# WormID-Benchmark: Extracting Whole-Brain Neural Dynamics of *C. elegans* At the Neuron Resolution

**Jason Adhinarta**[1] **Jizheng Dong**[2*] **Tianxiao He**[2*] **Junxiang Huang**[1*] **Daniel Sprague**[4*] **Jia Wan**[5] **Hyun Jee Lee**[6] **Zikai Yu**[6] **Hang Lu**[6] **Eviatar Yemini**[7] **Saul Kato**[3†] **Erdem Varol**[2†] **Donglai Wei**[1†]

[1] Boston College [2] New York University [3] UC San Francisco [4] Harvard Medical School [5] Harbin Institute of Technology, Shenzhen [6] Georgia Tech [7] UMass Chan Medical School

**The nematode C. elegans is a well-studied model organism for characterizing the structure, connectivity, and function of a complete nervous system. Recent technical breakthroughs in 3D light microscopy and fluorescent protein tagging of individual neurons have brought us closer to capturing the neural dynamics of the worm at whole-brain resolution. Nevertheless, capturing a complete map of neural dynamics using these high-resolution recordings requires solving three specific challenges: i) detection of individual neurons in fluorescence videos, ii) identification of these neurons according to their anatomically defined classes, and iii) tracking of neural positions over time. Successfully addressing these challenges with high sensitivity, specificity, and throughput can enable us to analyze a large population sample, providing unprecedented insights into the structure and function of an entire brain at single-neuron resolution—a feat previously unaccomplished in any organism. To facilitate this scientific goal, we have curated publicly available annotated datasets from 118 worms across five distinct laboratories and established systematic benchmarks, decomposing the overarching objective into three well-defined tasks: i) neural detection, ii) identification, and iii) spatiotemporal tracking. Our preliminary analysis has revealed considerable room for improvement in existing state-of-the-art computational methods. Consequently, we envision that our WormID-Benchmarkcan catalyze efforts by a broad audience specializing in computer vision to develop robust and accurate methods that significantly enhance the throughput of generating annotated whole-brain neural dynamics datasets. We make our benchmark results reproducible; our code is publicly available at https://github.com/focolab/WormND.**

## 1. Introduction

Recent developments in 3D whole-brain microscopy imaging have enabled the comprehensive characterization of neural dynamics in complete nervous systems of small model organisms Schrödel et al. (2013); Venkatachalam et al. (2016); Yemini et al. (2021); Haspel et al. (2023). The nematode *C. elegans*, a millimeter-sized worm, is particularly suitable for such studies due to its cells' unique and stereotyped identities, facilitating comparisons across multiple trials and animals.

Thanks to hardware breakthroughs, it is now relatively easy and quick to capture whole-brain microscopy videos of hundreds, if not thousands, of worms in raw form. However, the critical bottleneck is the annotation of these datasets with individually defined neurons across time. The current approach adopted by experimentalists—to manually detect and identify each neuron, then track it in videos for extraction of neural dynamics—has very low throughput, limiting the effective sample size of usable datasets for testing downstream scientific hypotheses. Thus, to match the throughput of annotation with the throughput of image acquisition, it is necessary to solve three subproblems: neural detection, neural identification, and spatiotemporal tracking of neural positions.

There has been limited success in automating these tasks due to the diversity in image format and appearance, biological variability, and batch effects across different labs and experimental setups Toyoshima et al. (2016, 2020); Varol et al. (2020); Nejatbakhsh et al. (2020b,a); Wu et al. (2022). Furthermore, the neural dynamics themselves can act as obstacles to automated annotation, as the brightness and appearance of neurons can widely vary over time based on their firing activity, making it challenging to accurately detect, identify, and track this neurons Nguyen et al. (2016).

