All methods use a single RTX 5090, with single-GPU serial execution. Training is for one system; test inference covers all 364 tomograms. Times are approximate. The combined budget adds one training run to full-test inference; totals are computed before rounding.

Method	Training (GPU·h)	Inference (s/tomo)	364 tomos (GPU·h)	Combined (GPU·h)	Params. (M)
TomoDETR + CWQR	1.16	5.33	0.54	1.70	7.45
DeepETPicker	0.605	13.45	1.36	1.96	2.43
DeepFinder	3.71	203.36	20.56	24.27	0.903
OCTOPI	7.70	84.51	8.54	16.24	3.26

ing run followed by inference on all 364 test tomograms, its total GPU time is approximately 1.70 GPU·h, a reduction of 13.5%, 89.5%, and 93.0%, respectively. DeepETPicker has the lowest training cost, while TomoDETR with CWQR has the lowest inference and combined costs in this comparison.

The detector and CWQR inference measurements are averaged first over the 364 test tomograms and then over three seeds. The detector-pipeline scope includes image reading and normalization, detector execution, export processing, and NMS. The separately measured refiner stage includes cache loading, graph construction, forward evaluation, and NMS. Both scopes exclude process start-up, warm-up, and persistence of outputs. CWQR adds approximately 2.9% to detector-pipeline inference time.

The six-class CWQR model adds 468,506 parameters to the 6,983,784-parameter detector, giving 7,452,290 parameters in total. In the shared-backbone comparison, the backbone has 5,585,600 parameters; the Dense and Direct heads have 943,750 and 1,398,184 parameters, respectively. Sparse neighborhoods contain at most $K = 32$ entries per candidate, avoiding all-to-all attention over the tomogram.

A.6 Supplementary Mechanism Evidence

Dense foreground and center extraction. Figure 6 traces DeepFinder's predicted foreground and native mean-shift center extraction in two crowded apo-ferritin regions. Here, merged foreground is followed by center-extraction errors, whereas TomoDETR resolves the individual particle centers.

Coordinate and class-score updates. Table 9 independently combines detector or CWQR coordinates with detector or CWQR class labels and scores. Coordinate choice also affects NMS. Updating both outputs achieves the highest score, supporting their joint use in CWQR.

Table 9: **Coordinate and class-score output combinations.** CZII Private364, 50% overlap; weighted F_4 (%), mean $\pm$ sample SD across three seeds in the RTX 5090 output-replay experiment. All rows reuse the same candidate identities and use separately calibrated Val40 thresholds. Boldface marks the best score.

Coordinate source	Class/score source	Weighted $F_4 \uparrow$
Detector	Detector	70.07 ± 0.39
CWQR	Detector	70.72 ± 0.54
Detector	CWQR	70.37 ± 0.40
CWQR	CWQR	$\mathbf{72.43 \pm 0.44}$

B Additional Post-processing and Localization Analyses

We evaluate suppression, window selection, and localization changes using frozen detector and CWQR candidate caches for seeds 17, 23, and 42. CZII uses the M0 reproduction caches on

