CRYOTRACER: A SYNTHETIC-TO-REAL SYSTEM FOR FILAMENT INSTANCE TRACING IN CRYO-EM

Anonymous authors
Paper under double-blind review

ABSTRACT

Particle picking in cryo-electron microscopy is commonly formulated as an object-detection problem in which individual particles are represented by bounding boxes, a natural representation for globular particles but a poor match for filaments that can extend across an entire micrograph. Existing methods reconstruct filament trajectories by linking local predictions in a separate post-processing stage, decoupling detection from instance formation and creating potential ambiguities at gaps and crossings. Directly preserving filament identity is particularly useful for polymorphic specimens, where segments originating from the same filament should remain associated. However, learning such instance-level representations is challenging because expert centerline annotations are scarce and costly to obtain. We introduce CryoTracer, a synthetic-to-real framework that addresses this limitation through procedural micrograph generation, transformer-based polyline prediction, and specimen-specific fine-tuning. Procedural pretraining exposes the model to diverse filament geometries, intersections, and imaging conditions before adaptation using a limited number of fully annotated experimental micrographs. CryoTracer represents filaments as ordered centerline polylines and optimizes their geometry using an orientation-invariant discrete Fréchet objective, enabling consistent learning regardless of tracing direction. Evaluation on established and newly expert-annotated datasets shows that CryoTracer improves instance-level filament recovery over existing particle-detection approaches. Complementary metrics distinguish regional coverage from complete-instance recovery, providing a more detailed assessment of filament-tracing performance. Together, these results establish end-to-end polyline prediction as a practical alternative to detection-and-linking pipelines for recovering complete filament instances with limited micrograph annotation budgets.

1 INTRODUCTION

Cryo-electron microscopy (cryo-EM) uses two-dimensional images of vitrified specimens to determine three-dimensional macromolecular structures. Software packages such as RELION Scheres (2012) and cryoSPARC Punjani et al. (2017) require an initial particle-picking step to identify particles in noisy micrographs. For filamentous specimens, this step must recover centerlines and distinguish individual filaments. Helical reconstruction workflows extract overlapping segments along these traces. Grouping segments by their parent filament allows an entire filament to be inspected or excluded and classification evidence to be combined across its segments. This grouping is particularly useful for specimens containing multiple structural polymorphs that must be separated before reconstruction Frey et al. (2024). For example, α-synuclein is a major component of the protein aggregates associated with Parkinson's disease and dementia with Lewy bodies Spillantini et al. (1997); Schweighauser et al. (2020); Yang et al. (2022); Enomoto et al. (2025). Its fibrils can occur in distinct structural polymorphs Li et al. (2018), and filament-level subset selection has been used when processing mixed α-synuclein polymorph datasets Lövestam et al. (2023). Preserving filament identity during picking requires following each filament through curves, crossings, and low-contrast regions. A fading signal can resemble an endpoint, while a crossing can leave several possible continuations. A tracer must consider the surrounding filament geometry to avoid splitting one filament into fragments or joining different filaments. Preparing real training data requires annotators to mark

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.

each filament's endpoints and place points along its centerline. At crossings or gaps in the visible signal, they must decide which segments belong to the same filament.

Established particle pickers such as crYOLO and Topaz assemble filament traces from local predictions. crYOLO's filament mode requires training on manually annotated filaments Wagner et al. (2020). It connects local bounding boxes into traces without directly predicting a complete ordered centerline. Topaz's filament extension Lövestam & Scheres (2022) selects regions above a threshold in a heatmap of likely particle locations and reduces them to thin lines. It then detects straight segments and merges nearby segments with similar orientations into longer traces. In both workflows, filament continuity is reconstructed through geometric rules with user-selected parameters, which required expert tuning in our comparisons. On α-synuclein, both methods produced more numerous, shorter traces than the annotated filaments, a pattern consistent with fragmentation despite substantial spatial coverage.

To address these limitations, we introduce CryoTracer, a synthetic-to-real system that predicts each filament as an ordered five-point centerline from a full micrograph. Direct centerline prediction makes filament continuity part of the learning task, removing the need to assemble traces from local detections using geometric linking rules. The model adapts RF-DETR Robinson et al. (2026) with curve proposals refined through deformable attention and successive decoder stages. An orientation-invariant discrete Fréchet objective supervises curve matching and regression. To reduce reliance on expert annotations, procedural micrograph generation provides centerlines, endpoints, and filament identities by construction, while varying curvature, density, crossings, width, and background appearance. This enables pretraining on diverse examples without tracing each micrograph manually. Fine-tuning on real annotations then adapts the model to the specimen and acquisition conditions.

Our contributions are as follows.

- We introduce a synthetic-to-real filament-tracing system that learns filament continuity through direct centerline prediction and reduces reliance on expert annotations through procedural pretraining followed by specimen-specific fine-tuning.
- We release filament annotations for EMPIAR-10022 and evaluate on an in-house α-synuclein dataset comprising 325 expert-annotated micrographs from 11 reconstruction sessions.
- We compare CryoTracer with established cryo-EM methods under standard and limited-annotation regimes, using instance-level and spatial-overlap metrics to distinguish complete filament recovery from local coverage.

2 Related work

Set detection and centerline representations. DETR Carion et al. (2020) formulates object detection as prediction of a set of instances with bipartite matching. RF-DETR builds on transformer detection with an architecture designed for efficient inference Robinson et al. (2026). CryoTracer adapts this detector to ordered curves. Centerline representations themselves are not unique to this work: Alonso and Kirkegaard Alonso & Kirkegaard (2023) predict centerlines of slender bodies in microscopy videos using a convolutional network trained on synthetic data. Their method uses temporal context and targets detection and tracking of moving organisms. CryoTracer instead conditions a transformer decoder on polyline references in individual cryo-EM micrographs and fine-tunes on specimen-specific annotations.

Particle picking in cryo-EM. Topaz Bepler et al. (2019) uses positive-unlabeled learning to train particle pickers with sparse annotations. crYOLO's filament mode Wagner et al. (2020) combines learned box detections with tracing post-processing. Other learned particle pickers include EPicker Zhang et al. (2022), which uses CenterNet and continual learning, and CryoTransformer Dhakal et al. (2024), which adapts transformer detection to particle picking. These methods illustrate the range of learned particle-localization approaches. Their particle-level outputs do not directly provide the ordered sequences of points along individual filaments considered here.

Filament tracing. FibrilFinder Thurber et al. (2021) uses classical image processing to identify amyloid fibrils and prepare coordinates for helical reconstruction. RELION 5.1's Amy-

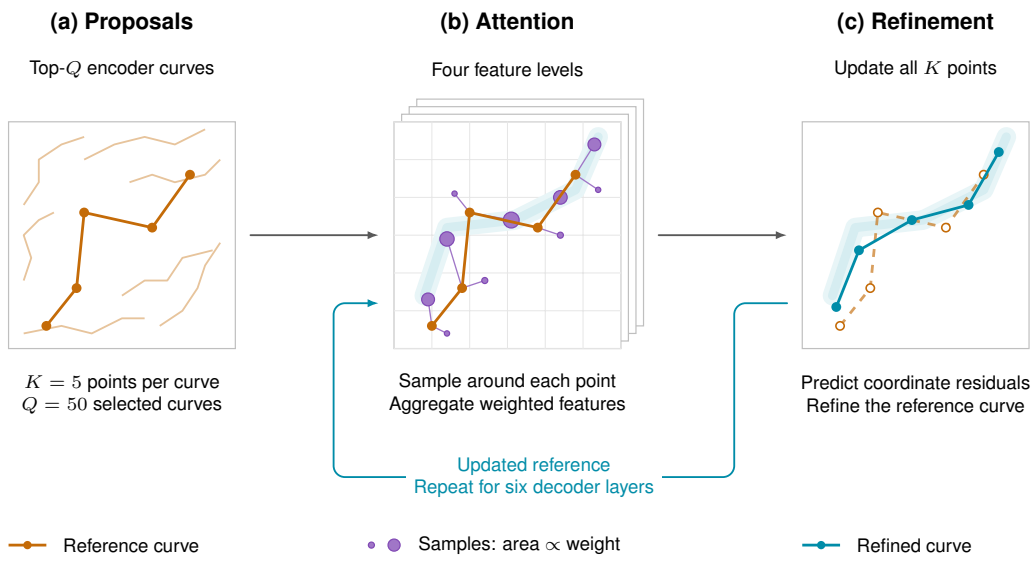

Figure 1: **Polyline prediction and iterative refinement in CryoTracer.** (a) The $Q = 50$ highest-scoring encoder proposals provide initial reference polylines with $K = 5$ ordered points. (b) Cross-attention updates each query using encoder features around its current polyline (orange). Purple markers indicate sampling locations, with marker area proportional to attention weight. (c) The updated query refines the polyline, and the resulting curve (teal) guides attention in the next decoder layer. This process progressively refines the predictions.

Tracer Lövestam et al. (2026) combines the characteristic 4.75 Å repeat signal of amyloid cross-β sheets with U-Net segmentation. This repeat reflects the spacing of β-strands along the filament axis and produces a directional Fourier signal. Filament centerlines are then extracted by thresholding and skeletonizing the predicted probability maps.

3 METHOD

Below, we present *CryoTracer*, a method for detecting filament instances in high-resolution, noisy micrographs. First, we describe our proposed architecture, which modifies RF-DETR to enable a polyline-based representation of detected structures. Next, we explain a two-stage training paradigm that addresses the problem of limited annotated real data. The key components of our method are illustrated in Figure 1.

Architecture overview. Our detection model consists of an encoder $\mathcal{E}_\theta$ and a decoder $\mathcal{D}_\phi$, parameterized by θ and ϕ, respectively. The decoder contains several learnable projection modules (ψ, Φ, Γ). Although these modules have separate parameter subsets, we denote all of them by ϕ for simplicity, emphasizing that they jointly constitute the decoder $\mathcal{D}_\phi$.

Feature extraction. Given an input micrograph $\mathbf{X} \in \mathbb{R}^{1\times W\times H}$, the encoder $\mathcal{E}_\theta$ extracts a set of feature maps $\{\mathbf{F}_l\}_{l=1}^{L}$ from L selected intermediate layers. Each spatial location of each feature map $\mathbf{F}_l$ defines one candidate polyline proposal.

Initial proposal construction. For a spatial location (x, y) in feature map $\mathbf{F}_l$, we obtain a latent proposal representation

$$\mathbf{z}_i = \psi_\phi(\mathbf{F}_{l,x,y}), \tag{1}$$

where, for simplicity, we use a single index i to enumerate proposals across all feature maps. Each proposal is additionally associated with a center $\mathbf{c}_i$, corresponding to the normalized spatial position (x, y), and a scale $\mathbf{r}_i$ determined by the feature-map level l. Coarser feature maps correspond to larger spatial scales.

From the latent representation $\mathbf{z}_i$, we predict a confidence score $\gamma_i = \Phi_\phi(\mathbf{z}_i)$. We retain the $Q = 50$ proposals with the highest confidence scores. This exceeds the observed filament counts in both datasets (means: 20.3 and 5.9), while limiting decoder cost. Given the image sizes and filament widths, substantially denser scenes would also increase overlap and tracing ambiguity. For each retained proposal, an MLP Γ_ϕ predicts offsets for K ordered polyline points:

$$\mathbf{D}_i = \Gamma_\phi(\mathbf{z}_i), \qquad \mathbf{D}_i \in \mathbb{R}^{K\times 2}. \tag{2}$$

The initial polyline is then constructed as

$$\mathbf{P}_i = (\mathbf{c}_i + \tanh(\mathbf{D}_{i,k}) \odot \mathbf{r}_i)_{k=1}^{K}. \tag{3}$$

Each row $\mathbf{D}_{i,k}$ specifies the offset of the k-th polyline point relative to the proposal center. The $\tanh$ function bounds the predicted offsets, while $\mathbf{r}_i$ scales them according to the spatial scale associated with the corresponding feature-map level. The resulting K ordered points form the initial polyline prediction $\mathbf{P}_i$.

