

TileBac - an annotated ultralow-dose cryo-EM montage of bacterial membranes and flagella

Lynnicia N Massenburg

Oak Ridge National Laboratory

Sita S. Madugula

Oak Ridge National Laboratory

Spenser R. Brown

Oak Ridge National Laboratory

Amber N. Bible

Oak Ridge National Laboratory

Chanda R. Harris

Oak Ridge National Laboratory

Scott T. Retterer

Oak Ridge National Laboratory

Jennifer L. Morrell-Falvey

Oak Ridge National Laboratory

Rama K. Vasudevan

Oak Ridge National Laboratory

Alexis N. Williams

`williamsan@ornl.gov`

Oak Ridge National Laboratory

data-descriptor

Keywords:

Posted Date: October 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11401837/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

TileBac - an annotated ultralow-dose cryo-EM montage of bacterial membranes and flagella

Lynnica N. Massenburg^{1,*}, Sita S. Madugula^{1,*}, Spenser R. Brown², Amber N. Bible², Chanda R. Harris¹, Scott T. Retterer¹, Jennifer L. Morrell-Falvey², Rama K. Vasudevan¹, and Alexis N. Williams^{1,†}

¹Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

²Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

*These authors contributed equally.

†Corresponding author: williamsan@ornl.gov

Abstract

Large-area cryogenic electron microscopy (cryo-EM) requires a tradeoff among field of view, acquisition time, and electron exposure. The goal is to mitigate this tradeoff by quickly collecting large fields of view with enough dose to detect features of interest using segmentation models but low enough dose to limit sample damage. The dataset presented here, TileBac, comprises two annotated cryo-EM datasets of *Pantoea* sp. YR343 collected with imaging conditions to achieve this goal. The first contains 2048×2048 pixel tiles derived from wide-field montages acquired at a total dose of $4 \text{ e}^-/\text{\AA}^2$ with manual pixel-level annotations of inner- and outer-membrane-enclosed regions. The second contains images acquired at a total dose of $40 \text{ e}^-/\text{\AA}^2$, with bacteria and flagella annotated as separate classes. These corresponding 40-frame movies retain individual $1 \text{ e}^-/\text{\AA}^2$ frames, allowing users to construct dose-dependent image series from the same exposure. Images and annotations are distributed in standard MRC, PNG, YOLO, and COCO formats. The records are intended to support development and evaluation of segmentation, registration, stitched inference, and dose-dependent image-analysis methods for low-signal cryo-EM.

Background and Summary

Contacts between cell-surface structures and material interfaces shape the early stages of bacterial colonization. Colonization can produce persistent contamination in clinical settings and biofouling in industrial systems[1, 2]. Gram-negative bacteria have inner and outer membranes, and changes in envelope morphology can indicate environmental stress or an antimicrobial mechanism of action [3, 4]. Bacterial flagella are micrometers long but nanometers thick and enable motility in many species. They can reflect cellular state, contribute to attachment on hydrophobic surfaces, and drive the transition from motile to sessile growth [5–7]. Characterizing these structures across a population in their near-native state requires cryo-EM imaging of many cells.

Wide-field cryo-EM montage acquisition addresses this sampling requirement by collecting many high-magnification tiles and registering them into one image. In these experiments for beam sensitive material, we try to maintain a practical collection rate while keeping a per-tile dose below $10\text{ e}^-/\text{\AA}^2$. This dose regime is well below the $40\text{--}50\text{ e}^-/\text{\AA}^2$ commonly used for low-dose imaging of proteins and the hundreds of electrons per $\text{\AA}^2$ often used in cryogenic electron tomography. However, at these lower doses cell membranes and flagella are difficult to detect and characterize.

Automated segmentation can extend membrane-morphology measurements to the population scale when these features remain detectable. Segment Anything (SAM) derivatives have been adapted for microscopy in μSAM and CellSAM[8, 9] and incorporated into electron-microscopy workflows in ACORN, developed in our laboratory [10]. U-Net architectures are well established for eukaryotic cell features[11, 12]. For bacteria specifically, we used a fine-tuned YOLOv11 segmentation model to measure envelope thickness and bacteria-flagella interactions in low-dose cryo-EM[13]. Tiling and stitching workflows such as SAHI are conventional for detection in large images[14] but registration of serial or tiled electron microscopy data remains a problem[15, 16]. These methods require representative training and evaluation data, as demonstrated for macromolecular cryo-EM[17]. Few bacterial cryo-EM datasets exist in EMPIAR[18], and, to our knowledge, no public dataset combines ultralow-dose imaging with montage tiling and pixel-level membrane annotation.

TileBac addresses this lack of training data. The first dataset is a set of 2048×2048 pixel montage tiles of *Pantoea* sp. YR343, a *Populus deltoides* rhizosphere isolate[19], acquired at a total dose of $4\text{ e}^-/\text{\AA}^2$ and manually annotated with the inner- and outer-membrane-enclosed regions as separate classes. Tiles were

selected from 4096×4096 pixel montage images collected as 7×7 arrays covering a single 200-mesh grid square. Each source image was divided into four non-overlapping 2048×2048 pixel quadrants before resizing to the 640- or 1024-pixel model inputs. This limits the downsampling applied to cell boundaries. The release includes stitched montages, source tiles, raw images, and dynamically normalized images. The second dataset supports flagella analysis. Low-dose images of *Pantoea* sp. YR343 acquired at $40 \text{ e}^-/\text{\AA}^2$ were analyzed for the presence of flagella and then annotated for flagella and bacteria as two separate classes. The same fields were recorded as 40-frame movies at $1 \text{ e}^-/\text{\AA}^2$ per frame. Figure 1 summarizes the relationship among a source montage image, its quadrants, and an annotated tile.

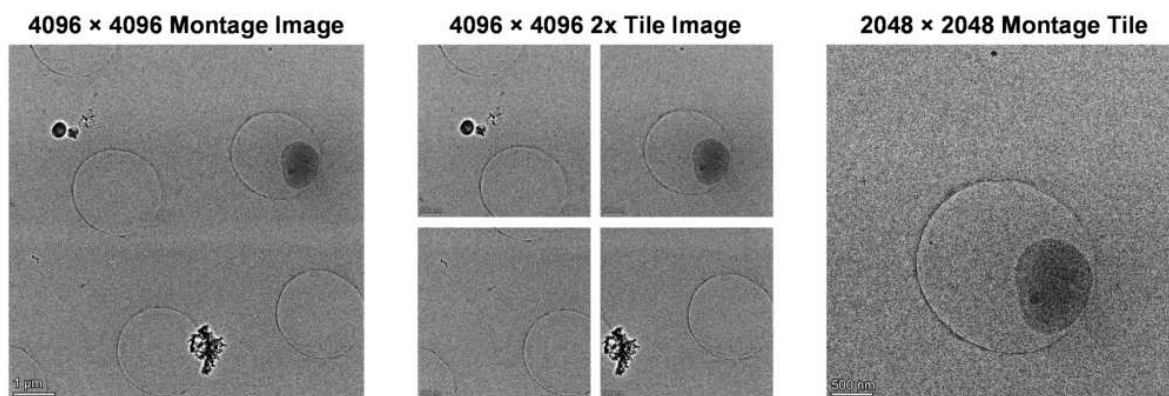


Figure 1: Ultralow-dose cryo-EM montage images and derived tiles. Left, a dynamically normalized 4096×4096 pixel montage image acquired at $4 \text{ e}^-/\text{\AA}^2$; scale bar, $1 \mu\text{m}$. Center, subdivision into 2048×2048 pixel tiles, which reduces compression of thin membrane features during model input resizing; scale bar, 500 nm. Right, a representative annotated tile from the TileBac benchmark dataset.