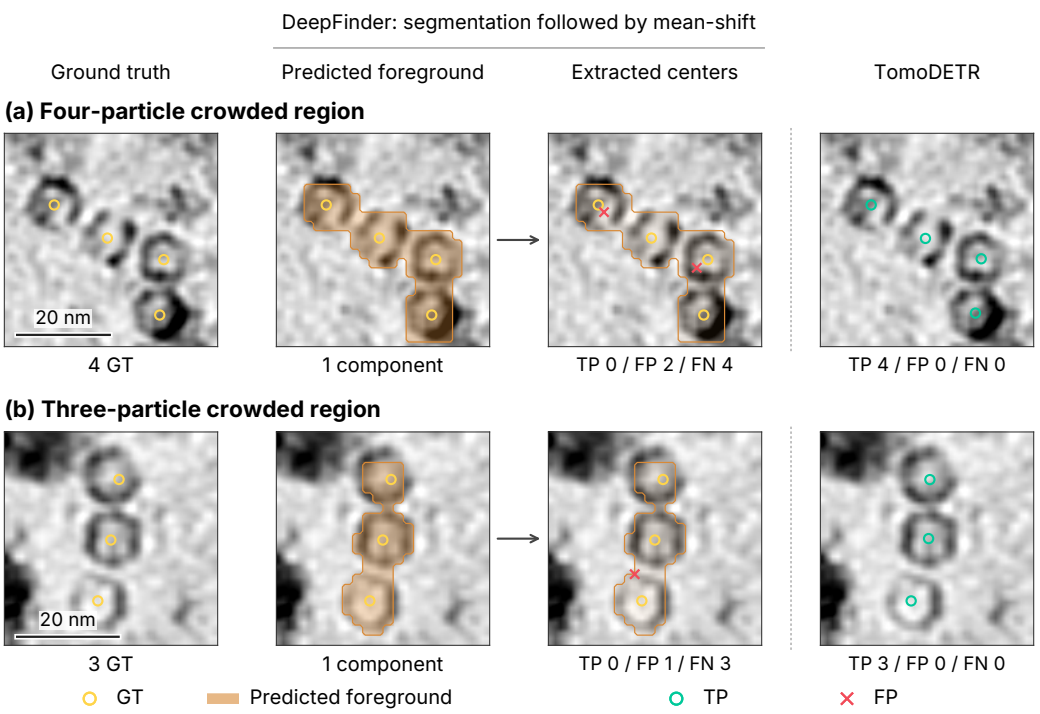

Figure 6: **Foreground merging in crowded apo-ferritin regions.** Yellow: GT; orange: DeepFinder foreground; green: TP; red: FP. Each foreground is one connected component; its mean-shift centers fall outside the 3 nm 3D tolerance. Rows (a,b) use 22- and 16-slice means with shared crops and grayscale mappings across methods.

Private364, with new thresholds calibrated on Val40; SHREC2021 uses model-9 predictions with thresholds calibrated on model 8. All radius configurations are prespecified and reported individually. Table 11 includes the CWQR scores from Table 1 for reference.

Dependence on final NMS. We further examine whether CWQR remains dependent on the final duplicate-suppression step (Table 10). Removing NMS from the local detector reduces weighted F_4 from 70.07 ± 0.40 to 49.29 ± 0.45, a drop of 20.78 points. In contrast, removing NMS after CWQR reduces performance only from 72.92 ± 0.45 to 71.63 ± 0.40, a 1.29-point drop. Notably, CWQR without NMS still exceeds the local detector with NMS by 1.56 points. These results suggest that cross-window refinement performs most of the coordination required to reconcile repeated predictions, while final NMS provides a smaller additional cleanup.

Table 10: **Effect of final NMS on CZII.** Weighted F_4 (%), mean $\pm$ sample SD over three seeds. Removing NMS substantially degrades the local detector, whereas CWQR retains most of its performance without final suppression. Score thresholding is retained in all configurations. Per-class score thresholds are separately calibrated on Val40 for each configuration, then frozen for evaluation on Private364. All runs use an NVIDIA A100 GPU.

Method	Final NMS	Weighted F_4 ↑
Local detector	✓	70.07 ± 0.40
Local detector	×	49.29 ± 0.45
CWQR	✓	$\mathbf{72.92 \pm 0.45}$
CWQR	×	71.63 ± 0.40

B.1 Suppression Radius and Window Selection

Table 11: **Suppression radius and window-selection controls.** Scores (%), mean ± sample SD over seeds 17, 23, and 42. CWQR scores are from Table 1. R_c is the class physical radius and $R_0 = 10$ voxels. Center-window selection is followed by native NMS: $0.5R_c$ on CZII and R_0 on SHREC2021.