Iterative polyline refinement. Each initial polyline is predicted independently from a single feature-map location $\mathbf{F}_{l,x,y}$. Consequently, the initial predictions have two limitations. First, multiple proposals may correspond to the same filament, while other filaments may receive no proposal. Second, predicting the entire polyline from a single feature representation may result in inaccurate localization or shape estimates. Following RF-DETR, we address both limitations through an iterative refinement procedure depicted in Figure 1 b) and c).

For each decoder query slot i, we initialize the query content $\mathbf{q}_i^{(0)}$ from a learned query embedding shared across images. The selected encoder proposal $\mathbf{P}_i$ supplies the initial reference geometry, $\mathbf{P}_i^{(0)} = \mathbf{P}_i$.

At decoder layer t, the queries first interact through self-attention,

$$\tilde{\mathbf{q}}_i^{(t)} = \mathrm{SelfAttn}_\phi\left(\mathbf{q}_i^{(t-1)}, \{\mathbf{q}_j^{(t-1)}\}_{j=1}^{Q}\right), \tag{4}$$

allowing information to be exchanged between proposals. This interaction helps the decoder reason jointly about the Q candidate objects rather than refining each proposal independently.

Next, cross-attention aggregates information from all encoder feature levels $\{\mathbf{F}_l\}_{l=1}^{L}$ into query $\tilde{\mathbf{q}}_i^{(t)}$. The current polyline $\mathbf{P}_i^{(t-1)}$ provides the reference locations for sampling features around each of its points:

$$\mathbf{q}_i^{(t)} = \mathrm{CrossAttn}_\phi\left(\tilde{\mathbf{q}}_i^{(t)}, \{\mathbf{F}_l\}_{l=1}^{L}; \mathbf{P}_i^{(t-1)}\right). \tag{5}$$

Consequently, the query is updated using features sampled around its current polyline across all feature levels.

Finally, the updated query predicts a correction to the current polyline,

$$\mathbf{P}_i^{(t)} = \mathrm{Refine}_\phi\left(\mathbf{P}_i^{(t-1)}, \mathbf{q}_i^{(t)}\right). \tag{6}$$

The refined polyline $\mathbf{P}_i^{(t)}$ and query $\mathbf{q}_i^{(t)}$ are passed to the next decoder layer, progressively improving the predictions. After the final layer, the resulting polylines $\{\hat{\mathbf{P}}_i\}_{i=1}^{Q}$ form the decoder outputs.

Training objective. Since we are representing objects with polylines, a simple regression from off-the-shelf DETR methods is not applicable. Instead, to compare the predicted polyline $\hat{\mathbf{P}}_i$ with a ground truth polyline $\mathbf{Y}_j$, we use the Fréchet Distance $d_F(\hat{\mathbf{P}}_i, \mathbf{Y}_j)$, which measures the spatial distance between two polylines rather than exact point-to-point alignment. To make the distance invariant to point direction (forward vs backward), we define it as

$$\bar{d}_F(\hat{\mathbf{P}}_i, \mathbf{Y}_j) = \min\{d_F(\hat{\mathbf{P}}_i, \mathbf{Y}_j), d_F(\mathrm{rev}(\hat{\mathbf{P}}_i), \mathbf{Y}_j)\}, \tag{7}$$

where rev reverses the order of points. To find a proper match between the sets of predictions $\{\hat{\mathbf{P}}_i\}_{i=1}^{Q}$ and ground truths $\{\mathbf{Y}_j\}_{j=1}^{N}$, we use Hungarian matching to find a minimum-cost one-to-one assignment that balances geometric distance $\bar{d}_F$ and prediction confidence. Predictions not

selected by the assignment are optimized as no-object (polyline confidence γ is minimized for them via the cross entropy loss term $\mathcal{L}_{CE}$). For matched polylines, we optimize the following criterion

$$\mathcal{L}_F = \frac{1}{\max(N,1)} \sum_{j=1}^{N} \bar{d}_F \left(\hat{\mathbf{P}}_{\pi(j)}, \mathbf{Y}_j \right), \tag{8}$$

where $\pi(j)$ is the index of the prediction matched to the j-th ground truth polyline. The cross-entropy loss $\mathcal{L}_{CE}$ also encourages high confidence for matched predictions. Combining classification and geometry gives

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_F \mathcal{L}_F, \tag{9}$$

with geometric loss weight $\lambda_F = 10$.

We apply Equation 9 independently to the final decoder predictions and the Q selected initial encoder proposals, obtaining $\mathcal{L}_{\text{dec}}$ and $\mathcal{L}_{\text{enc}}$, respectively. Each prediction set has its own Hungarian assignment. Encoder supervision directly trains the proposal heads because the initial coordinates are detached before decoder refinement. We therefore compute $\mathcal{L}_{\text{enc}}$ from an undetached copy of the selected proposals.

The decoder and encoder losses are combined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{dec}} \mathcal{L}_{\text{dec}} + \lambda_{\text{enc}} \mathcal{L}_{\text{enc}}, \tag{10}$$

where $\lambda_{\text{dec}} = \lambda_{\text{enc}} = 1$.

Synthetic pretraining. Starting from RF-DETR weights, we pretrain on procedural micrographs with exact centerline labels. The generator varies filament geometry, crossings, density, and image appearance, and includes empty fields. Procedural generation supplies diverse examples without additional manual tracing, providing supervision for learning filament shape and instance separation before adaptation to real data.

Real data finetuning. The synthetic checkpoint initializes finetuning on annotated real micrographs. This stage uses the available centerline annotations to adapt the pretrained representation to specimen-specific filament morphology and experimental image appearance. Both stages use the same architecture and objective, with all parameters trainable. Data generation and preprocessing are specified in Appendix A; optimization and checkpoint selection are specified in Appendix B.

Inference and non-maximum suppression. At inference, we retain predictions with confidence at least 0.5 and apply polyline non-maximum suppression (NMS) to remove duplicate filament detections. Duplicates are identified by similar unoriented directions, substantial projected overlap, and small perpendicular separation, preserving the geometry of the higher-confidence curve. We use the same thresholds across both datasets and evaluation regimes (Appendix B.2).

4 Experiments

We evaluate how accurately CryoTracer and established cryo-EM methods recover complete filaments on EMPIAR-10022 and α-synuclein. We use a 70/15/15% train/validation/test split in the standard regime.

A typical reconstruction session contains about 2,000 micrographs, but experts may annotate only 30 to 50 to adapt a tracing model to a new specimen. We study this limited-annotation setting with a practical regime, using a 15/15/70% train/validation/test split.

We report precision, recall, and F1 for filament detection, with Box IoU and coverage as complementary overlap metrics. We detail the splits, baseline configurations, and prediction conversions in Appendix B.3.

4.1 Evaluation metrics

We measure complete-instance recovery using orientation-invariant discrete Fréchet matching between predicted and annotated centerlines. Precision, recall, and F1 are computed per micrograph

and averaged across the evaluation set. Matching is sensitive to fragmentation and endpoint errors. Our primary metric, mF1, averages these macro F1 scores equally over distance tolerances of 25, 50, 100, 150, and 200 Å. We compute mRecall and mPrecision by the same averaging procedure. Main results summarize five training seeds for confirmation, while ablations use only one seed. Appendix tables give per-tolerance scores and sample standard deviations.

Box IoU and coverage provide complementary measures of spatial agreement between predicted and annotated filament regions. Pick counts and mean pick lengths in nanometres describe predictions and ground truth on the same evaluation images. Appendix B.4 specifies matching, resampling, and overlap definitions.

4.2 Main results

CryoTracer achieves the highest mF1 on both datasets and in both evaluation regimes (Tables 1 and 2).

EMPIAR-10022. CryoTracer achieves mF1 of 0.63 and 0.54 in the standard and practical regimes, compared with 0.08 and 0.09 for Topaz, the strongest baseline by mF1. CryoTracer leads both overlap metrics in the standard regime, whereas Topaz leads them in the practical regime.

Table 1: Filament detection on EMPIAR-10022 in the standard and practical regimes. mF1, mRecall, and mPrecision average per-micrograph scores over five distance tolerances (Section 4.1). Box IoU measures spatial overlap; coverage is the fraction of annotated area recovered. Pick count and mean polyline length (nm) summarize filament traces, with ground truth as a reference. CryoTracer values are means over five training seeds. Bold marks the highest unrounded score per metric and regime; pick statistics are descriptive.

Method	mF1	mRecall	mPrecision	Box IoU	Coverage	Pick count	Mean pick length (nm)
Standard regime: ground truth						347	181
CryoTracer (ours)	**0.63**	**0.63**	**0.63**	**0.84**	**0.90**	337	176
Topaz	0.08	0.13	0.06	0.82	0.88	947	56
crYOLO	0.03	0.03	0.03	0.66	0.69	329	122
Practical regime: ground truth						1,578	190
CryoTracer (ours)	**0.54**	**0.53**	**0.56**	0.81	0.86	1,474	184
Topaz	0.09	0.16	0.06	**0.86**	**0.92**	4,607	61
crYOLO	0.02	0.02	0.02	0.65	0.68	1,559	118

α-synuclein. CryoTracer achieves mF1 of 0.40 and 0.34 in the standard and practical regimes, compared with 0.11 and 0.09 for the RELION Amyloid model, the strongest baseline by mF1. CryoTracer leads Box IoU in both regimes; Topaz has higher practical-regime coverage.

Implications for filament processing. On standard α-synuclein, Topaz produces 1,659 picks with mean pick length 33 nm, compared with 272 annotated filaments averaging 203 nm. CryoTracer produces 278 picks averaging 176 nm. The shorter, more numerous Topaz picks are consistent with fragmentation, although false positives and endpoint errors can also contribute. Thus, substantial spatial overlap can coexist with poor recovery of individual filaments. Figure 7 illustrates these differences in the picked-filament length distributions.

Figures 2 and 3 show selected predictions, with additional galleries in Appendix C.

4.3 Ablations

In this section, we study how synthetic initialization, curve supervision, and architecture affect filament detection. We compare each variant with the main model in Table 3. We describe the training changes in Appendix B.1 and report results at each matching tolerance in Appendix C.

Synthetic initialization. With identical real-data partitions, 20,000 synthetic steps followed by 5,000 real-data steps yield standard-regime mF1 of 0.62 on EMPIAR-10022 and 0.44 on α-

Table 2: Filament detection on α-synuclein in the standard and practical regimes. mF1, mRecall, and mPrecision average per-micrograph scores over five distance tolerances (Section 4.1). Box IoU measures spatial overlap; coverage is the fraction of annotated area recovered. Pick count and mean polyline length (nm) summarize filament traces, with ground truth as a reference. CryoTracer values are means over five training seeds. Bold marks the highest unrounded score per metric and regime; pick statistics are descriptive.

Method	mF1	mRecall	mPrecision	Box IoU	Coverage	Pick count	Mean pick length (nm)
Standard regime: ground truth						272	203
CryoTracer (ours)	**0.40**	**0.43**	**0.39**	**0.75**	**0.84**	278	176
Topaz	0.01	0.04	0.01	0.59	0.80	1,659	33
crYOLO	0.01	0.01	0.01	0.60	0.66	540	60
RELION Amyloid	0.11	0.11	0.11	0.59	0.63	342	94
Practical regime: ground truth						1,364	203
CryoTracer (ours)	**0.34**	**0.36**	**0.33**	**0.69**	0.80	1,374	181
Topaz	0.02	0.05	0.01	0.63	**0.83**	7,891	34
crYOLO	0.00	0.00	0.00	0.58	0.62	2,498	61
RELION Amyloid	0.09	0.09	0.09	0.61	0.66	1,914	88

Table 3: CryoTracer ablations of synthetic initialization, curve supervision, and architecture. Synthetic pretraining uses up to 20,000 steps; real-data training uses up to 5,000 steps unless marked 25k. Bold labels identify the main model configuration. Higher mF1 and Box IoU are better. Training and architecture details appear in Appendix B.1.