The dataset supports three primary uses: training and evaluating segmentation architectures on low-signal cryo-EM images; developing registration and stitched-inference pipelines that avoid seams through cells; and measuring dose-dependent image-analysis performance. Because the flagella movies are released as unaveraged frames, users can construct images spanning $1 \text{ e}^-/\text{\AA}^2$ to $40 \text{ e}^-/\text{\AA}^2$. The specimen, instrument settings, and field of view remain fixed within each movie, allowing users to evaluate segmentation performance as a function of accumulated dose. All images, annotations, and the fine-tuned models used for technical validation are openly available through the records cited in the reference section of this work.

Methods

Bacterial growth

Pantoea sp. YR343 wild-type cells were grown on R2A agar at 29 °C. Single colonies were used to inoculate liquid cultures in R2A broth, minimal MOPS medium supplemented with 0.4% glucose, or minimal MOPS medium supplemented with 0.4% succinate. Liquid cultures were shaken at 250 rpm and cells were harvested after 24 h at stationary phase[13]. Cultures for the ultralow-dose montage dataset were grown in R2A or MOPS with glucose. Cultures for the flagella datasets were grown in all three conditions.

Cryo-EM sample preparation

Quantifoil or C-Flat 2/2 gold grids were glow discharged immediately before use at 5 mA for 45 s. Aliquots of 3 μ L of culture were applied to each grid and blotted in a Vitrobot Mark IV (Thermo Fisher Scientific) at a blot force of -5 for 5–10 s at 100% relative humidity and 4.5 °C. Grids were vitrified by plunge freezing into a liquid ethane/propane mixture cooled with liquid nitrogen and stored under liquid nitrogen until imaging.

Ultralow-dose montage acquisition

Montage images were collected on a Thermo Fisher Scientific Krios G4 scanning transmission electron microscope operating at 300 kV in nanoprobe TEM mode with a 70 μ m C2 aperture. A Falcon 3EC direct electron detector recorded images in counting mode. Tomography 5 (Thermo Fisher Scientific) automated acquisition with the Search Map workflow. Search-map montages were collected as 7×7 image arrays at 3.8 k $\times$ magnification, 2.128 nm per pixel and 4096×4096 pixels per tile, covering the area of a single grid square on a 200-mesh grid[13]. Each tile received a total electron dose of $4 \text{ e}^-/\text{\AA}^2$. These are referred to throughout as ultralow-dose montage (ULDM) images. Tomography 5 generated 12005×12005 pixel stitched montages automatically. The 2048×2048 quadrants are crops rather than resampled images and retain the source sampling of 2.128 nm per pixel. Resampling to 640 and 1024 pixels gives 6.8096 and 4.256 nm per pixel, respectively.

Low-dose bacteria and flagella acquisition

Images for the flagella dataset were collected manually rather than through montage automation on the same Krios G4 with the same Falcon 3EC detector in counting mode. Images were acquired in nanoprobe TEM mode at 8.7 k $\times$ magnification, 4096×4096 pixels, 0.8986 nm per pixel, -5μ m objective-lens underfocus,

and a total electron dose of $40 \text{ e}^-/\text{\AA}^2$ [13]. Fields of view were selected by screening for the presence of flagella.

Frame-partitioned movie acquisition

The Falcon 3EC recorded flagella-containing fields as 40-frame movies in counting mode, with $1 \text{ e}^-/\text{\AA}^2$ delivered per frame. For these multi-frame MRC files individual frames are recorded so that any subset can be summed. Summing four frames reproduces the $4 \text{ e}^-/\text{\AA}^2$ total dose used for the montage tiles, and summing all forty reproduces the conventional low-dose condition.

Image preprocessing

The microscope recorded cryo-EM micrographs as MRC files were converted to PNGs. Per-image adaptive normalization (e.g., dynamic normalization) set the display range from each image’s intensity distribution rather than from a fixed global range. The release includes both unnormalized and normalized PNG files so users can apply alternative preprocessing steps. Normalization has a larger visual effect on the ULDM tiles than on the $40 \text{ e}^-/\text{\AA}^2$ flagella images (Figure 2).

Tiling

Each 4096×4096 montage image was divided into four non-overlapping 2048×2048 tiles. Tiles without bacterial cells were removed before annotation to prevent background-only images from dominating training. The published dataset includes both the source montage images and the tiled data.

Annotation

Annotators manually delineated polygons in Roboflow[20]. In the ultralow-dose montage tile dataset, inner membranes are class 0 and outer membranes are class 1. In the flagella dataset, flagella are class 0 and bacteria are class 1 and only the outer membrane was annotated for the bacterial class. The exports store all annotations as polygon masks (Figure 3).

Flagella in these images frequently overlap bacterial cells but are not visibly continuous with the cells. As such, we did not attempt to assign flagella to individual cells and the flagella annotations only signify overlap and not attachment.