Equal contributions
[†]Corresponding authors: saul.kato@ucsf.edu, ev2240@nyu.edu, donglai.wei@bc.edu

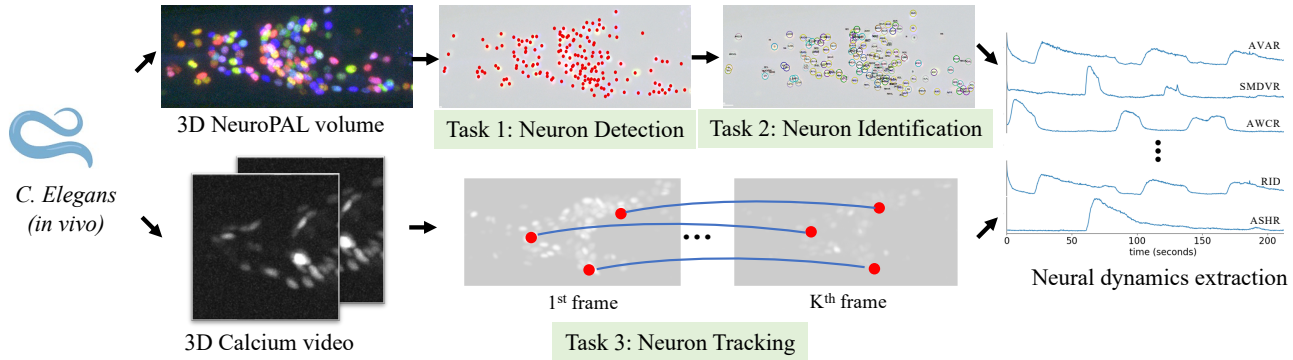



**Figure 1.** Task overview. To extract neural dynamics from microscopy images of *in vivo C. elegans*, **WormID-Benchmark** benchmarks each step of the pipeline: neuron *detection*, *identification*, and *tracking*.

To propel a community effort to improve the accuracy and throughput of annotating microscopy videos of *C. elegans* at the single-neuron resolution, we have curated several publicly available datasets with complete or semi-complete ground truth annotations that span multiple imaging setups and laboratories. Recent efforts have pre-processed the data acquired from various microscopy setups, experimental conditions, and protocols across labs, demonstrating that retraining automatic approaches on large and diverse datasets significantly improves the generalizability and accuracy of these tools Sprague et al. (2024). However, there remains a lack of clearly defined tasks, standardized evaluation metrics, and benchmarks for state-of-the-art computer vision methods in the context of worms.
To this end, we introduce WormID-Benchmark, a carefully designed benchmark for all three sub-problems (i) detection, ii) identification, and iii) tracking on curated publicly available datasets. We hope this benchmark will catalyze the development of next-generation tools to robustly extract neural dynamics from 3D microscopy videos with high throughput.

## 2. Related Work

Although numerous benchmark datasets exist for cell analysis in 2D images Ma et al. (2024), there are few available for 3D light microscopy image volumes and videos due to the costly annotation process. We limit our review to light microscopy which is standard for *in vivo* live imaging.
**3D Cell Detection and Segmentation Benchmarks.** 3D Cell Detection is always the first step for any neuron-level imaging-related task. Previously Alwes *et al.* Alwes et al. (2016) collected and annotated volumes of 3D images of histone fluorescent protein expression in Parhyale. Another valuable resource is the BBBC050 dataset, consisting of nuclei of mouse embryonic cells Tokuoka et al. (2020) to study cell dynamics and developmental processes. However, there are few large-scale datasets covering neuronal volumes in *C. elegans* Ruszczycki et al. (2019) with comprehensive annotations for cell detection, identification, and cell tracking. Unlike previous datasets, *C. elegans* microscopy images tend to have lower resolution and have more non-cellular structure, making the segmentation task challenging.
**3D Neuron Identification Benchmarks.** Neuron identification from fluorescent microscopy images is notoriously difficult. Thanks to the recently developed NeuroPAL Yemini et al. (2021), every single neuron is deterministically colored by a stereotyped fluorescent RGB color barcode in living animals that is identical across all stages of development in both sexes. While this does not enable fully automatic identification of each neuron since there may be similar colored neurons in close proximity, it greatly simplifies the identification problem since the color code helps eliminate a wide range of potential labels for a neuron.
**3D Cell Tracking Benchmarks.** The popular Cell Tracking Challenge (CTC) Maška et al. (2023) consists of ten 3D time-lapse microscopy datasets with various non-neuron cells. The challenge provides dense segmentation annotations at each time frame, along with tracking annotations that indicate correspondences between frames and cell-splitting events. For neural dynamics extraction, annotating cell center points is sufficient while saving much manual labeling labor. To the best of our knowledge, our dataset of 36 worm videos is the only dataset with annotations for neuron subtype identities, their tracked positions, and their neural activity.