					EMPIAR-10022		α-synuclein	
Experiment	Training	Distance	Points	Pyramid levels	mF1	Box IoU	mF1	Box IoU
Standard regime								
Main	**Synthetic + real**	**Fréchet**	**5**	**4**	0.62	0.85	0.44	0.75
Training strategy	Synthetic only	Fréchet	5	4	0.00	0.05	0.00	0.01
	Real only (5k)	Fréchet	5	4	0.16	0.49	0.18	0.51
	Real only (25k)	Fréchet	5	4	0.59	0.83	0.39	0.71
Distance metric	Real only (5k)	Chamfer	5	4	0.00	0.42	0.00	0.35
	Synthetic + real	Chamfer	5	4	0.03	0.71	0.01	0.69
Number of points	Synthetic + real	Fréchet	3	4	0.65	0.85	0.41	0.73
	Synthetic + real	Fréchet	10	4	0.60	0.83	0.42	0.74
Feature pyramid	Synthetic + real	Fréchet	5	1	0.56	0.81	0.36	0.73
Original attention	Synthetic + real	Fréchet	5	4	0.57	0.82	0.33	0.69
Practical regime								
Main	**Synthetic + real**	**Fréchet**	**5**	**4**	0.53	0.80	0.36	0.70
Training strategy	Synthetic only	Fréchet	5	4	0.00	0.05	0.00	0.01
	Real only (5k)	Fréchet	5	4	0.07	0.45	0.02	0.28
	Real only (25k)	Fréchet	5	4	0.40	0.68	0.11	0.44

synuclein. Training directly from RF-DETR for 25,000 real-data steps gives 0.59 and 0.39, respectively. The corresponding practical-regime scores are 0.53 and 0.36 with synthetic initialization, versus 0.40 and 0.11 for real-only training. The observed advantage is therefore larger in the practical regime.

With only 5,000 real-only steps, mF1 is 0.16 and 0.18 in the standard regime and 0.07 and 0.02 in the practical regime. Synthetic initialization also improves spatial overlap relative to both real-only

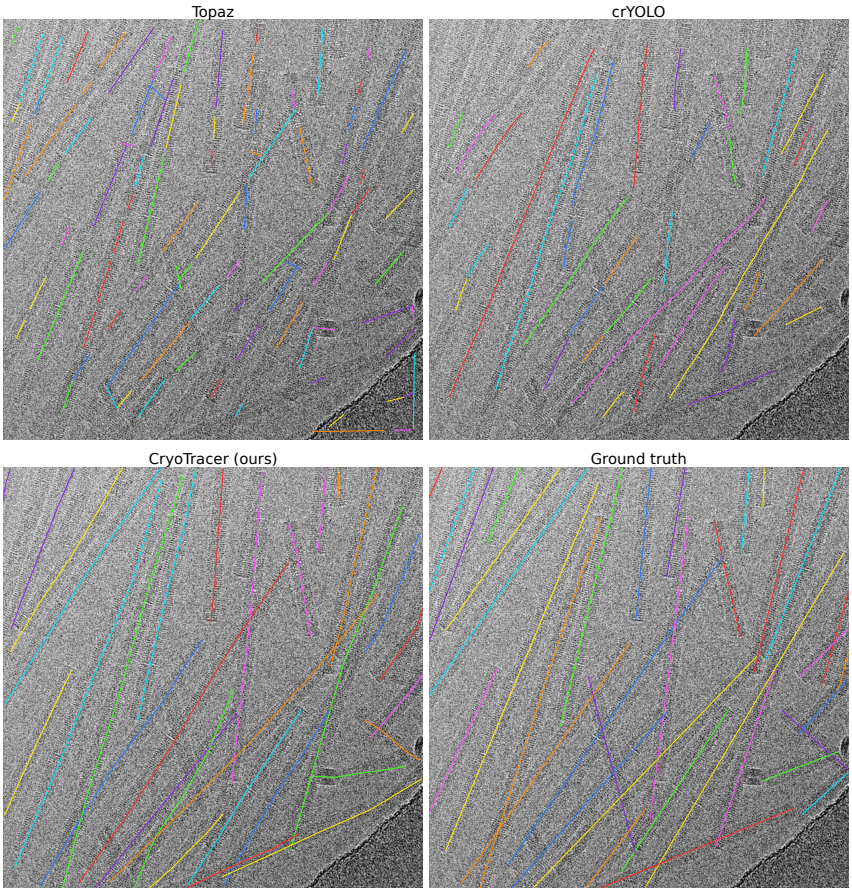

Figure 2: EMPIAR-10022, standard regime. One selected micrograph from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.

budgets. Real-data adaptation remains necessary, as the frozen synthetic-only model yields mF1 of 0.00 in all four settings.

Curve supervision. Replacing Fréchet distance with Chamfer distance in both Hungarian matching and regression reduces standard-regime mF1 from 0.62 to 0.03 on EMPIAR-10022 and from 0.44 to 0.01 on α-synuclein after synthetic pretraining. Box IoU remains 0.71 and 0.69, respectively. Substantial spatial overlap can therefore coexist with poor complete-instance agreement. This is consistent with Chamfer's insensitivity to point order.

Architecture. Restoring original RF-DETR box attention, while retaining five-point polyline outputs, lowers standard-regime mF1 to 0.57 on EMPIAR-10022 and 0.33 on α-synuclein, compared with 0.62 and 0.44 for the main model. Box IoU decreases to 0.82 and 0.69, respectively.

Reducing the feature pyramid from four levels to one lowers standard-regime mF1 from 0.62 to 0.56 on EMPIAR-10022 and from 0.44 to 0.36 on α-synuclein. Point-count effects vary by dataset. Three points score highest on EMPIAR-10022, whereas five score highest on α-synuclein. The simpler three-point representation may suit EMPIAR-10022's relatively clean, straight filaments. However, its two line segments limit the representation of more strongly curved filaments, so this advantage may not generalize to more complex geometries. Ten points improve neither mF1 nor Box IoU over five.

We find that synthetic pretraining followed by real-data adaptation improves mF1 over real-only training, with larger gains when annotations are limited. Our results also support Fréchet supervi-

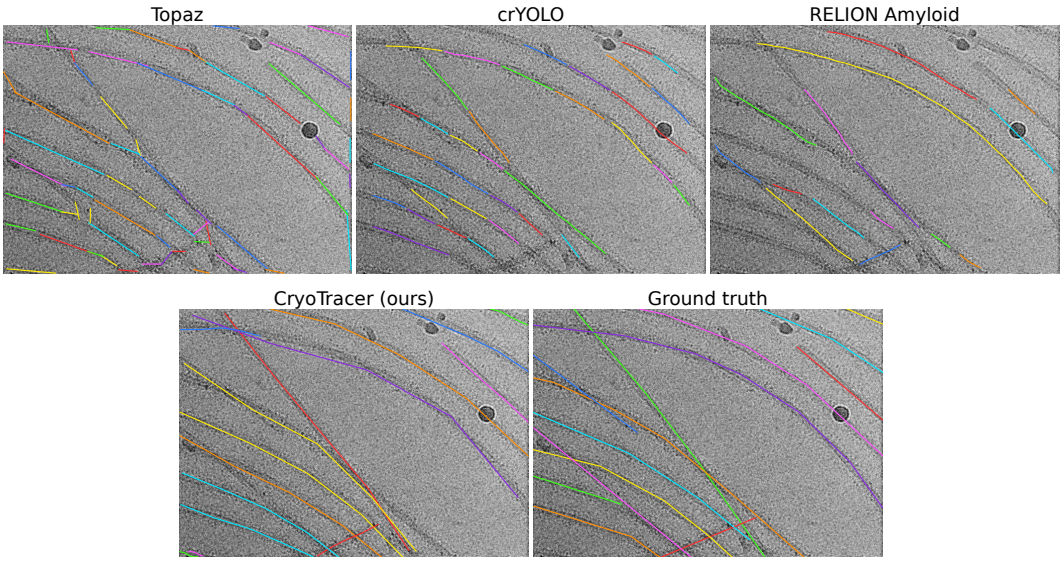

Figure 3: α-synuclein, standard regime. One selected micrograph from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.

sion, the modified attention design, and a four-level feature pyramid. Although the preferred point count depends on the dataset, we find five points to offer a reasonable compromise across the two datasets.

4.4 Limitations

CryoTracer remains sensitive to ambiguous filament geometry. At crossings and overlaps, predicted traces can connect distinct filaments or recover only part of an instance. The fixed five-point representation may also limit the fidelity of traces with complex curvature. Synthetic pretraining reduces reliance on real training annotations but does not eliminate it. Our results cover two datasets with specimen-specific fine-tuning. Generalization to unseen specimens and acquisition conditions requires further evaluation. The next step is to test whether improved filament recovery translates into more effective downstream classification and reconstruction.

5 Conclusion

CryoTracer learns to trace individual cryo-EM filaments by predicting ordered centerline polylines, making geometry and instance identity explicit in the model output. The framework combines procedural synthetic pretraining, orientation-invariant Fréchet supervision, and specimen-specific fine-tuning. On EMPIAR-10022 and α-synuclein, it achieves higher complete-instance mF1 than Topaz and crYOLO in both standard and limited-annotation regimes, and outperforms the RELION Amyloid baseline on α-synuclein. Ablations support the value of synthetic initialization, with larger gains when real training annotations are scarce. Together with the overlap metrics, these results show why filament tracing should be evaluated for both spatial coverage and instance continuity. We release new EMPIAR-10022 centerline annotations to support further evaluation.

AI USE STATEMENT

In this work, we did not use generative AI tools to generate synthetic datasets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, writing proofs, and translation were not applicable to this work.

We used generative AI tools for creating or editing software code, drafting parts of the research paper and polishing writing. We have reviewed all AI-assisted work. LLM-generated code was verified and the research paper was reviewed and edited by the authors to ensure accuracy, clarity, and consistency with the reported research.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide a ZIP archive containing the PyTorch Lightning training implementation, Hydra configurations for the main experiments and ablations, and a DVC pipeline for data splitting, baseline training, and prediction and evaluation of CryoTracer and the comparison methods. Appendices A and B document data preprocessing and synthetic generation, training settings, hardware, checkpoint selection, random seeds, and the evaluation protocol.

REFERENCES

Albert Alonso and Julius B Kirkegaard. Fast detection of slender bodies in high density microscopy data. *Communications Biology*, 6(1):754, 2023.

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0575-8. URL https://doi.org/10.1038/s41592-019-0575-8.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. Cryotransformer: A transformer model for picking protein particles from cryo-em micrographs. *Bioinformatics*, 2024. doi: 10.1093/bioinformatics/btae109. URL https://doi.org/10.1093/bioinformatics/btae109.

Masahiro Enomoto, Ivan Martinez-Valbuena, Shelley L. Forrest, Xiaoxiao Xu, Renato P. Munhoz, Jun Li, Ekaterina Rogaeva, Anthony E. Lang, and Gabor G. Kovacs. Lewy-msa hybrid fold drives distinct neuronal α-synuclein pathology. *Communications Biology*, 8(1), 2025. ISSN 2399-3642. doi: 10.1038/s42003-025-08355-7. URL http://dx.doi.org/10.1038/s42003-025-08355-7.