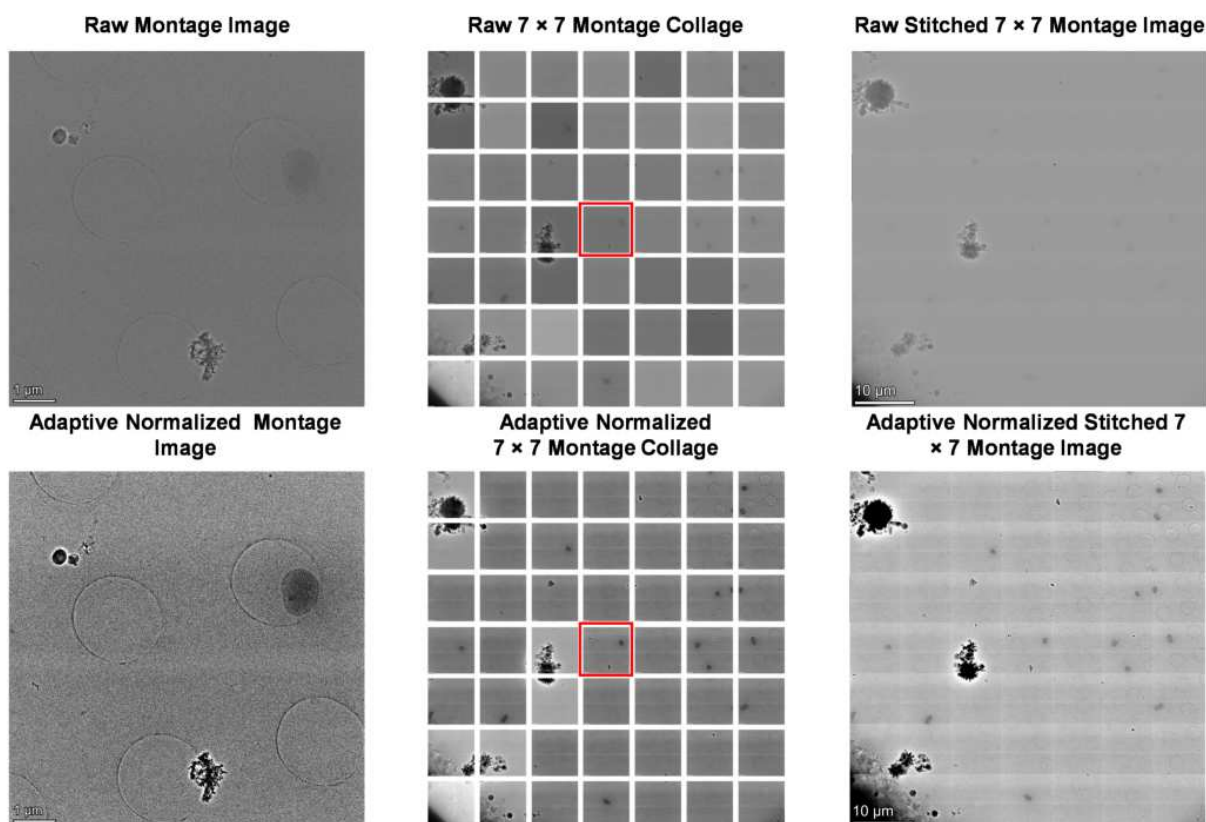


Figure 2: **Dynamic normalization and montage stitching.** Left, a 4096×4096 pixel montage tile before (top) and after (bottom) per-image adaptive normalization. Center, the montage array before registration; the tile outlined in red corresponds to the image in the left panel. Right, the 12005×12005 pixel stitched montage generated by the Tomography 5 Search Map workflow. Bacterial cells occupy fewer pixels in the stitched image than in the source tiles.

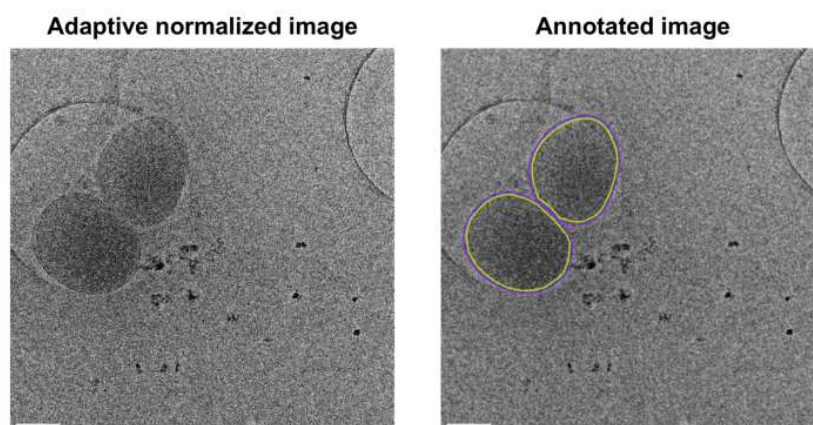


Figure 3: **Annotation of bacterial inner and outer membranes.** Left, a 2048×2048 pixel ULDM tile. Right, the same tile with inner-membrane (lime) and outer-membrane (purple) polygon annotations. Scale bar, 500 nm.

Dataset partitioning and augmentation

The ultralow-dose montage dataset was partitioned into training, validation, and test sets with 272, 18, and 52 tiles, respectively. Test tiles were held out at the 7×7 level. The training and validation shared eight arrays and three source images and the validation subset included no MOPS-glucose datasets. Partition and instance counts are reported in Table 1.

Table 1 | Composition of the ultralow-dose montage tile dataset.

Quantity	Train	Valid	Test	Total
Search Map arrays (7×7)	16	8	3	19
Source 4096×4096 images	210	18	36	261
2048×2048 tiles	272	18	52	342
Inner membrane	307	19	57	383
Outer membrane	306	19	57	382
Images after augmentation	816	18	52	886

Augmentation of the training partition included vertical and horizontal flips, crops with zoom between 0% and 20%, rotations between -15° and $+15^\circ$, horizontal and vertical shear of $\pm 10^\circ$, and saturation adjustment between -25% and $+25\%$. Images are resized to either 640×640 or 1024×1024 pixels, and both versions are released. Using this augmentation triples the training partition to 816 images. The validation and test partitions are not augmented in this dataset. Table 1 summarizes the partition sizes and the augmentation settings.

Segmentation models used for technical validation

Five segmentation architectures were fine-tuned from pretrained weights to test whether the annotations support model training. These five architectures are YOLOv11, YOLO26, U-Net, Detectron2, and SAM3[11, 21–25]. Each was trained at both 640 and 1024 pixel input sizes and then ten fine-tuned models. Training was performed in Google Colab on an A100 GPU. Table 2 lists the optimizer, learning rate, batch size, training length, input size, and model selection criterion for each architecture. Annotations were exported in YOLO text format for the YOLO models and in COCO format for U-Net, Detectron2, and SAM3. All five architectures used identical partitions for model testing.

Table 2 | Training configuration for the baseline segmentation models. YOLOv11 and YOLO26 refer to the YOLOv11n-seg and YOLO26n-seg checkpoints, respectively. Each model was trained at 640 and 1024 pixel input size.

Model	Optimizer	Learning rate	Batch	Length	Checkpoint
YOLOv11	AdamW	1×10^{-3}	16	100 epochs	Best validation fitness
YOLO26	AdamW	1×10^{-3}	16	100 epochs	Best validation fitness
U-Net	Adam	1×10^{-4}	4	180 epochs	Best validation Dice
Detectron2	SGD	2.5×10^{-4} to 1×10^{-4}	2	8,000 iterations	Highest AP
SAM3	AdamW	5×10^{-7} global; 2.5×10^{-5} backbone	4	20 epochs	Final epoch, no selection

All ten models were evaluated using a common procedure rather than framework-specific validators. Predictions were converted to COCO format at full resolution and evaluated against a single ground-truth file using one `COCOeval` segmentation call (`pycocotools` 2.0.11). For these models, AP50 and AP50–95 were taken from summary statistics. This standardization was necessary for cross model comparison because Ultralytics evaluates YOLO masks at the 160×160 prototype resolution unless JSON export is enabled, and the Detectron2 script reports pixel-level average precision rather than instance AP. For U-Net, instances were reconstructed as connected components of the sigmoid mask thresholded at 0.5 and scored by mean component probability. Pixel-level F1 was calculated for every model at the fixed probability threshold of 0.5. There was no threshold tuning on either evaluation partition. The evaluation script is released with the dataset and Figures 4 and 5 show the qualitative model outputs and unified evaluation curves.

Data Records

All datasets described here are openly available through the records listed in the references [26–30]. Separate records contain the raw and processed images, annotations, and fine-tuned models so users can retrieve only the components required for a given analysis.