## 3. WormID-Benchmark

### A. Dataset Overview

For this benchmark, we use the wormID corpus of NeuroPAL and Calcium Imaging datasets curated in Sprague *et al.* Sprague et al. (2024). This corpus consists of seven datasets from five different labs around the world comprising

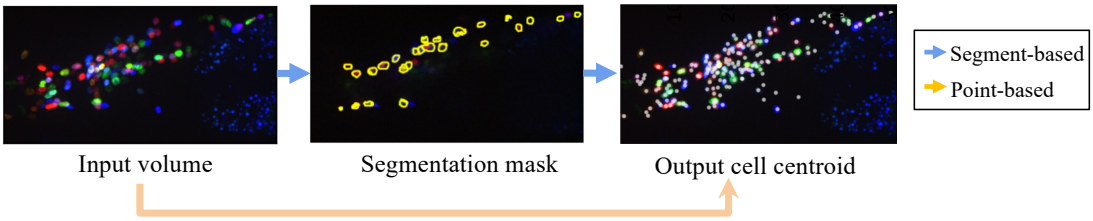



**Figure 2.** Neuron detection paradigms (Task 1). Given input color microscopy imaging volumes, cell centroids can be directly inferred (orange arrow) or cell segmentation can be executed first and then cell centroids can be inferred from segmentation maps. Note the presence of non-neuronal objects (blue points on the right side) in the field of view that are prone to be detected as false positives.

118 total worms with a mixture of NeuroPAL structural images and whole-brain calcium activity time series. The NP, HL, SF, and EY datasets come from previous publications Yemini et al. (2021); Chaudhary et al. (2021); Atanas et al. (2023); Nejatbakhsh et al. (2020b) while KK, SK1, and SK2 are published for the first time as part of the wormID corpus. These datasets represent a curated set of publication quality worms from each lab: i.e. images with high enough quality to extract meaningful information about neural structure and/or activity. These datasets are presented as a singular corpus with data associated with each worm stored in the standardized Neurodata Without Borders (NWB) file format and are publicly available for download or streaming on the DANDI archive Rübel et al. (2022); Halchenko et al. (2024). Further details and code for accessing these datasets can be found at wormID.org and in Sprague *et al.* Sprague et al. (2024). In Tab. 1, we list the statistics for the large corpus of datasets we use to train and benchmark the models across the three tasks presented here.

| Dataset | resolution ($\mu m/px$) | 3D NeuroPAL | | | 3D Calcium Video | | |
|---|---|---|---|---|---|---|---|
| | | #samples | #DET | #ID | #samples | #frames | #TRAJ |
| NP Yemini et al. (2021) | 0.21×0.21× 1.02 | 10 | 193 | 190 | × | × | × |
| HL Chaudhary et al. (2021) | 0.33 ×0.33× 1.0 | 10 | 119 | 64 | × | × | × |
| SF Atanas et al. (2023) | 0.54 ×0.54× 0.54 | 38 | 70 | 70 | × | × | × |
| EY Nejatbakhsh et al. (2020b) | 0.27 ×0.27× 1.5 | 21 | 177 | 175 | 21 | 961 | 177 |
| KK Wen et al. (2021) | 0.32 ×0.32× 1.5 | 9 | 154 | 154 | 9 | 1646 | 155 |
| SK1 Sprague et al. (2024) | 0.32 ×0.32× 1[1] | 20 | 111 | 48 | 4 | 1500 | 108 |
| SK2 Sprague et al. (2024) | 0.32 ×0.32× 0.75[1] | 10 | 173 | 49 | 2 | 3000 | 131 |