Lukas Frey, Dhiman Ghosh, Bilal M. Qureshi, David Rhyner, Ricardo Guerrero-Ferreira, Aditya Pokharna, Witek Kwiatkowski, Tetiana Serdiuk, Paola Picotti, Roland Riek, and Jason Greenwald. On the ph-dependence of α-synuclein amyloid polymorphism and the role of secondary nucleation in seed-based amyloid propagation. *eLife*, 12:RP93562, 2024. doi: 10.7554/eLife.93562.4. URL https://elifesciences.org/articles/93562.

Simon A. Fromm, Tanmay A.M. Bharat, Arjen J. Jakobi, Wim J.H. Hagen, and Carsten Sachse. Seeing tobacco mosaic virus through direct electron detectors. *Journal of Structural Biology*, 189(2):87–97, 2015. ISSN 1047-8477. doi: 10.1016/j.jsb.2014.12.002. URL https://www.sciencedirect.com/science/article/pii/S1047847714002652.

Binsen Li, Peng Ge, Kevin A. Murray, Phorum Sheth, Meng Zhang, Gayatri Nair, Michael R. Sawaya, Woo Shik Shin, David R. Boyer, Shulin Ye, David S. Eisenberg, Z. Hong Zhou, and

Lin Jiang. Cryo-em of full-length α-synuclein reveals fibril polymorphs with a common structural kernel. *Nature Communications*, 9:3609, 2018. doi: 10.1038/s41467-018-05971-2. URL https://www.nature.com/articles/s41467-018-05971-2.

Sofia Lövestam and Sjors H. W. Scheres. High-throughput cryo-em structure determination of amyloids. *Faraday Discussions*, 240:243–260, 2022. doi: 10.1039/D2FD00034B.

Sofia Lövestam, David Li, Jane L. Wagstaff, Abhay Kotecha, Dari Kimanius, Stephen H. McLaughlin, Alexey G. Murzin, Stefan M. V. Freund, Michel Goedert, and Sjors H. W. Scheres. Disease-specific tau filaments assemble via polymorphic intermediates. *Nature*, 625(7993):119–125, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06788-w. URL http://dx.doi.org/10.1038/s41586-023-06788-w.

Sofia Lövestam, Jenny Shi, David Li, Kiarash Jamali, and Sjors H.W. Scheres. Cryo-em image processing of amyloid filaments in relion-5.1. *bioRxiv*, 2026. doi: 10.64898/2026.03.17.712386. URL https://www.biorxiv.org/content/early/2026/03/17/2026.03.17.712386.

Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. RF-DETR: Neural Architecture Search for Real-Time Detection Transformers. In *International Conference on Learning Representations*, 2026.

Sjors H.W. Scheres. Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519–530, December 2012. ISSN 1047-8477. doi: 10.1016/j.jsb.2012.09.006. URL http://dx.doi.org/10.1016/j.jsb.2012.09.006.

Manuel Schweighauser, Yang Shi, Airi Tarutani, Fuyuki Kametani, Alexey G. Murzin, Bernardino Ghetti, Tomoyasu Matsubara, Taisuke Tomita, Takashi Ando, Kazuko Hasegawa, Shigeo Murayama, Mari Yoshida, Masato Hasegawa, Sjors H. W. Scheres, and Michel Goedert. Structures of α-synuclein filaments from multiple system atrophy. *Nature*, 585(7825):464–469, May 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2317-6. URL http://dx.doi.org/10.1038/s41586-020-2317-6.

Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M.-Y. Lee, John Q. Trojanowski, Ross Jakes, and Michel Goedert. α-Synuclein in lewy bodies. *Nature*, 388(6645):839–840, August 1997. ISSN 1476-4687. doi: 10.1038/42166. URL http://dx.doi.org/10.1038/42166.

Kent R Thurber, Yi Yin, and Robert Tycko. Automated picking of amyloid fibrils from cryo-em images for helical reconstruction with relion. *Journal of structural biology*, 213(2):107736, 2021.

Thorsten Wagner, Luca Lusnig, Sabrina Pospich, Markus Stabrin, Fabian Schönfeld, and Stefan Raunser. Two particle-picking procedures for filamentous proteins: Sphire-cryolo filament mode and sphire-striper. *Acta Crystallographica Section D: Structural Biology*, 76(7):613–620, 2020. doi: 10.1107/S2059798320007342.

Yang Yang, Yang Shi, Manuel Schweighauser, Xianjun Zhang, Abhay Kotecha, Alexey G. Murzin, Holly J. Garringer, Patrick W. Cullinane, Yuko Saito, Tatiana Foroud, Thomas T. Warner, Kazuko Hasegawa, Ruben Vidal, Shigeo Murayama, Tamas Revesz, Bernardino Ghetti, Masato Hasegawa, Tammaryn Lashley, Sjors H. W. Scheres, and Michel Goedert. Structures of α-synuclein filaments from human brains with lewy pathology. *Nature*, 610(7933):791–795, 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05319-3. URL http://dx.doi.org/10.1038/s41586-022-05319-3.

Xinyu Zhang, Tianfang Zhao, Jiansheng Chen, Yuan Shen, and Xueming Li. Epicker is an exemplar-based continual learning approach for knowledge accumulation in cryoem particle picking. *Nature Communications*, 13(1):2468, 2022.

CryoTracer
Appendix

A DATA

We use two real cryo-electron microscopy datasets and a procedurally generated dataset. All annotations are represented as ordered polylines, where each polyline describes the centerline of one filament.

EMPIAR-10022. EMPIAR-10022 is a public dataset of tobacco mosaic virus micrographs Fromm et al. (2015). The original helical annotations were of insufficient quality for reliable filament-level evaluation, so we reannotated the micrographs with ordered filament centerlines. We make these new annotations publicly available. Our annotated dataset contains 109 images at a native resolution of 4096×4096 pixels and a pixel size of $1.06\,\text{Å/px}$.

α-synuclein. The α-synuclein dataset is a private, in-house collection of 11 reconstruction sessions comprising 325 cryo-EM micrographs annotated by a cryo-EM expert. All sessions were collected by the same person in the same laboratory. Each micrograph has a native resolution of 5760×4092 pixels and a pixel size of $0.65\,\text{Å/px}$.

Real-image preprocessing. Micrographs are downsampled by a factor of 4 using bicubic interpolation with antialiasing (`torchvision` resize with `antialias=True`). EMPIAR-10022 images are mapped from 4096×4096 to 1024×1024, and α-synuclein images from 5760×4092 to 1440×1023. Polyline coordinates are divided by 4 and normalized by the resulting image dimensions. Pixel-size metadata is multiplied by 4. Training applies horizontal and vertical flips independently with probability 0.5 and uniformly sampled in-plane rotations. A radial Fourier low-pass mask at $0.2\,\text{Å}^{-1}$ filters the images, which are standardized over valid pixels and clipped to $[-5, 5]$. Validation and test preprocessing uses the same filtering and normalization without geometric augmentation.

Synthetic pretraining data. To expose the model to a broad range of filament geometries before adaptation to real data, we construct a procedural micrograph dataset. Each example is defined in a native 4096×4096 coordinate system and rendered directly at 1024×1024 resolution. Non-empty images contain between one and 30 curved filaments. Filaments are generated from smooth parametric curves, clipped at the image boundary, and include explicit crossing configurations. Ten percent of examples are empty, allowing the model to learn from background-only micrographs.

We randomize both imaging and specimen properties. The native pixel size is sampled uniformly from 0.65 to $1.5\,\text{Å/px}$, corresponding to an effective input sampling of 2.6-$6.0\,\text{Å/px}$ at the rendered resolution. Filament widths are sampled from 100-$200\,\text{Å}$. Appearance variation includes both contrast polarities, spatial background gradients, Gaussian blur, and additive noise. These perturbations are intended to encourage transfer across datasets rather than reproduce a particular microscope contrast-transfer function.

A.1 POLYLINE SAMPLING

Synthetic and real annotations use the same five-point target conversion. For annotations with more than five points, we select five vertices at uniformly spaced indices, rounded down, including both endpoints. For annotations with at most five points, we retain every annotated point and add any needed points along the existing segments in proportion to their lengths. Thus, short annotations are only augmented rather than resampled, so no annotated bend is inadvertently omitted or replaced by a chord.

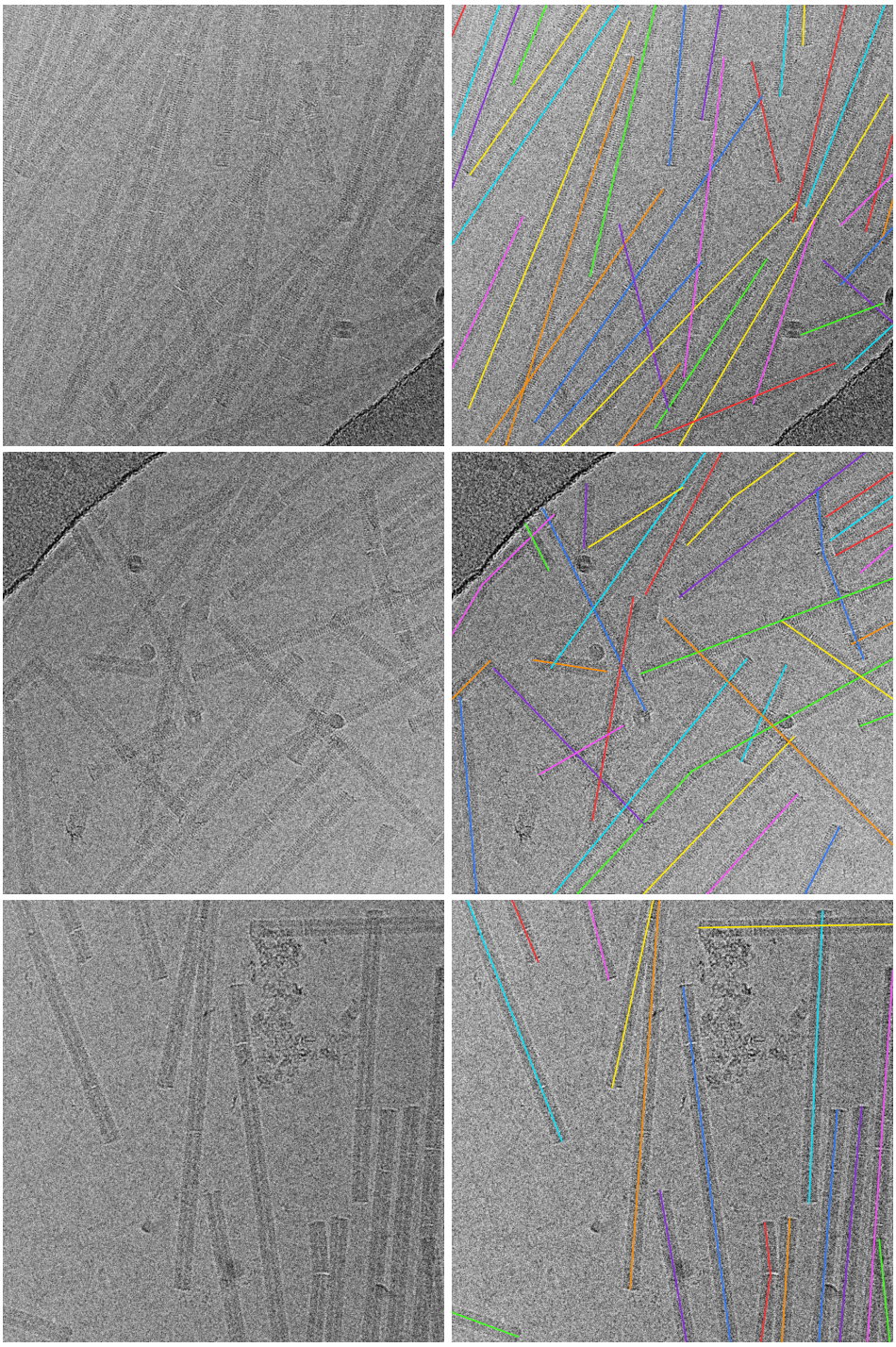

Figure 4: EMPIAR-10022 dataset examples. Each row pairs an image (left) with a low-pass-filtered image and its ground-truth centerlines (right). Colors distinguish annotated filament instances.