Ultralow-dose montage tile dataset

The benchmark [26] contains annotated 2048×2048 pixel montage tiles as PNGs in unnormalized and dynamically normalized forms and two annotation formats. YOLO labels have one text file per image with a class index followed by normalized polygon coordinates while COCO annotations are stored as one JSON file per partition using polygon coordinates in absolute pixels. For both formats, class 0 is the inner-membrane-enclosed region and class 1 is the outer-membrane-enclosed region. Each Roboflow

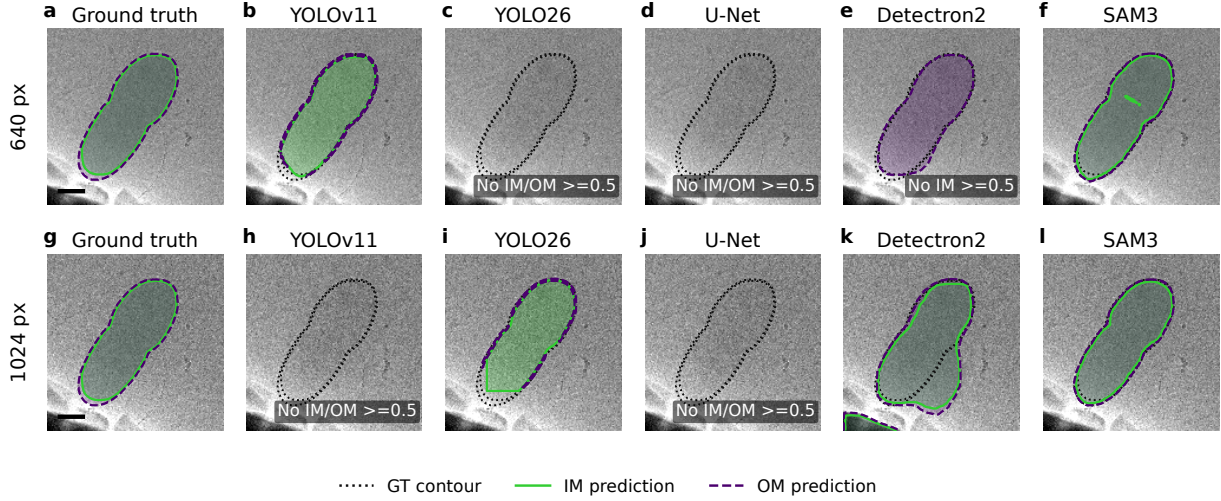


Figure 4: **Predicted inner-membrane (IM) and outer-membrane (OM) masks for five architectures on a fully visible ULDM test cell.** The tile was selected as the largest fully visible test OM instance using ground truth only, without conditioning on model success. (a–f) 640-pixel export; (g–l) 1024-pixel export. Ground-truth contours are dotted; IM predictions are solid and OM predictions are dashed. YOLOv11 and YOLO26 render OM as a thin boundary, whereas U-Net, Detectron2, and SAM3 reproduce the filled-region annotation convention. Scale bar, 500 nm.

export contains `train`, `valid`, and `test` partitions. YOLO partitions include `images` and `labels` directories, whereas COCO partitions contain the images and an `_annotations.coco.json` file. The YOLO export includes a root-level `data.yaml` file showing class names and partitions.

Source montage images and stitched montages

These datasets [27, 28] contain the 4096×4096 pixel source images from which the annotated tiles were extracted, in MRC and PNG formats, with 12005×12005 pixel stitched montages generated with Thermo Fisher Scientific Tomography 5. The images are unannotated and support registration, full-montage inference, and construction of alternative tilings. An accompanying CSV manifest maps each 2048×2048 pixel tile filename in every export to its source image, Search Map array, position within the 7×7 array, quadrant, growth condition, and partition.

Low-dose bacteria and flagella dataset

This released dataset [29] contains the manually collected $40 \text{ e}^-/\text{\AA}^2$ images screened for flagella in raw and dynamically normalized PNGs with annotations in YOLO and COCO exports. Class index 0 is flagella and

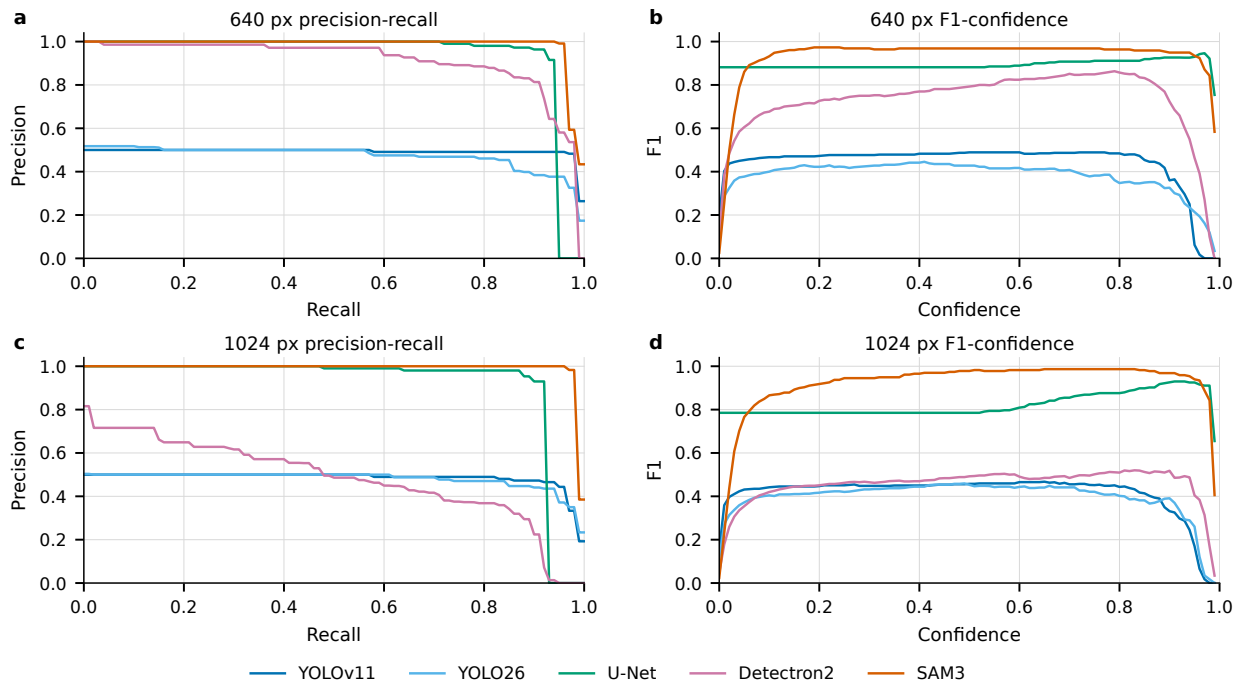


Figure 5: **Precision–recall and F1–confidence curves for five architectures under a unified evaluation of the 52-tile test partition (114 instances).** (a,b) 640-pixel export; (c,d) 1024-pixel export. Precision–recall curves use IoU 0.50, and F1–confidence curves use instance matching at IoU 0.50. Curves are averaged over the inner- and outer-membrane classes. The effect of boundary-v-region predictions for these metrics can be seen in Figure 4.

class index 1 is bacteria. Two augmented training variants, generated at $2\times$ and $3\times$ the original training set size, are provided to increase the representation of flagella from the source images. These variants are released in both YOLO and COCO formats at $40\text{ e}^{-}/\text{\AA}^2$, with outputs resized to 1024×1024 pixels. Augmentation was applied only to the training partition and used horizontal and vertical flips, crop with 0–20% zoom, rotation from -15° to $+15^\circ$, horizontal and vertical shear of $\pm 10^\circ$, and saturation adjustment from -25% to $+25\%$. Figure 6 shows a representative field and Table 3 summarizes the native 4096×4096 export.