	CZII weighted F_4	SHREC2021 localization F_1	
Radius multiplier	R_c	R_c	R_0
0.5	70.07 ± 0.40	76.72 ± 0.83	82.10 ± 3.07
0.75	69.94 ± 0.62	79.94 ± 4.68	86.14 ± 2.08
1	69.89 ± 0.62	81.28 ± 3.85	86.99 ± 2.31
1.25	69.93 ± 0.56	83.60 ± 2.96	88.28 ± 1.88
1.5	69.87 ± 0.57	85.86 ± 2.39	87.63 ± 1.99
Center-window	68.30 ± 0.16	—	88.33 ± 0.81
Class-aware NMS (R_0)	—	—	77.15 ± 2.41
CWQR (Table 1)	$\mathbf{72.92 \pm 0.45}$	$\mathbf{93.08 \pm 0.16}$	

Radius sweep. For CZII, class-aware NMS uses suppression radii from $0.5R_c$ to $1.5R_c$, where R_c is the class physical radius. For SHREC2021, we sweep class-agnostic NMS using both class physical radii and multiples of the native scalar radius $R_0 = 10$ voxels. In the physical-radius sweep, the retained candidate's predicted class determines the suppression radius. Thresholds are calibrated separately for each configuration on validation data.

CZII scores range from 69.87 to 70.07, with no improvement over native $0.5R_c$ NMS. On SHREC2021, the largest score among the prespecified native-radius settings is 88.28 at $1.25R_0$, compared with 86.99 at R_0. These controls quantify the effect of changing suppression strength without learned refinement.

Center-window selection. For each candidate location, we identify all windows that geometrically cover it and retain the candidate only if its source window has the nearest center. Equal-distance ties are retained for subsequent native NMS. This uses actual window boundaries, including edge windows, without GT information. With separately calibrated thresholds, center-window selection yields 68.30 on CZII and 88.33 on SHREC2021 (Table 11).

SHREC2021 suppression type. The SHREC2021 detector endpoint in this analysis already uses class-agnostic NMS at 10 voxels. To isolate suppression type, we keep its candidate coordinates, predicted labels, foreground-sum scores, radius, and threshold search grid fixed, changing only whether suppression is restricted to the same class. Each arm is calibrated on model 8. Class-aware NMS obtains 77.15 ± 2.41, compared with 86.99 ± 2.31 for class-agnostic NMS.

B.2 False-positive Decomposition

CZII categories. We first maximize the number of valid one-to-one prediction–GT matches, then minimize their total distance. Unmatched predictions within the matching tolerance of a same-class GT are *duplicate-like*; those without such support are *unsupported*. Unsupported predictions are partitioned in order: predictions within another class's matching tolerance; remaining predictions within twice the same-class matching tolerance; and far-background predictions. Thus, unsupported errors include class and localization errors as well as background predictions.

Official CZII scoring permits one prediction to cover multiple GT centers. Let B be the difference between official covered-GT counts and the diagnostic one-to-one TP count. The corresponding official FP count is $N_{\text{duplicate}} + N_{\text{unsupported}} - B$. Table 12(a) reports this coverage correction separately and excludes beta-amylase, whose aggregate weight is zero.

Table 12: **False-positive decomposition.** Counts are summed over seeds 17, 23, and 42. CZII excludes the zero-weight beta-amylase class. The coverage correction reconciles CZII one-to-one diagnostic matching with official coverage-based scoring. SHREC2021 separately lists predictions assigned to excluded objects.

Error category	Without CWQR	With CWQR
(a) CZII		
Duplicate-like	4,094	543
Unsupported, total	202,751	135,563
Far background	168,951	126,577
Near other-class GT	2,623	1,805
Near same-class GT, outside tolerance	31,177	7,181
Coverage correction (subtracted)	59	111
(b) SHREC2021		
Duplicate-like	123	19
Background	529	195
Excluded-object assignments	165	35

SHREC2021 categories and observed changes. SHREC2021 categories follow the official occupancy lookup: repeated assignments to an already covered scored object are duplicate-like, assignments to empty occupancy are background, and assignments to excluded objects are recorded separately. Across three seeds, CZII duplicate-like predictions decrease from 4,094 to 543 and unsupported predictions from 202,751 to 135,563. SHREC2021 duplicate-like predictions decrease from 123 to 19 and background predictions from 529 to 195. The changes span multiple error categories, extending beyond removal of duplicate-like predictions.