For evaluation, predicted and reference polylines are sampled uniformly by arc length to 32 points. This gives curves with different original point counts a common representation for matching and metric computation while preserving a denser approximation of local bends.

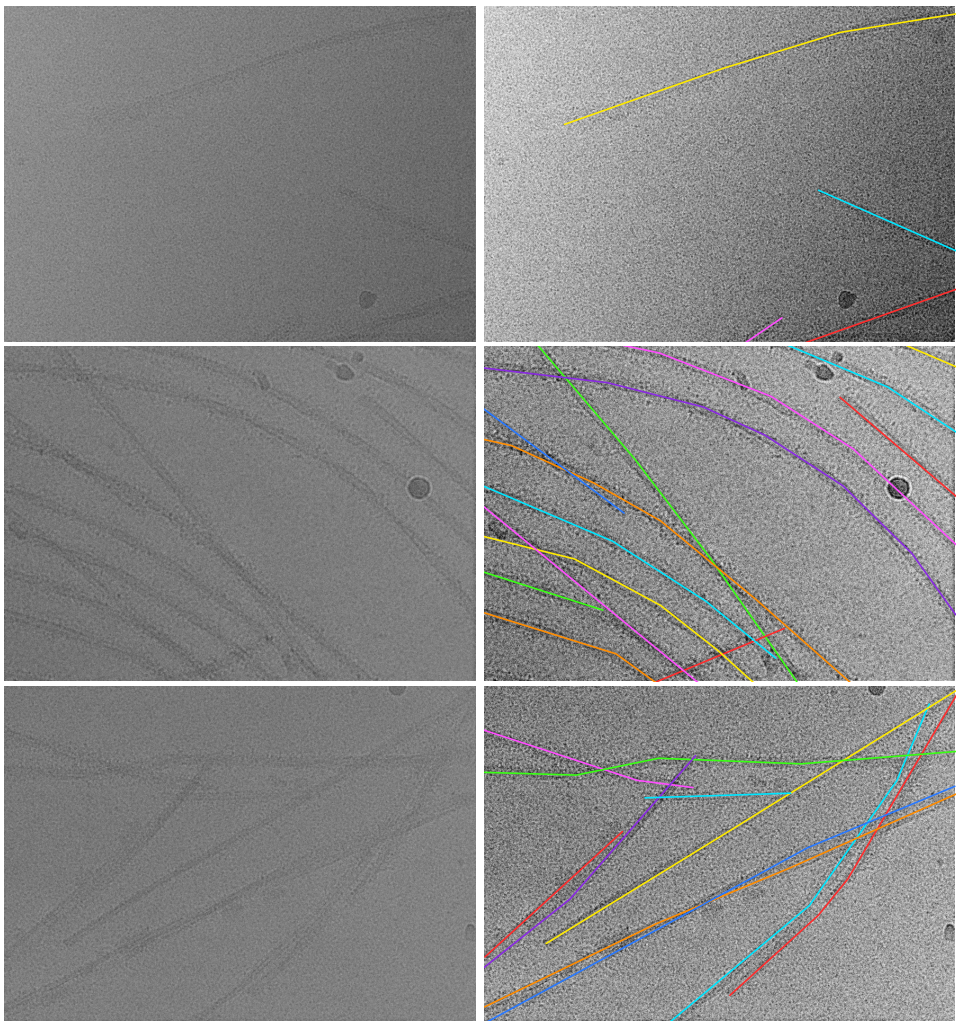

Figure 5: α-synuclein dataset examples. Each row pairs an image (left) with a low-pass-filtered image and its ground-truth centerlines (right). Colors distinguish annotated filament instances.

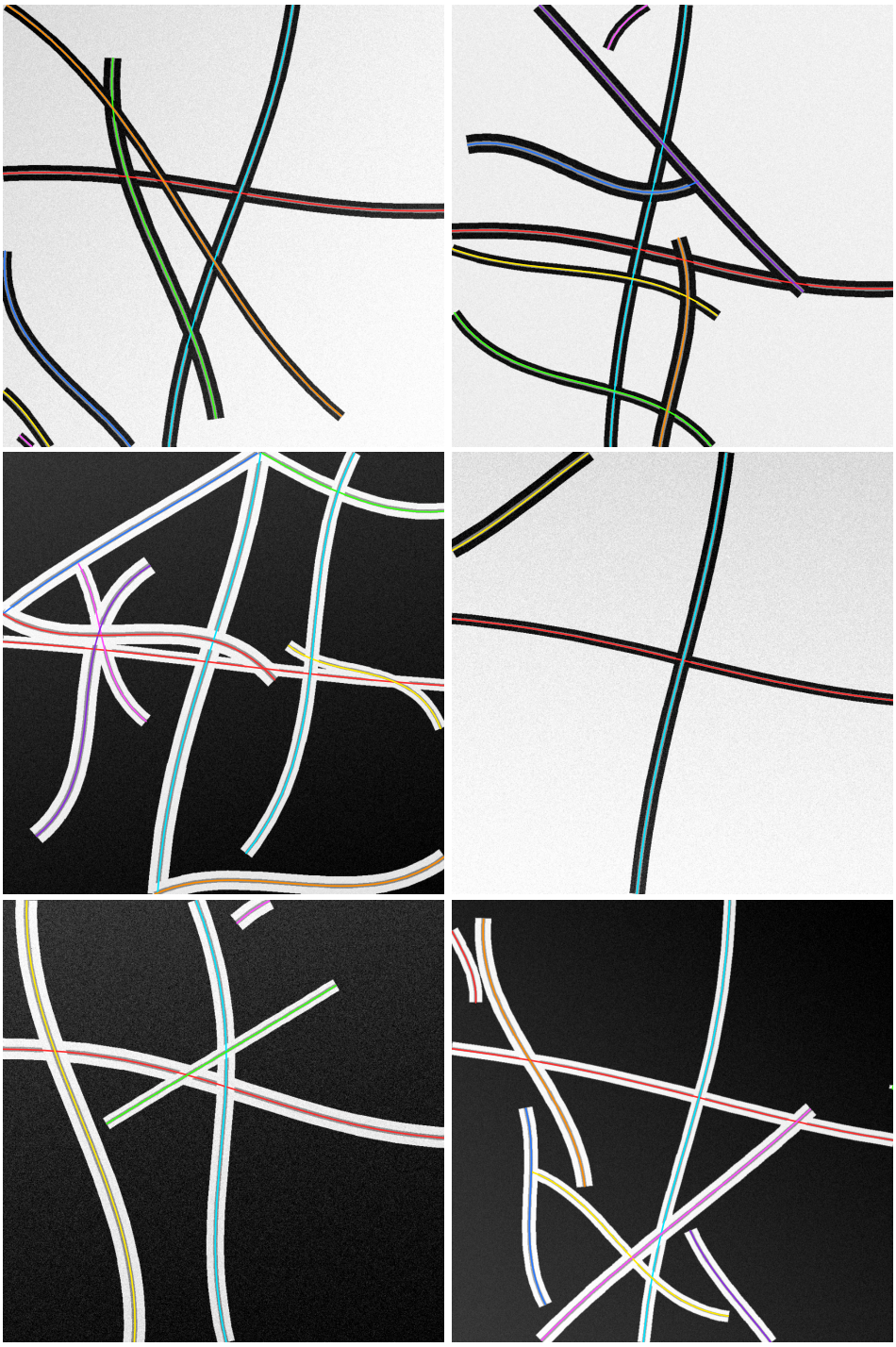

Figure 6: Synthetic dataset examples. Six synthetic micrographs are shown with their ground-truth centerlines. Colors distinguish annotated filament instances.

B TRAINING AND EVALUATION DETAILS

B.1 TRAINING CONFIGURATION

Hardware and optimization. Each CryoTracer run uses one NVIDIA H100 GPU, bfloat16 mixed precision, gradient checkpointing, and a batch size of 64 micrographs. AdamW uses weight decay 10^{-4} and initial learning rates of 10^{-4} for the detector and 10^{-5} for the backbone. A cosine schedule decays both towards 10^{-6} over each stage, with gradient clipping at 1.0. The complete model remains trainable throughout.

Training stages and validation. Synthetic pretraining starts from RF-DETR weights adapted to single-channel input and runs for at most 20,000 steps, with validation every 500 steps. The selected synthetic model initializes dataset-specific fine-tuning for at most 5,000 steps, with validation every 100 steps on the corresponding real validation partition. Fine-tuning starts a new optimizer and learning-rate schedule. Within each seed, the same selected synthetic initialization is used for both real datasets and their standard and practical regimes.

Checkpoint selection. At each stage, we retain the checkpoint minimizing the final decoder's classification loss plus weighted polyline regression loss on the validation set. Encoder and intermediate decoder terms are excluded from this selection criterion. Final evaluation uses the selected real-data checkpoint. Test metrics do not influence model selection.

Loss implementation. Batch regression losses are normalized by the total target count, clamped to at least one. Distributed training synchronizes this normalization across workers. Squared distances are floored at 10^{-12} before taking square roots. The hard minimum and maximum operations in the Fréchet recurrence are piecewise differentiable. Automatic differentiation follows the selected branches with the framework's subgradient convention at ties. Reversal invariance applies to the objective, not to the decoder's internal features.

Ablation configurations. The ablation reference uses the main model's configuration, including synthetic pretraining, Fréchet matching and regression, auxiliary decoder supervision, five polyline points, and a four-level feature pyramid. The no-synthetic condition starts directly from RF-DETR weights with a 5,000-step real-data training budget. Additional real-only controls train for 25,000 steps with the cosine schedule extended over that budget, matching the total optimizer steps of 20,000-step synthetic pretraining plus 5,000-step fine-tuning. These controls retain the same model, data partitions, and validation-selection criterion in both regimes. Optimizer-step matching does not imply identical wall time or computational cost. The Chamfer condition changes both matching and regression in pretraining and fine-tuning, using its own validation-selected synthetic initialization. The architecture ablations use three or ten polyline points, or a single feature-pyramid level. Each is pretrained and fine-tuned with the corresponding architecture.

The original-attention ablation restores RF-DETR's box-based deformable cross-attention and box positional encoding. At each decoder layer, the reference polyline supplies its axis-aligned enclosing box. This variant samples four locations per head and feature level, compared with four per polyline point (20 total) in the main model. It retains five-point outputs, polyline refinement and losses, and the four-level feature pyramid, using its own 20,000-step synthetic pretraining and 5,000-step real-data fine-tuning. The comparison therefore changes reference geometry, positional encoding, and sampling budget together.

B.2 INFERENCE AND PARTICLE EXPORT

Duplicate suppression. For filament evaluation, polyline NMS identifies duplicates using their unoriented directions, projected overlap relative to the shorter projected trace, and perpendicular separation. We use a 12° angular threshold, minimum overlap ratio 0.8, and maximum separation $140\,\text{Å}$ across both datasets and regimes. Suppression favors the higher-confidence curve and preserves its geometry. Fragments remain separate.

RELION particle export. The optional export samples confidence-filtered curves uniformly by arc length, with physical spacing equal to the user-specified helical rise times the number of asym-

metric units between samples. Centers whose extraction boxes would extend outside the micrograph are omitted. Each particle records its native coordinates, parent filament identifier, local tangent-angle prior, and cumulative distance along the original trace. Export-specific duplicate suppression and optional crossing exclusion operate on sampled particles. Removed samples leave gaps without resetting track distances. Filament identifiers restart within each micrograph. The resulting STAR file includes the associated optics and CTF metadata from the supplied micrographs STAR file and supports subsequent particle extraction in RELION. CryoTracer writes particle metadata, and RELION performs image extraction.