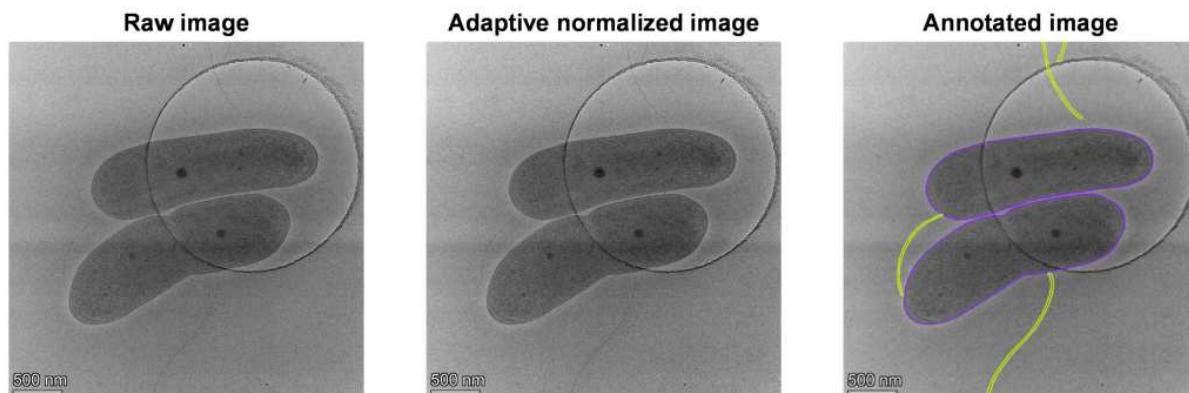


Figure 6: **Low-dose cryo-EM images of bacteria and flagella.** Left, raw image. Center, the image after adaptive normalization. Right, annotations of bacteria (purple) and flagella (lime).

Table 3 | Low-dose bacteria and flagella dataset counts computed from the public Hugging Face COCO annotation JSONs for the native 4096×4096 export.

Split	Images	Flagella instances	Bacteria instances
Train	219	657	347
Validation	28	70	24
Test	31	80	58
Total	278	807	429

Frame-partitioned flagella movies

This record [29] contains the 40-frame movies acquired at $1 \text{ e}^-/\text{\AA}^2$ per frame and distributed as multi-frame MRC files. The files retain the frames in acquisition order without annotations. Users can construct a specified total dose by summing the corresponding number of frames and then applying the preprocessing described in Methods. Figure 7 shows sums of 1, 4, 10, and 40 frames from the same movie.

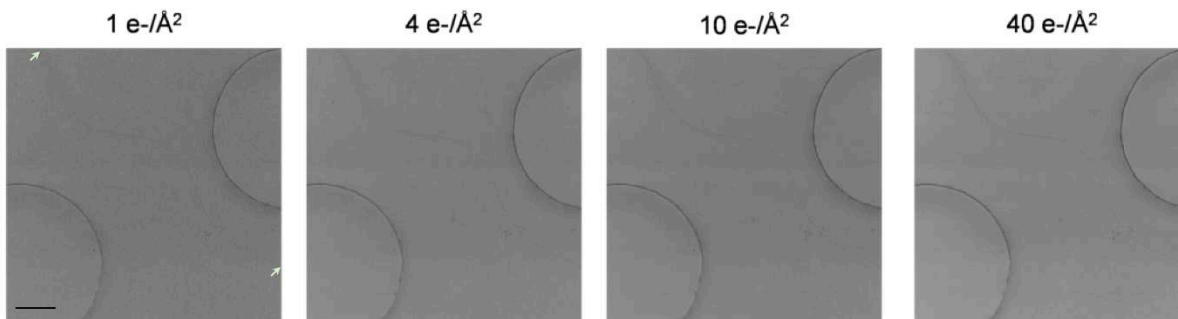


Figure 7: **Frame summation across the dose range.** Images generated from one 40-frame movie by summing 1, 4, 10, and 40 frames, corresponding to total doses of $1 \text{ e}^-/\text{\AA}^2$, $4 \text{ e}^-/\text{\AA}^2$, $10 \text{ e}^-/\text{\AA}^2$, and $40 \text{ e}^-/\text{\AA}^2$. The $4 \text{ e}^-/\text{\AA}^2$ image matches the total dose used for the montage tiles. Green arrows denote the start and end of the flagellum. Scale bar, 500 nm.

Fine-tuned models

The ten fine-tuned segmentation models described in Technical Validation are released as separate files [30]. The file names specify the input size, architecture, and dataset (e.g., 640-YOLOv11-ULDM). Each record contains model weights and the corresponding training configuration. These models provide reproducible baselines and are not recommendations for a specific application.

Technical Validation

Technical validation addresses three questions: what the annotations represent; what the overlap and boundary metrics can and cannot show; and whether the annotations support training across segmentation architectures. All results below were calculated from the released dataset using the code described in the Code Availability section.

Annotation geometry and labeling convention

For each bacterial cell the regions enclosed by the inner and outer membranes were annotated as separate filled polygons and across the 383 matched inner-outer annotation pairs in the full dataset the median IoU was 0.890. This shows that the two regions contain many of the same pixels. The annotations were convex to within rasterization error (median solidity, 1.00). Across the 228 test annotations at both input sizes the median fill ratio, defined as the annotated area divided by the bounding-box area, was 0.759. The median local width was 90 pixels at 640 input size and 142 pixels at 1024, corresponding to approximately 600 nm for both sizes with no test annotation more narrow than 5 pixels.

These labels of inner and outer membrane do not represent the actual bilayer. They are representative of the area enclosed by each membrane. This means a model that predicts a narrow membrane trace can correctly place the boundary on the respective bacterial cell membrane but still receive a low region-overlap score because the trace covers less area than the filled annotation. This distinction is evaluated in the next section using both region and boundary metrics.

One training tile (Tile_29481199_2, quadrant 1_1) contains one outer-membrane polygon enclosing two inner-membrane polygons. This accounts for the difference between the 383 inner-membrane and 382 outer-membrane annotations. Solidity values slightly above 1 arise from rasterization and are reported as 1.00.

Capabilities of overlap and boundary metrics

We first compared the model results with three geometry-based reference masks constructed from each ground-truth annotation: a filled bounding box, an ellipse inscribed in the bounding box, and the annotation’s convex hull. Their mean per-instance Dice coefficients on the test partition were 0.845, 0.859, and 0.991, respectively. Table 4 compares these values with the fine-tuned models, and Figure 8a–f shows the corresponding masks and scores.

Table 4 | Results for reference masks and fine-tuned models in the test partition.

Method	Type	Cells	Input	Dice, mean (95% CI)
Convex hull	Reference	57	640	0.991 (0.987–0.993)
Convex hull	Reference	57	1024	0.992 (0.989–0.994)
Inscribed ellipse	Reference	57	640	0.859 (0.838–0.878)
Inscribed ellipse	Reference	57	1024	0.859 (0.838–0.879)
Filled bounding box	Reference	57	640	0.845 (0.828–0.861)
Filled bounding box	Reference	57	1024	0.845 (0.827–0.861)
SAM3	Fine-tuned	57	640	0.919 (0.851–0.970)
SAM3	Fine-tuned	57	1024	0.956 (0.916–0.979)
U-Net	Fine-tuned	57	640	0.911 (0.846–0.961)
U-Net	Fine-tuned	57	1024	0.896 (0.826–0.951)
Detectron2	Fine-tuned	57	640	0.894 (0.830–0.946)

Method	Type	Cells	Input	Dice, mean (95% CI)
Detectron2	Fine-tuned	57	1024	0.807 (0.716–0.889)
YOLOv11	Fine-tuned	57	640	0.522 (0.500–0.537)
YOLOv11	Fine-tuned	57	1024	0.515 (0.478–0.545)
YOLO26	Fine-tuned	57	640	0.424 (0.371–0.473)
YOLO26	Fine-tuned	57	1024	0.508 (0.468–0.539)

The convex hull scored higher than every fine-tuned model. For comparison, SAM3, the best-performing fine-tuned model by Dice, scored 0.919 at 640 input size and 0.956 at 1024. Because the convex hull was constructed from the annotation geometry without using image information, this comparison shows that Dice alone does not establish accurate boundary prediction.