B.3 Paired Localization Errors

Table 13: **Paired localization errors.** Error is normalized by class physical radius (e/R_c). Entries give without/with-CWQR errors for common GT, and before/after coordinate updates for the same query. N counts paired occurrences pooled over three seeds.

Dataset	Pairing	N	Mean (pre/post)	Median (pre/post)
CZII	Common GT	104,313	0.2207 / 0.1335	0.1983 / 0.1130
CZII	Same query	112,791	0.2411 / 0.1442	0.2075 / 0.1190
SHREC2021	Common GT	4,170	0.4201 / 0.3561	0.3642 / 0.3063
SHREC2021	Same query	4,234	0.3931 / 0.3611	0.3371 / 0.3096

Pairing protocols. The *common-GT* analysis compares final predictions for particles successfully matched by both endpoints. The *same-query* analysis fixes each query and its GT match among the final CWQR true positives, then compares that query's original and refined coordinates against the same GT. The latter measures coordinate changes conditional on CWQR TP selection. Errors are normalized by class physical radius R_c; counts and distributions pool paired occurrences over the three seeds, with repeated GT identities across seeds.

Both pairings show lower mean and median error after refinement (Table 13). For the same-query cohort, mean normalized error decreases from 0.2411 to 0.1442 on CZII and from 0.3931 to 0.3611 on SHREC2021. Figure 7 shows the corresponding empirical distributions. Alongside the FP decomposition, these results document both coordinate improvement and changes in final prediction selection.