B.3 Evaluation protocol

Repeated runs and aggregation. Main-text CryoTracer results report means over five training runs (seeds 42-46). Appendix tables add the sample standard deviation. Ablations use seed 42, and each baseline has one run. Within each seed, instance scores are averaged over the same evaluation micrographs within each regime. mF1, mRecall, and mPrecision additionally average over distance tolerances before aggregation across runs. Pick counts and mean pick lengths are also computed separately for each seed before reporting their mean and sample standard deviation. Standard deviations describe variation across seeds, not across tolerances, micrographs, or individual filament lengths.

Data regimes. We evaluate two data regimes. In the standard regime, EMPIAR-10022 is split into 76 training, 16 validation, and 17 test micrographs, while the α-synuclein dataset is split into 227 training, 48 validation, and 50 test micrographs. These correspond approximately to a 70/15/15% train/validation/test split.

The practical regime reduces the amount of annotated data available for training to approximately 15% of each dataset while retaining approximately 15% for validation and leaving about 70% for evaluation. We obtain this split by rotating the standard partitions. The validation set is used for training, the test set for validation, and the training set for evaluation. This results in $16/17/76$ train/validation/evaluation micrographs for EMPIAR-10022 and $48/50/227$ for α-synuclein.

Partitions are defined at the micrograph level. For α-synuclein, reconstruction sessions overlap across training, validation, and evaluation partitions in both regimes, while individual micrographs remain disjoint within each regime. This protocol evaluates held-out micrographs from the collection described in Appendix A.

Comparison methods. Topaz and crYOLO use the same supervised partitions as CryoTracer. Their configuration required substantial manual tuning by a cryo-EM expert, including dataset-specific picking and filament-tracing settings, which were then fixed for evaluation. Topaz is trained for 10 epochs with an expected particle count of 200, using downsampling factors of four and six and extraction thresholds of -3 and -4 for EMPIAR-10022 and α-synuclein, respectively. crYOLO uses filament mode with expert-selected box sizes of 400 and 512 pixels, approximately twice the filament width to include surrounding image context. The filament-width argument (`-fw`) is set to the corresponding box size, with tracing box distances of 40 and 51 pixels, respectively. These expert-selected picking and tracing settings are independent of the tighter physical filament widths used for spatial-overlap evaluation below. CryoTracer uses the same optimization recipe and inference thresholds across both datasets and regimes, without corresponding dataset-specific manual tuning. Its specimen-specific model-weight fine-tuning remains part of training. This comparison of tuning effort is qualitative. The experiments compare configured systems without matched hyperparameter-search or compute budgets, and do not quantify expert time or the number of tuning trials.

RELION Amyloid. RELION 5.1's AmyTracer pipeline combines an amyloid-specific Fourier signal with a pretrained neural network and subsequent tracing Lövestam et al. (2026). It is therefore a learned, pretrained reference, even when used without dataset-specific training. We evaluate the official pretrained AmyTracer v2.0 checkpoint only on α-synuclein. The RELION-5.1.0 release supplies inference code but no AmyTracer training or fine-tuning code, so we perform no dataset-specific weight updates. The same checkpoint is evaluated on the original test partition in the standard regime and the original training partition in the practical regime, matching the evalu-

ation micrographs used by the other methods within each regime. We also retain the RELION-5.1 default tracing parameters. Its amyloid cross-beta signal makes it applicable to α-synuclein, but not the tobacco mosaic virus filaments in EMPIAR-10022.

Inference settings and representations. CryoTracer uses a confidence threshold of 0.5 and the polyline NMS described in Appendix B.2. Topaz predictions are interpreted as consecutive endpoint pairs. crYOLO box centers and RELION traces are grouped by filament identifier. These conversions preserve the methods' exported representations. Short fragments can therefore achieve high spatial overlap while failing to match a complete annotated instance. All coordinates are mapped to native pixels and converted to physical units for evaluation.

B.4 Evaluation metric definitions

Instance matching. Predicted and reference curves are sampled uniformly by arc length to 32 points. The distinction from the five-point training construction is explained in Appendix A.1. Following the notation in Section 3, let $\mathcal{P} = \{\hat{\mathbf{P}}_i\}$ and $\mathcal{G} = \{\mathbf{Y}_j\}$ denote the predicted and ground truth polylines, respectively, and define

$$D_{ij} = \bar{d}_F(\hat{\mathbf{P}}_i, \mathbf{Y}_j), \tag{11}$$

using the orientation-invariant distance in Equation 7, measured in Å. At tolerance τ, only pairs with $D_{ij} \leq \tau$ are eligible. Hungarian assignment with this tolerance constraint first maximizes the number of valid one-to-one matches, then minimizes their total distance. Matched pairs are true positives. Unmatched predictions and ground truth polylines are false positives and false negatives, respectively. The per-image counts give

$$\begin{aligned} P_\tau &= \frac{\mathrm{TP}_\tau}{\mathrm{TP}_\tau + \mathrm{FP}_\tau}, \\ R_\tau &= \frac{\mathrm{TP}_\tau}{\mathrm{TP}_\tau + \mathrm{FN}_\tau}, \\ F1_\tau &= \frac{2P_\tau R_\tau}{P_\tau + R_\tau}. \end{aligned} \tag{12}$$

We report the mean of each per-image score, with undefined ratios assigned zero, including images with no predictions or annotations. Macro F1 is therefore the mean of per-image F1 scores. With the fixed tolerance set $\mathcal{T} = \{25, 50, 100, 150, 200\}$ Å, our primary metric is

$$\mathrm{mF1} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{1}{M} \sum_{i=1}^{M} F1_{i,\tau}, \tag{13}$$

where M is the number of evaluation micrographs. mRecall and mPrecision use the same equal-weight averaging of the corresponding per-image scores. mF1 averages F1 directly and is not the harmonic mean of mPrecision and mRecall. Scores at individual tolerances are secondary diagnostics.

Spatial overlap. For each image, let $B_\mathcal{P}$ and $B_\mathcal{G}$ be the unions of axis-aligned boxes swept along the predicted and reference curves. The metric box width is independent of the particle-extraction box. We use 170 native pixels for EMPIAR-10022, corresponding to the established 180 Å TMV diameter, and 215 pixels for α-synuclein, corresponding to an assumed 140 Å filament width. These tight widths measure agreement around the filament itself and are shared by all methods within each dataset. Define

$$\begin{aligned} \mathrm{BoxIoU} &= \frac{|B_\mathcal{P} \cap B_\mathcal{G}|}{|B_\mathcal{P} \cup B_\mathcal{G}|}, \\ \mathrm{Coverage} &= \frac{|B_\mathcal{P} \cap B_\mathcal{G}|}{|B_\mathcal{G}|}. \end{aligned} \tag{14}$$

Box IoU penalizes both missing and excess predicted area. Coverage measures the fraction of reference area recovered. We average Box IoU over all images, assigning one when both masks are empty, and coverage over images with non-empty ground truth. These union-based metrics do not measure instance identity and depend on the specified box width.

C ADDITIONAL RESULTS

Table 4: Detailed filament-detection results on EMPIAR-10022. Precision (P), recall (R), and F1 are averaged over micrographs at each distance tolerance (in Å). Box IoU and coverage measure spatial overlap. Pick count and mean polyline length (nm) summarize filament traces, with ground truth as a reference. CryoTracer values are mean ± sample standard deviation over five seeds, with statistics computed separately per seed. Bold marks the highest unrounded score per metric and regime. Pick statistics are descriptive. Regimes use different evaluation sets.

Regime	Method	P@25	R@25	F1@25	P@50	R@50	F1@50	P@100	R@100	F1@100	P@150	R@150	F1@150	P@200	R@200	F1@200	Box IoU	Coverage	Pick count	Mean pick length (nm)
standard	CryoTracer (ours)	**0.25 ± 0.04**	**0.25 ± 0.04**	**0.25 ± 0.04**	**0.55 ± 0.02**	**0.55 ± 0.02**	**0.55 ± 0.02**	**0.74 ± 0.02**	**0.74 ± 0.02**	**0.74 ± 0.02**	**0.80 ± 0.02**	**0.79 ± 0.02**	**0.79 ± 0.02**	**0.82 ± 0.02**	**0.82 ± 0.01**	**0.82 ± 0.02**	**0.84 ± 0.01**	**0.90 ± 0.01**	337 ± 4	176 ± 3
	Topaz	0.01	0.02	0.01	0.02	0.04	0.03	0.06	0.13	0.08	0.09	0.22	0.13	0.11	0.25	0.15	0.82	0.88	947	56
	crYOLO	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.04	0.03	0.04	0.06	0.05	0.06	0.66	0.69	329	122
	Ground truth	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	347	181
practical	CryoTracer (ours)	**0.14 ± 0.04**	**0.13 ± 0.04**	**0.14 ± 0.04**	**0.45 ± 0.06**	**0.43 ± 0.05**	**0.44 ± 0.05**	**0.67 ± 0.05**	**0.64 ± 0.04**	**0.65 ± 0.04**	**0.74 ± 0.04**	**0.71 ± 0.03**	**0.72 ± 0.04**	**0.78 ± 0.04**	**0.75 ± 0.03**	**0.76 ± 0.03**	0.81 ± 0.01	0.86 ± 0.01	1,474 ± 34	184 ± 5
	Topaz	0.01	0.04	0.02	0.03	0.09	0.05	0.06	0.17	0.09	0.09	0.24	0.13	0.10	0.28	0.15	**0.86**	**0.92**	4,607	61
	crYOLO	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.03	0.04	0.65	0.68	1,559	118
	Ground truth	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1,578	190

Table 5: Detailed filament-detection results on α-synuclein. Precision (P), recall (R), and F1 are averaged over micrographs at each distance tolerance (in Å). Box IoU and coverage measure spatial overlap. Pick count and mean polyline length (nm) summarize filament traces, with ground truth as a reference. CryoTracer values are mean $\pm$ sample standard deviation over five seeds, with statistics computed separately per seed. Bold marks the highest unrounded score per metric and regime. Pick statistics are descriptive. Regimes use different evaluation sets.

Regime	Method	P@25	R@25	F1@25	P@50	R@50	F1@50	P@100	R@100	F1@100	P@150	R@150	F1@150	P@200	R@200	F1@200	Box IoU	Coverage	Pick count	Mean pick length (nm)
standard	CryoTracer (ours)	**0.02 ± 0.01**	**0.03 ± 0.02**	**0.03 ± 0.02**	**0.18 ± 0.04**	**0.20 ± 0.04**	**0.19 ± 0.04**	**0.47 ± 0.05**	**0.51 ± 0.06**	**0.48 ± 0.05**	**0.62 ± 0.04**	**0.67 ± 0.04**	**0.63 ± 0.04**	**0.68 ± 0.03**	**0.73 ± 0.01**	**0.69 ± 0.02**	**0.75 ± 0.01**	**0.84 ± 0.01**	278 ± 11	176 ± 7
	Topaz	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.05	0.01	0.01	0.07	0.02	0.01	0.08	0.02	0.59	0.80	1,659	33
	crYOLO	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.60	0.66	540	60
	RELION Amyloid	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.12	0.12	0.20	0.20	0.20	0.23	0.23	0.22	0.59	0.63	342	94
	Ground truth	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	272	203
practical	CryoTracer (ours)	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.12 ± 0.03**	**0.12 ± 0.03**	**0.12 ± 0.03**	**0.40 ± 0.05**	**0.43 ± 0.03**	**0.40 ± 0.04**	**0.54 ± 0.04**	**0.58 ± 0.01**	**0.54 ± 0.03**	**0.61 ± 0.05**	**0.66 ± 0.01**	**0.62 ± 0.03**	**0.69 ± 0.01**	0.80 ± 0.01	1,374 ± 97	181 ± 9
	Topaz	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.05	0.02	0.02	0.07	0.03	0.02	0.11	0.04	0.63	**0.83**	7,891	34
	crYOLO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.58	0.62	2,498	61
	RELION Amyloid	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.08	0.07	0.16	0.18	0.16	0.19	0.21	0.20	0.61	0.66	1,914	88
	Ground truth	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1,364	203

Table 6: CryoTracer ablations on EMPIAR-10022. Precision (P), recall (R), and F1 are averaged over micrographs at each distance tolerance (in Å). Box IoU and coverage measure spatial overlap. Synthetic pretraining uses up to 20,000 steps. Real-data training uses up to 5,000 steps unless marked 25k. Synthetic-only rows evaluate the same frozen checkpoint in both regimes. Bold labels identify the main model configuration. Bold scores mark the highest unrounded value per metric and regime. Higher scores are better. Training and architecture details appear in Appendix B.1.