The difference between region and boundary predictions is most evident in the YOLO results. The YOLO models predict the outer membrane as a narrow trace, whereas the other three architectures predict the filled region represented by the annotation. All five architectures predict filled regions for the inner membrane. The YOLO outer-membrane traces had mean widths of 4.00 pixels at 640 input size and 6.32 pixels at 1024, corresponding to 27.2 and 26.9 nm, respectively. The similar physical widths at both input sizes indicate that the traces are not caused by image resampling. Because a narrow trace has little area in common with a filled annotation, the YOLO models had near-zero outer-membrane AP50 despite inner-membrane AP50 values above 0.98. Morphological closing and hole filling increased YOLOv11 outer-membrane AP50 from 0.000 to 0.138 at 640 input size and from 0.000 to 0.140 at 1024. For YOLO26 at 640, AP50 increased from 0.005 to 0.109. Figure 8g–j shows this effect for one cell. The same operation did not alter the outputs of the three region-predicting architectures or any inner-membrane output.

However, filling did not recover every YOLO prediction because approximately 57% of test instances were truncated at a tile boundary which left open traces in these instances. Across both YOLO models and input sizes, none of the 138 traces on truncated cells could be filled, compared with 41 of 77 traces on complete cells. When a trace could be filled, the resulting region had a mean IoU of 0.978 with the annotation.

Boundary measurements provide a different assessment of the same predictions. Table 5 reports the symmetric surface distance between the predicted and annotated outer-membrane outlines. YOLOv11 had

a mean surface distance of 21.0 nm at 640 input size and 20.4 nm at 1024, ranking second among the five architectures at 640 and first at 1024 despite an outer-membrane AP50 of 0.000 at both sizes. Results on TileBac should state whether a method predicts the enclosed region or the membrane boundary. Boundary-predicting methods should be evaluated using a boundary metric in addition to region-overlap metrics.

Finally, the dataset does not contain overlapping cells. Excluding the expected nested inner–outer annotations, none of the 189 intercell annotation pairs intersected. The dataset does contain closely adjacent cells with 70 pairs within 15 pixels and 25 of the 47 outer-membrane pairs within 5 pixels. No overlaps were found for any of the 969 possible choices of three held-out arrays, confirming that this is a property of the full dataset rather than the released partition. This means TileBac can be used to evaluate segmentation of separated cells but not overlapping ones.

Table 5 | Boundary localization results showing how close each model’s predicted membrane outline is to the annotated outline in nm.

Model	OM, 640	OM, 1024	IM, 640	IM, 1024
SAM3	14.3	61.8	21.3	70.9
U-Net	32.1	62.9	36.5	68.2
Detectron2	119.2	346.5	82.4	284.8
YOLOv11	21.0	20.4	36.4	37.6
YOLO26	259.5	45.9	137.5	65.0

Model training across segmentation architectures

Ten models spanning five architectures were fine-tuned on the training partition and evaluated on the validation and test partitions using the common evaluator described in Methods. Table 6 reports AP50 and AP50–95 on the test partition, AP50 on the validation partition, and pixel-level F1 on the test partition at a fixed probability threshold of 0.5. Values are reported separately for each model and class.

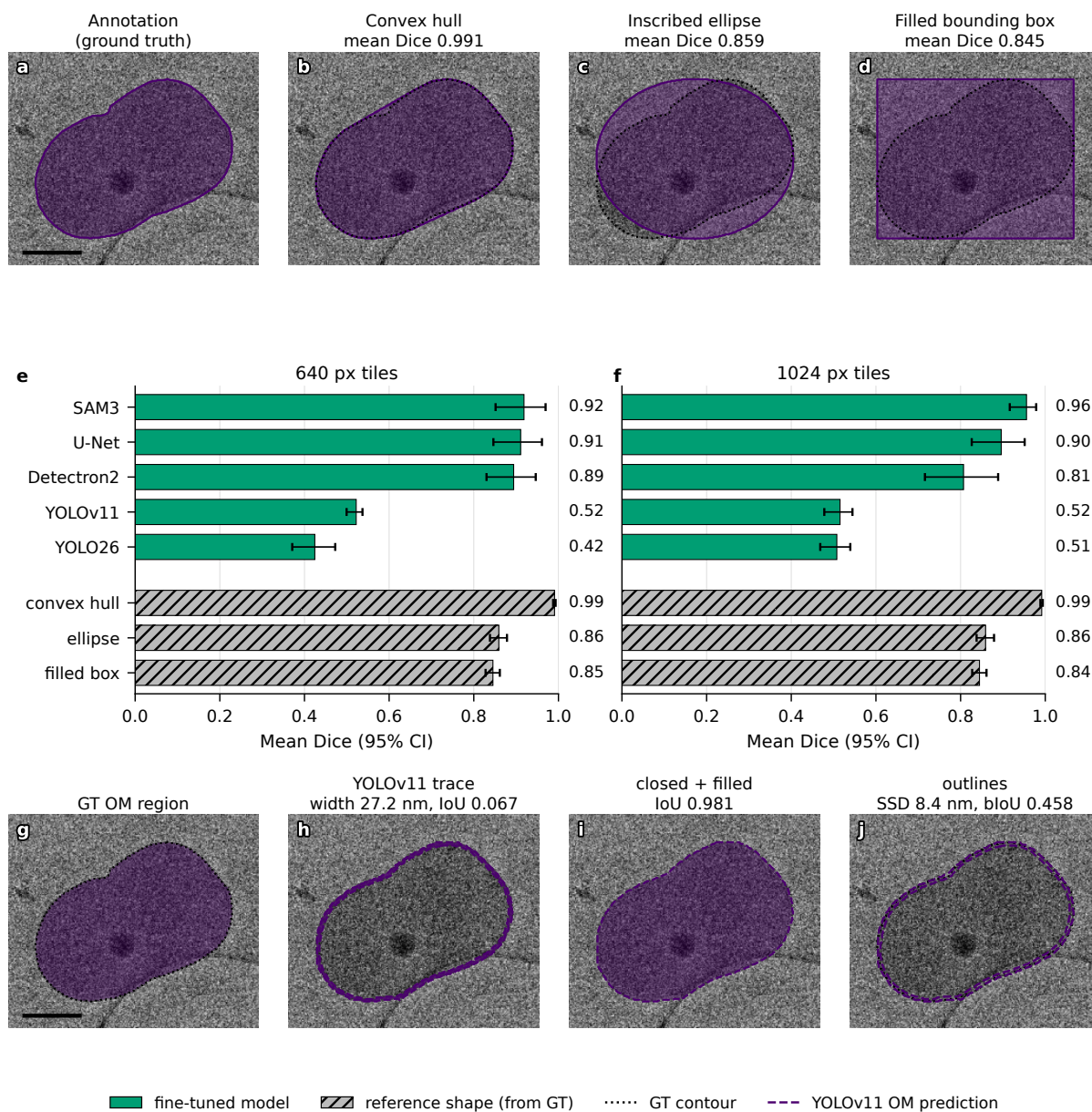


Figure 8: Annotation geometry and geometry-based reference masks. (a–d) One fully visible outer-membrane annotation and its convex hull, inscribed ellipse, and filled bounding box. The annotation contour is shown as a dotted line in the three derived panels. Values in panels a–d are dataset-level mean Dice coefficients over all 114 test annotations at 640 input size. (e,f) Mean Dice coefficients and 95% confidence intervals for the five fine-tuned models (green) and three reference masks (gray, hatched) at 640 (e) and 1024 (f) input size. Values beside the bars are mean Dice coefficients. The reference masks use annotation geometry but no image intensities. (g–j) Region and boundary measurements for one fully visible test cell selected for high overlap after filling: (g) filled ground-truth outer-membrane annotation; (h) YOLOv11 outer-membrane prediction, a 27.2 nm-wide trace with region IoU 0.067; (i) the same trace after morphological closing and hole filling, with region IoU 0.981; and (j) superimposed outlines, with symmetric surface distance 8.4 nm and boundary IoU 0.458. This example illustrates the effect of filling and is not a dataset-level performance estimate. Across the test partition, YOLOv11 had a mean outer-membrane surface distance of 21.0 nm, and filling increased outer-membrane AP50 from 0.000 to 0.138 at 640 input size. Scale bars, 500 nm.