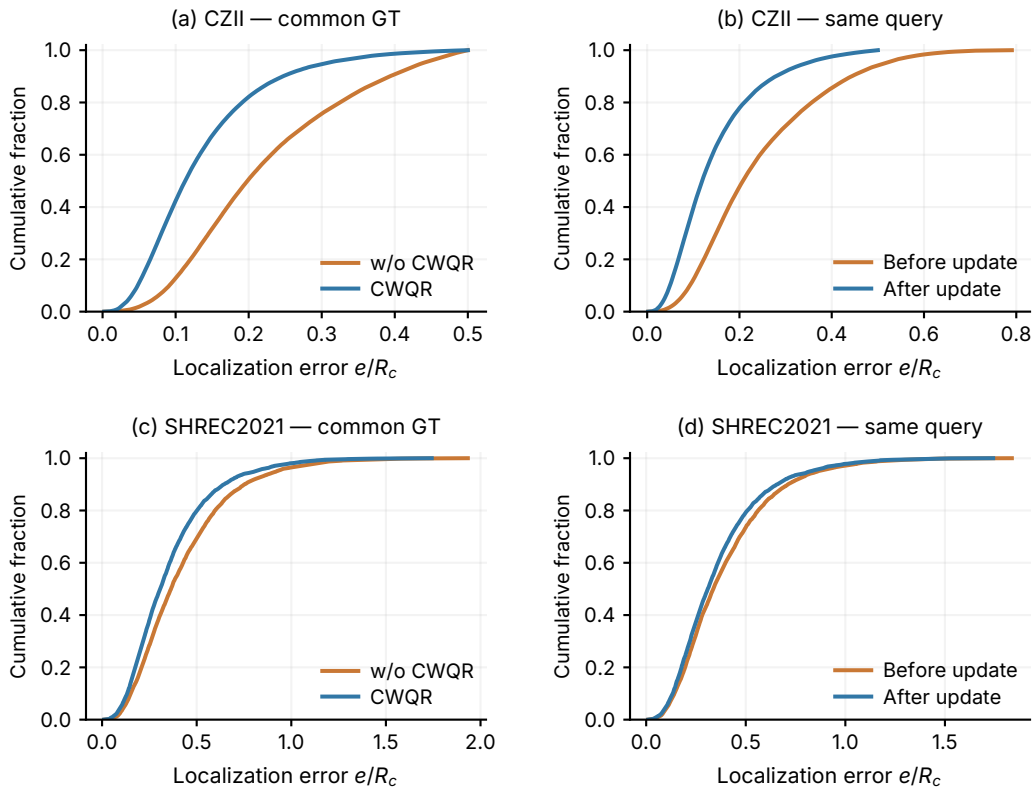

Figure 7: **Paired localization-error distributions.** CZII (a,b) and SHREC2021 (c,d), pooled over three seeds. Left: final-prediction errors on GT matched by both endpoints. Right: original and refined coordinates of the same queries, conditioned on final CWQR TP matches. Curves show empirical cumulative fractions; R_c is the class physical radius. Higher curves at a given error indicate a larger fraction of predictions within that error.

C Window Occupancy and Query Capacity

The local detector uses $Q = 64$ prediction slots per 96^3 window. To relate this budget to particle density, Table 14 reports the number of GT particle centers visible within each window, distinguishing experimentally acquired CZII and DeePiCt data from synthetic SHREC2021 data.

Counting protocol. Centers are counted using the loaders' half-open crop bounds, including all detector classes: six for CZII, fourteen for SHREC2021, and ribosome plus FAS for DeePiCt. Training statistics replay the configured samplers for epochs 0–4, with 1,024 crops per epoch. CZII Train7 and SHREC2021 models 0–8 use seeds 17, 23, and 42; DeePiCt uses the three current 6/2/2 splits within each imaging condition, with the corresponding configured seeds. Each training row therefore summarizes 15,360 sampled crops. Counts describe visible centers before augmentation and any supervision filtering. The replay characterizes the sampling distribution rather than enumerating every crop from a full training run.

Inference-grid statistics enumerate all 96^3 windows at 50% overlap, including windows aligned to volume boundaries. For DeePiCt, each of the ten tomograms per imaging condition is counted once, independently of its role across splits. Overlapping windows are separate counting units.

Experimentally acquired data. The training-sampler means are 5.00 particles per window for CZII, 25.02 for DeePiCt VPP, and 23.76 for DeePiCt defocus. Their 95th percentiles are 11, 45, and 42, respectively; observed maxima are 33, 60, and 56. None of these sampled windows exceeds

Table 14: **GT particle counts per 96^3 window.** Counts include all visible annotated centers for detector classes. P95 is the 95th percentile. Training rows use five-epoch sampler replays; inference rows enumerate the 50%-overlap grid including edge windows.

Dataset / scope	Windows	Mean	P95	Max
(a) Training-sampler replay				
CZII	15,360	5.00	11	33
SHREC2021 (synthetic)	15,360	30.76	56	79
DeePiCt VPP	15,360	25.02	45	60
DeePiCt defocus	15,360	23.76	42	56
(b) Exhaustive inference-grid windows				
CZII / Train7	3,549	2.81	8	30
CZII / Val40	20,280	1.84	6	24
SHREC2021 / models 0–8	9,000	12.77	45	73
SHREC2021 / model 8	1,000	9.69	32	42
SHREC2021 / model 9	1,000	11.23	38	50
DeePiCt VPP / all 10 tomograms	54,150	3.79	28	58
DeePiCt defocus / all 10 tomograms	39,710	5.20	30	54

64 particles. The audited inference grids likewise contain no window above 64. These observations support the query budget for the measured particle densities in the experimentally acquired datasets.

Synthetic data. SHREC2021 has a denser tail: the training replay has mean 30.76, 95th percentile 56, and maximum 79 particles per window; 181 of 15,360 sampled windows (1.18%) exceed 64. On the models 0–8 inference grid, 31 of 9,000 windows (0.34%) exceed 64, whereas the model-9 test grid has maximum 50. A window containing more than 64 GT particles has a local cardinality limit, although this does not directly determine tomogram-level recall because particles also appear in overlapping windows. Thus, 64 covers the observed experimentally acquired cohorts and the SHREC2021 test grid, while a small fraction of dense synthetic training windows exceeds the budget.