Regime	Experiment	training	distance	architecture	P@25	R@25	F1@25	P@50	R@50	F1@50	P@100	R@100	F1@100	P@150	R@150	F1@150	P@200	R@200	F1@200	Box IoU	Coverage
standard	**Main**	**Synthetic + real**	**Fréchet**	**5 points, 4 levels**	0.23	0.23	0.23	0.54	0.54	0.54	0.74	0.73	0.73	0.79	0.79	0.79	0.83	0.82	0.82	0.85	0.90
	Training strategy	Synthetic only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05
		Real only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.03	0.02	0.02	0.16	0.12	0.13	0.33	0.23	0.27	0.45	0.31	0.37	0.49	0.56
		Real only (25k)	Fréchet	5 points, 4 levels	0.25	0.24	0.24	0.51	0.50	0.50	0.70	0.68	0.69	0.77	0.74	0.75	0.80	0.78	0.79	0.83	0.88
	Distance metric	Real only	Chamfer	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.60
		Synthetic + real	Chamfer	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.05	0.05	0.05	0.08	0.08	0.08	0.71	0.81
	Number of points	Synthetic + real	Fréchet	3 points, 4 levels	**0.28**	**0.28**	**0.28**	**0.57**	**0.57**	**0.57**	**0.76**	**0.76**	**0.76**	**0.81**	**0.81**	**0.81**	**0.83**	**0.83**	**0.83**	**0.85**	**0.90**
				10 points, 4 levels	0.18	0.18	0.18	0.50	0.50	0.49	0.74	0.74	0.74	0.79	0.79	0.79	0.82	0.82	0.82	0.83	0.88
	Feature pyramid	Synthetic + real	Fréchet	5 points, 1 level	0.19	0.19	0.19	0.47	0.49	0.48	0.66	0.68	0.67	0.72	0.74	0.73	0.75	0.77	0.75	0.81	0.88
	Original attention	Synthetic + real	Fréchet	5 points, 4 levels	0.13	0.14	0.14	0.42	0.42	0.42	0.68	0.69	0.68	0.77	0.78	0.77	0.82	0.82	0.82	0.82	0.89
practical	**Main**	**Synthetic + real**	**Frechet**	**5 points, 4 levels**	**0.12**	**0.12**	**0.12**	**0.42**	**0.40**	**0.41**	**0.66**	**0.63**	**0.64**	**0.74**	**0.71**	**0.72**	**0.77**	**0.74**	**0.75**	**0.80**	**0.86**
	Training strategy	Synthetic only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05
		Real only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.04	0.05	0.13	0.09	0.11	0.22	0.16	0.18	0.45	0.56
		Real only (25k)	Fréchet	5 points, 4 levels	0.06	0.06	0.06	0.27	0.23	0.25	0.53	0.45	0.48	0.62	0.54	0.57	0.69	0.59	0.63	0.68	0.74

Table 7: CryoTracer ablations on α-synuclein. Precision (P), recall (R), and F1 are averaged over micrographs at each distance tolerance (in Å). Box IoU and coverage measure spatial overlap. Synthetic pretraining uses up to 20,000 steps. Real-data training uses up to 5,000 steps unless marked 25k. Synthetic-only rows evaluate the same frozen checkpoint in both regimes. Bold labels identify the main model configuration. Bold scores mark the highest unrounded value per metric and regime. Higher scores are better. Training and architecture details appear in Appendix B.1.

Regime	Experiment	training	distance	architecture	P@25	R@25	F1@25	P@50	R@50	F1@50	P@100	R@100	F1@100	P@150	R@150	F1@150	P@200	R@200	F1@200	Box IoU	Coverage
standard	**Main**	**Synthetic + real**	**Fréchet**	**5 points, 4 levels**	0.02	0.03	0.02	**0.22**	0.24	**0.23**	**0.54**	**0.56**	**0.54**	**0.64**	0.67	0.64	**0.75**	**0.77**	**0.75**	**0.75**	0.84
	Training strategy	Synthetic only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02
		Real only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.04	0.04	0.04	0.15	0.14	0.15	0.29	0.28	0.27	0.48	0.46	0.45	0.51	0.63
		Real only (25k)	Fréchet	5 points, 4 levels	0.02	0.02	0.02	0.15	0.16	0.15	0.48	0.49	0.48	0.62	0.62	0.61	0.70	0.69	0.68	0.71	0.81
	Distance metric	Real only	Chamfer	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.48
		Synthetic + real	Chamfer	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.69	**0.87**
	Number of points	Synthetic + real	Fréchet	3 points, 4 levels	**0.04**	**0.05**	**0.04**	0.21	**0.25**	0.22	0.47	0.56	0.50	0.59	0.68	0.61	0.66	0.74	0.68	0.73	0.83
				10 points, 4 levels	0.01	0.01	0.01	0.20	0.22	0.21	0.49	0.55	0.51	0.63	**0.68**	**0.65**	0.70	0.76	0.72	0.74	0.84
	Feature pyramid	Synthetic + real	Fréchet	5 points, 1 level	0.01	0.01	0.01	0.10	0.12	0.10	0.43	0.47	0.44	0.58	0.64	0.59	0.66	0.73	0.68	0.73	0.83
	Original attention	Synthetic + real	Fréchet	5 points, 4 levels	0.01	0.01	0.01	0.09	0.10	0.10	0.34	0.36	0.34	0.56	0.59	0.56	0.66	0.69	0.67	0.69	0.79
practical	**Main**	**Synthetic + real**	**Fréchet**	**5 points, 4 levels**	**0.01**	**0.01**	**0.01**	**0.13**	**0.12**	**0.12**	**0.45**	**0.43**	**0.43**	**0.61**	**0.58**	**0.58**	**0.70**	**0.66**	**0.66**	**0.70**	**0.79**
	Training strategy	Synthetic only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
		Real only	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.04	0.03	0.03	0.07	0.06	0.06	0.28	0.41
		Real only (25k)	Fréchet	5 points, 4 levels	0.00	0.00	0.00	0.01	0.01	0.01	0.09	0.10	0.09	0.20	0.19	0.18	0.27	0.26	0.25	0.44	0.56

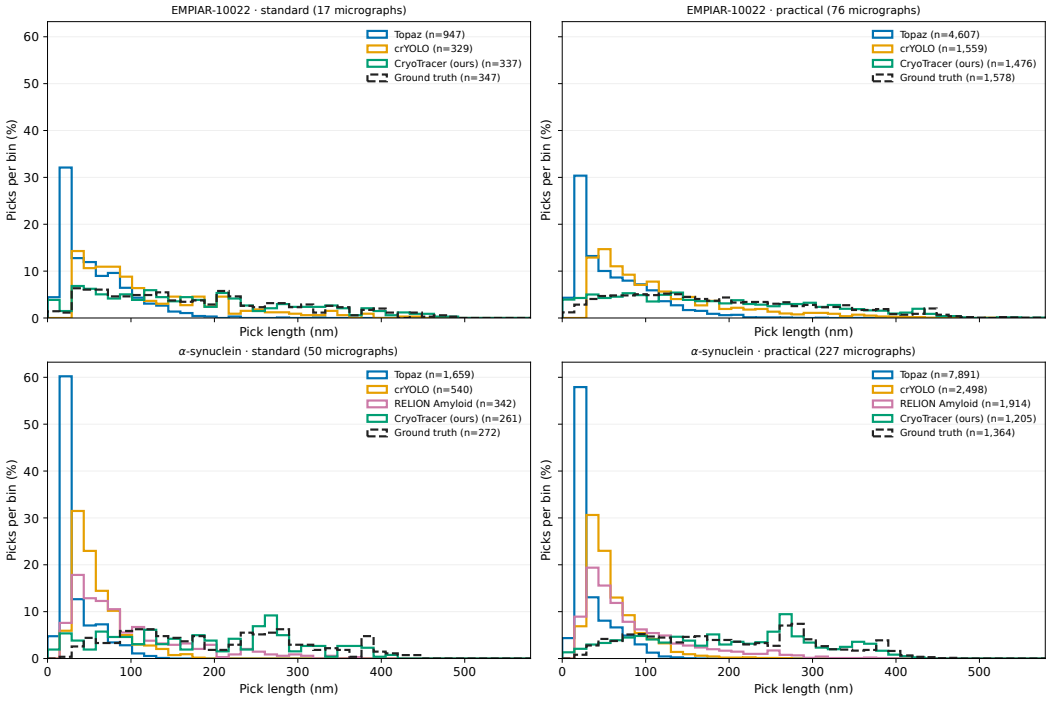

Figure 7: Picked-filament length distributions for EMPIAR-10022 (top) and α-synuclein (bottom), under the standard (left) and practical (right) evaluation regimes. Length is measured along each polyline in nanometres. Histograms share 40 equal-width bins and show the percentage of each method's picks in each bin, summing to 100% per method within a panel. The dashed curve denotes ground truth. Legend counts give the total number of picks.

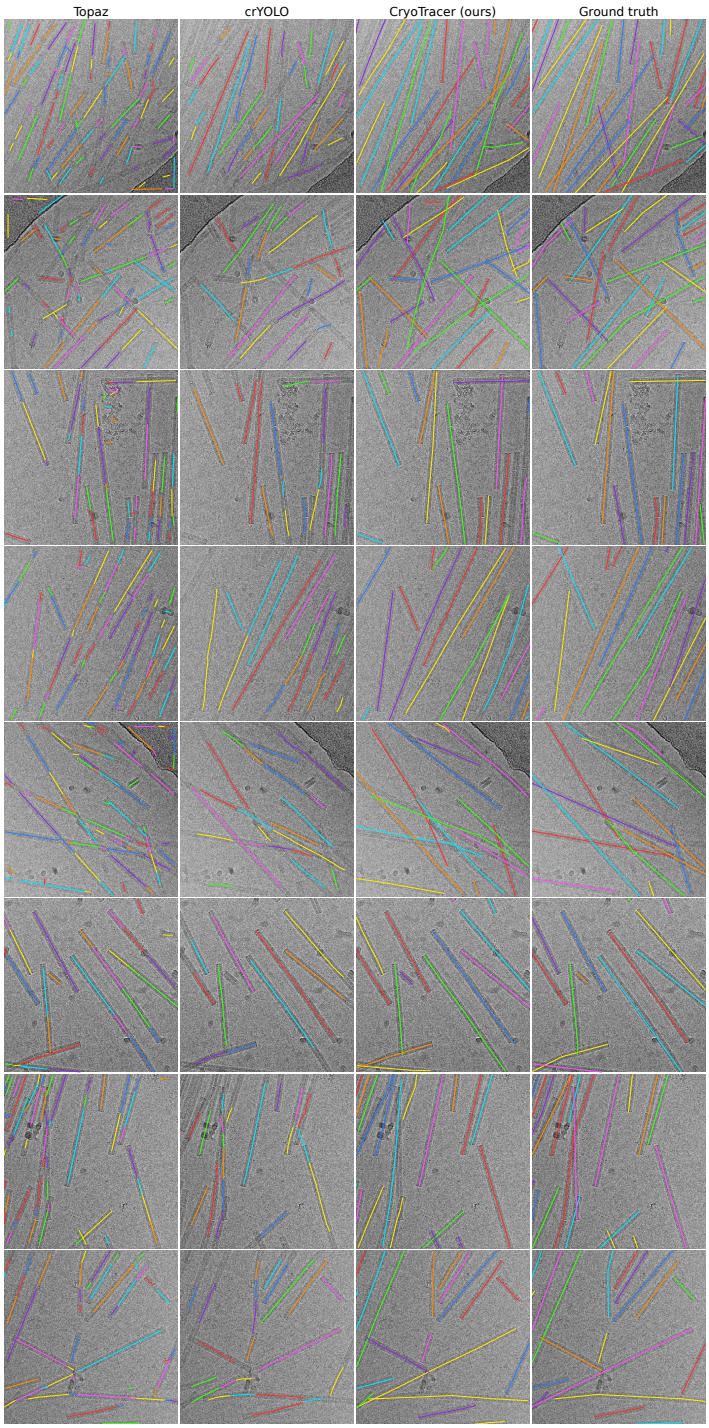

Figure 8: EMPIAR-10022, standard regime. Eight selected micrographs from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.