Table 6 | Baseline model performance results that show SAM3, U-Net, and Detectron2 generally predict the filled annotated regions well, while YOLO performs well for the inner membrane but fails region-based outer-membrane AP since it predicts a narrow boundary trace instead of a filled region.

Model	Class	AP50	AP50– 95	Val AP50	Pixel F1
640-SAM3-ULDM	All	0.981	0.919	1.000	0.943
	IM	0.972	0.895	1.000	0.935
	OM	0.989	0.943	1.000	0.951
640-U-Net-ULDM	All	0.935	0.844	0.941	0.887
	IM	0.934	0.826	0.941	0.883
	OM	0.936	0.861	0.941	0.889
640-Detectron2-ULDM	All	0.906	0.769	0.969	0.740
	IM	0.900	0.717	0.939	0.750
	OM	0.912	0.821	1.000	0.732
640-YOLOv11-ULDM	All	0.491	0.414	0.500	0.620
	IM	0.983	0.828	1.000	0.927
	OM	0.000	0.000	0.000	0.106
640-YOLO26-ULDM	All	0.471	0.393	0.500	0.517
	IM	0.938	0.784	1.000	0.803
	OM	0.005	0.002	0.000	0.080
1024-SAM3-ULDM	All	0.988	0.912	1.000	0.962
	IM	0.980	0.887	1.000	0.952
	OM	0.996	0.937	1.000	0.971
1024-U-Net-ULDM	All	0.912	0.826	0.910	0.866
	IM	0.913	0.815	0.880	0.863
	OM	0.910	0.836	0.941	0.869

continued on next page

Table 6 continued

Model	Class	AP50	AP50– 95	Val AP50	Pixel F1
1024-Detectron2-ULDM	All	0.482	0.399	0.729	0.356
	IM	0.497	0.415	0.636	0.358
	OM	0.467	0.384	0.821	0.353
1024-YOLOv11-ULDM	All	0.484	0.411	0.500	0.590
	IM	0.967	0.822	1.000	0.883
	OM	0.000	0.000	0.000	0.116
1024-YOLO26-ULDM	All	0.479	0.416	0.500	0.585
	IM	0.957	0.832	1.000	0.874
	OM	0.000	0.000	0.000	0.115

For class C , AP50 is average precision at an IoU threshold of 0.5. Mean AP50 over C classes is

$$\text{mAP}_{50} = \frac{1}{C} \sum_{c=1}^C \text{AP}_c(\text{IoU} = 0.5). \quad (1)$$

Average precision was calculated by COCOeval using interpolated precision at 101 recall thresholds:

$$\text{AP} = \frac{1}{101} \sum_{r \in \{0, 0.01, \dots, 1\}} P_{\text{interp}}(r). \quad (2)$$

Precision and recall are

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad (3)$$

where TP , FP , and FN denote true positives, false positives, and false negatives. Pixel-level F1 is

$$\text{F1} = \frac{2PR}{P + R}. \quad (4)$$

Here, P and R are calculated over pixels. Pixel-level F1 is equivalent to the Dice coefficient.

The annotations supported training across all five architectures. At 640 input size, every architecture achieved an inner-membrane AP50 of at least 0.90 on the test partition. The near-zero outer-membrane AP50 values of the YOLO models reflect their narrow boundary predictions, as described above. Detectron2 achieved an AP50 of 0.482 at 1024 input size. In the COCO size breakdown, its AP50–95 was 0.079 for medium instances and 0.408 for large instances. At 640 input size, the corresponding values were 0.765 and 0.785.

The validation partition contains 19 instances per class and is sensitive to individual predictions. In these cases one instance represents approximately 5% of the class total. Several models achieved a validation AP50 of 1.000. Validation data were not used to select thresholds for any result in Table 6. Figure 5 shows the precision–recall and F1–confidence curves at both input sizes.

Usage Notes

Recommended partitions

Users comparing against the baselines in Table 6 should use the released partition unchanged. Its test partition is held out at the parent array level, so no grid square contributes to both test and the other partitions. Validation is not held out in the same way (see Methods) as it shares arrays with training, contains no MOPS glucose tiles, and has 19 instances per class. Users constructing their own partition should group by array, using the manifest supplied with the source montage record, which maps every 2048×2048 tile to its source image, array, position in the array, and quadrant. Grouping by source image alone keeps quadrants of one image together but still places neighboring images from the same grid square (e.g., images that share ice and imaging conditions) on both sides of a split.

Choosing evaluation metrics

Region-overlap metrics are useful on TileBac but they should not be interpreted in isolation. Because the annotations are compact, convex filled regions simple geometry-based masks can score highly without using information from the actual image. A filled bounding box reaches a Dice coefficient of 0.845 and a convex hull reaches 0.991. These high Dice/IoU values do not mean the method has successfully learned membrane contrast or bounding area. Results from this section should report instance AP from a stated evaluator and compare region-overlap scores with the reference masks in Table 4. Methods that predict membranes as boundaries should be evaluated with a boundary metric, such as symmetric surface distance between

predicted and annotated outlines. Authors using this dataset should state whether their method predicts membranes as filled regions or as boundaries. Centerline Dice (clDice) should not be used as a connectivity measure for these annotations because a filled bounding box scores 0.990.

Dose-dependent analysis

The frame-partitioned flagella movies allow dose-dependent analysis within a single field of view. Comparisons between independent datasets acquired at $40 \text{ e}^-/\text{\AA}^2$ and $4 \text{ e}^-/\text{\AA}^2$ confound dose with differences in specimen, ice thickness, defocus, and field of view. In contrast, cumulative-dose images constructed from one movie show the same cell during the same exposure, and the full-dose annotation remains registered to the component frames. These movies support several analyses. Segmentation performance can be measured as a function of cumulative dose and models trained at one dose can be tested at another. Denoising and restoration methods can also use a partial frame sum as input and the full frame sum as a reference. Because the full-dose annotation remains registered to each partial sum the same reference annotation can be used at doses where the structure is difficult to identify manually.

Open problems

The dataset supports four open analysis problems. (1) Users can compare direct inference on the 4096×4096 -pixel source images with tiled, sliding-window, and stitched-montage inference. Tiled methods require predictions to be mapped back to source-image coordinates, while full-montage analysis increases memory and computational requirements. The source-image record and manifest provide the coordinate mappings needed to compare these workflows. (2) Users can develop and evaluate montage-registration methods. Registration errors can produce artifacts in tile-overlap regions or shift membranes when cells cross stitching seams [16]. The source tiles and stitched montages can be used to measure these errors and determine how they affect downstream segmentation. (3) Users can test model performance across image scales. Bacteria occupy different numbers of pixels in the source tiles and re-tiled montages, so a model trained at one scale may not transfer directly to another. Multiscale architectures may reduce this sensitivity [31–33]. Because both representations come from the same specimens, the records support direct measurement of cross-scale transfer. And (4) users can compare preprocessing and normalization methods with the linked raw, normalized, tiled, and annotated records.

Code Availability

Code for image conversion, dynamic normalization, tiling, partitioning, and evaluation is available in the TileBac code repository cited in the references [26–30]. The unified evaluation used PyTorch 2.5.1, Detectron2 0.6, Ultralytics 8.4.122, SAM3 0.1.4, and pycocotools 2.0.11. Model fine-tuning used the published implementation of each architecture at the version recorded in the training configuration files distributed with the model records. Annotations were created in Roboflow [20], a commercial web service.

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science FWP ERKCZ64, Structure Guided Design of Materials to Optimize the Abiotic-Biotic Material Interface, as part of the Biopreparedness Research Virtual Environment (BRaVE) initiative. This research used sample preparation, electron microscopy, and image-analysis resources at the Center for Nanophase Materials Sciences, a U.S. Department of Energy Office of Science User Facility at Oak Ridge National Laboratory. Electron microscopy data were collected using instrumentation within ORNL’s Materials Characterization Core provided by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the DOE and sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy.

Author Contributions

A.N.W. conceived and supervised the study. L.N.M. and S.S.M. collected and annotated the data, trained and evaluated the models, and contributed equally. C.R.H. supported sample preparation and microscope operation. S.R.B. and A.N.B. prepared bacterial cultures. S.T.R., J.L.M.-F. and R.K.V. provided project direction. L.N.M., S.S.M. and A.N.W. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

References

- [1] S. Zheng et al. Implication of surface properties, bacterial motility, and hydrodynamic conditions on bacterial surface sensing and their initial adhesion. *Frontiers in Bioengineering and Biotechnology*, 9: 643722, 2021.

- [2] L. Porter et al. How long do nosocomial pathogens persist on inanimate surfaces? a scoping review. *Journal of Hospital Infection*, 147:25–31, 2024.
- [3] E. H. L. Chen et al. Visualizing the membrane disruption action of antimicrobial peptides by cryo-electron tomography. *Nature Communications*, 14:5464, 2023.
- [4] T. P. T. Cushnie, N. H. O’Driscoll, and A. J. Lamb. Morphological and ultrastructural changes in bacterial cells as an indicator of antibacterial mechanism of action. *Cellular and Molecular Life Sciences*, 73:4471–4492, 2016.
- [5] R. S. Friedlander, N. Vogel, and J. Aizenberg. Role of flagella in adhesion of *Escherichia coli* to abiotic surfaces. *Langmuir*, 31:6137–6144, 2015.
- [6] S. B. Guttenplan and D. B. Kearns. Regulation of flagellar motility during biofilm formation. *FEMS Microbiology Reviews*, 37:849–871, 2013.
- [7] R. Belas. Biofilms, flagella, and mechanosensing of surfaces by bacteria. *Trends in Microbiology*, 22: 517–527, 2014.
- [8] A. Archit et al. Segment anything for microscopy. *Nature Methods*, 22:579–591, 2025.
- [9] U. Israel et al. A foundation model for cell segmentation, 2023. URL <https://arxiv.org/abs/2311.11004>.
- [10] VitrifyAI. ACORN: Annotate, curate, observe, review, navigate. GitHub software repository, 2026. URL https://github.com/vitrifyai/acorn_software. Accessed 25 September 2026.
- [11] O. Ronneberger, P. Fischer, and T. Brox. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
- [12] I. Oztel, G. Yolcu, I. Ersoy, T. White, and F. Bunyak. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1195–1200. IEEE, 2017.
- [13] S. S. Madugula et al. Automated bacterial identification and morphological feature analysis in low-dose cryo-EM using YOLOv11. *Advanced Intelligent Discovery*, page e202500241, 2026.
- [14] F. C. Akyon, S. O. Altinuc, and A. Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing*, pages 966–970. IEEE, 2022.
- [15] Z. Zhang et al. Unsupervised trajectory optimization for 3d registration in serial section electron microscopy using neural ODEs. In *Advances in Neural Information Processing Systems*, volume 38, pages 127997–128043, 2025.
- [16] F. Feng et al. Electron microscopy image registration with transformers. In *Neural Information Processing*, pages 14–25. Springer, 2023.
- [17] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng. A large expert-curated cryo-EM image dataset for machine learning protein particle picking. *Scientific Data*, 10:392, 2023.
- [18] A. Iudin, P. K. Korir, J. Salavert-Torres, G. J. Kleywegt, and A. Patwardhan. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods*, 13:387–388, 2016.

- [19] A. N. Bible et al. A carotenoid-deficient mutant in *Pantoea* sp. YR343, a bacteria isolated from the rhizosphere of *Populus deltoides*, is defective in root colonization. *Frontiers in Microbiology*, 7:491, 2016.
- [20] F. Ciaglia et al. Roboflow 100: a rich, multi-domain object detection benchmark, 2022. URL <https://arxiv.org/abs/2211.13523>.
- [21] R. Khanam and M. Hussain. YOLOv11: an overview of the key architectural enhancements, 2024. URL <https://arxiv.org/abs/2410.17725>.
- [22] G. Jocher and J. Qiu. Ultralytics YOLO11, 2024.
- [23] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee. YOLO26: key architectural enhancements and performance benchmarking for real-time object detection, 2025. URL <https://arxiv.org/abs/2509.25164>.
- [24] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2, 2019. URL <https://github.com/facebookresearch/detectron2>.
- [25] N. Carion et al. SAM 3: Segment anything with concepts, 2025. URL <https://arxiv.org/abs/2511.16719>.
- [26] L. N. Massenburg et al. TileBac ultralow-dose benchmark dataset. Hugging Face Dataset, revision 1545f0662c8fc85783793b05627b87435e7df79e, 2026. URL <https://huggingface.co/datasets/LynnMass/tilebac-ULDM-benchmark-dataset>.
- [27] L. N. Massenburg et al. TileBac ultralow-dose source tile images. Hugging Face Dataset, revision 59affd070acaa4af70ecfcb7a4752b28246e7807, 2026. URL <https://huggingface.co/datasets/LynnMass/tilebac-ULDM-tiles>.
- [28] L. N. Massenburg et al. TileBac stitched montage images. Hugging Face Dataset, revision 283085ec1e14bd9195a508def9cc98273c2c3c93, 2026. URL <https://huggingface.co/datasets/LynnMass/tilebac-stitched-montage>.
- [29] L. N. Massenburg et al. TileBac low-dose bacteria and flagella dataset, including frame-partitioned MRC movies. Hugging Face Dataset, revision a9ef6021d7af7af4ebc38ff4dcc35901dfb0df28, 2026. URL <https://huggingface.co/datasets/LynnMass/tilebac-flag-dataset>.
- [30] L. N. Massenburg et al. TileBac fine-tuned segmentation model records. Hugging Face Model repositories, 2026. URL <https://huggingface.co/LynnMass>. Model revisions and checksums are recorded in the associated code-release metadata.
- [31] J. Wang et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2020.
- [32] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [33] T.-Y. Lin et al. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944. IEEE, 2017.