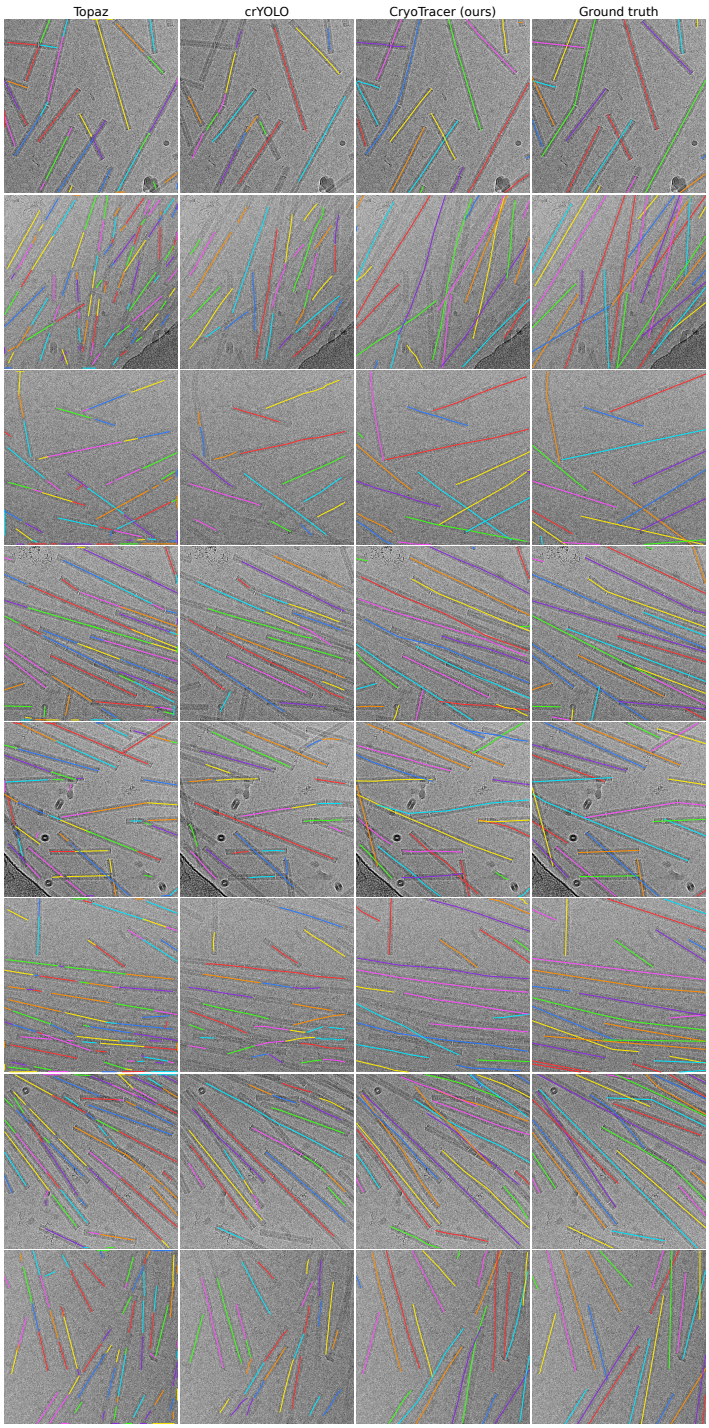

Figure 9: EMPIAR-10022, practical regime. Eight selected micrographs from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.

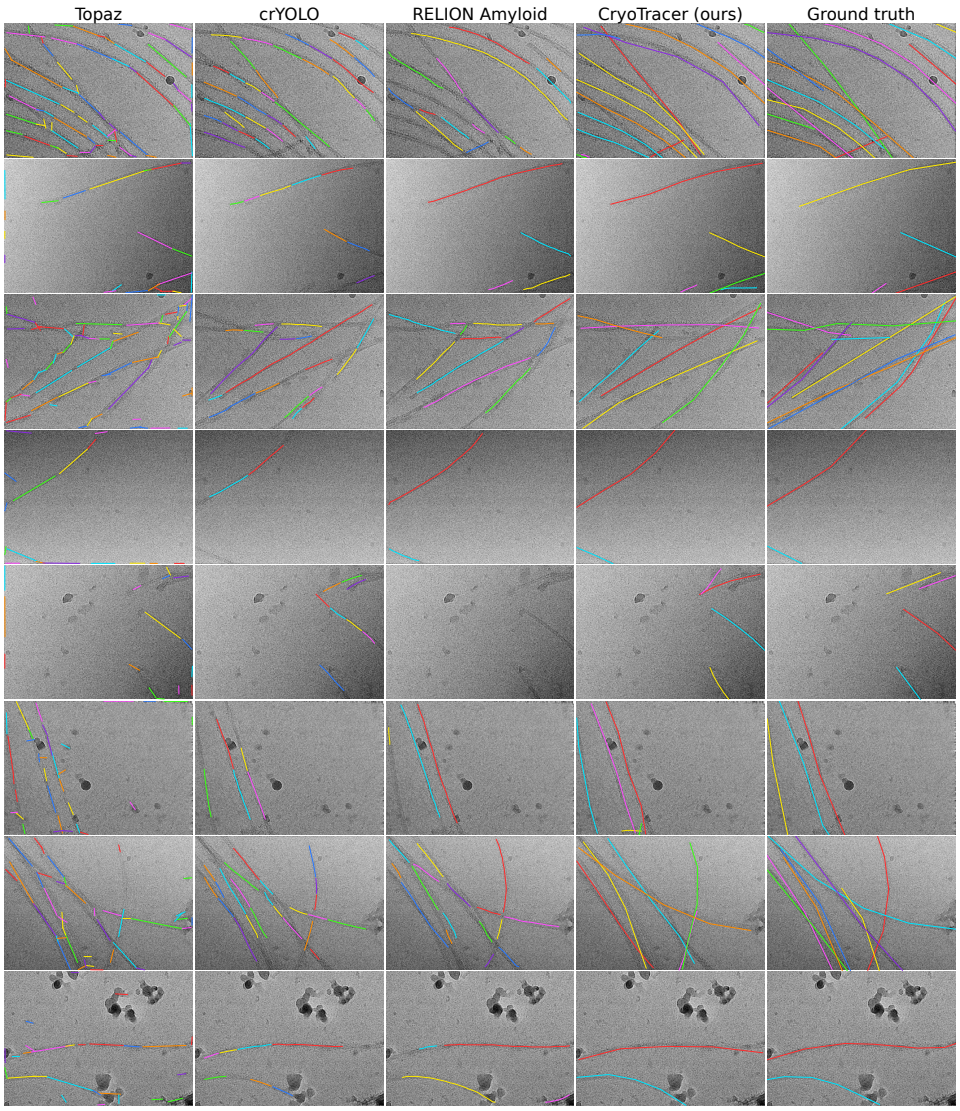

Figure 10: α-synuclein, standard regime. Eight selected micrographs from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.

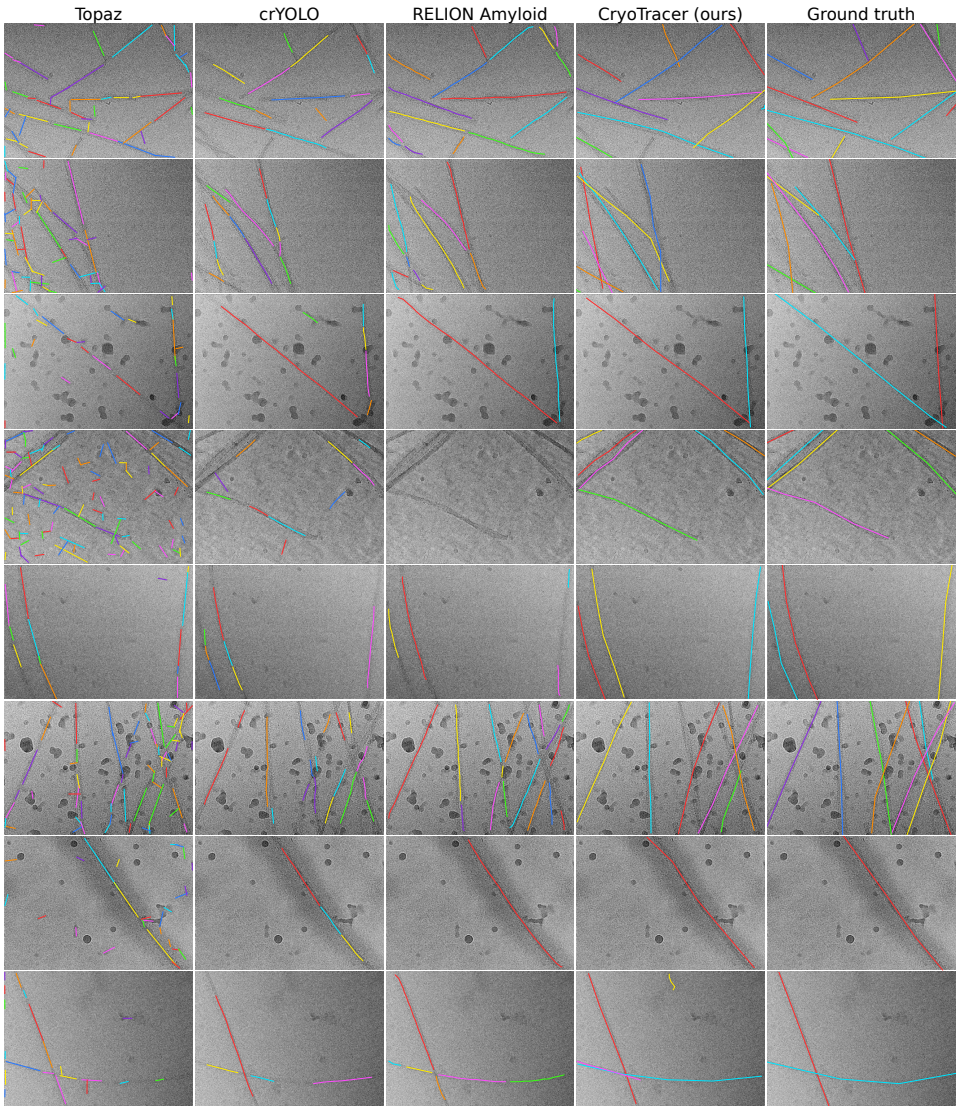

Figure 11: α-synuclein, practical regime. Eight selected micrographs from the corresponding evaluation set, with method predictions and ground truth labeled in the panels. Images are low-pass filtered and contrast adjusted for display. Colors distinguish curves within a panel and do not indicate matched identities across methods.