**Table 1.** Summary of WormID-Benchmark datasets and their parameters. #DET, #ID, and #TRAJ represent the average number of annotated detection, identification, and trajectory across samples respectively.

**Benchmark exclusion criteria.** For the detection and ID tasks, we omit 14 worms that contain significant non-linear deformities or color artifacts since many of the models we test assume that inputs are roughly aligned along the principal axes of the worm body and rely on color information. This leaves 104 remaining datasets that we use to train and benchmark each of the detection and ID approaches. For the tracking task, we omit 38 worm videos in SF Atanas et al. (2023) which does not provide enough information to extract ground truth annotations.

**Dataset split.** Most previous approaches to the tasks outlined here typically only train and test on a small subset of data, usually all collected by one lab, limiting the generalizability of these approaches. We split the dataset into five equal groups with equal representation from each dataset in each group. For each task, we report cross-validated results for each metric. For each group, we train the model on the four other groups and then test on the held-out group. For the detection and ID tasks, the whole NP dataset, which is the standard reference dataset for NeuroPAL imaging, is used for training all groups but is not used in evaluation. Thus, we report an average of cross-validated metrics across 94 worms for the detection and ID task and 36 worms for the tracking task.

## B. Task 1: 3D Neuron Detection in 3D NeuroPAL Volumes

The first step in curating fully annotated neural dynamics datasets in *C. elegans* is to detect and segregate neurons from other non-neural structures in fluorescence microscopy images of the nematode. As there are 302 neurons in the worm, amongst a total of 959 somatic cells Strange (2006), this problem can be thought of as a 3D instance segmentation problem, where we first generate 3D segmentation mask from raw images, and then detect the centroids of objects for each instance.

**Baseline methods.** We approach the detection problem using 3D instance segmentation and we compare it with both classical methods for cell segmentation in microscopy images and general-purpose segmentation tools from computer

[1] Videos in SK1 and SK2 have varying z-resolutions and we refer to Sprague et al. (2024) for details.

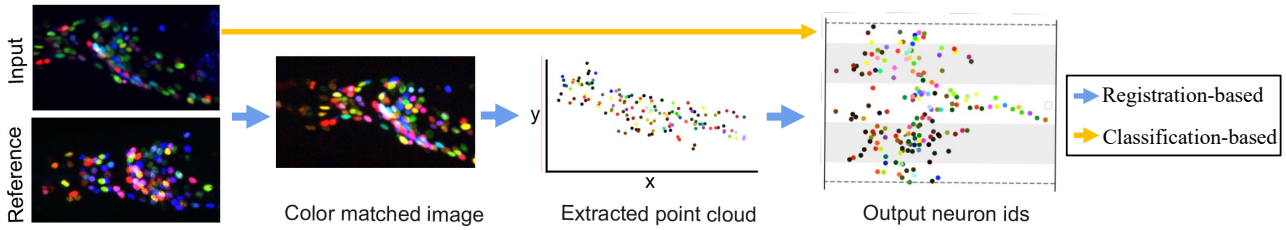



**Figure 3.** Neuron identification paradigms (Task 2). It can be cast either as a point cloud registration problem, using a fully annotated template atlas as a reference, or a multi-way classification problem, labeling each neuron into one of 302 classes.

vision. Here we first compare against the generalist pre-trained *cyto3* model available in CellPose3 Stringer and Pachitariu (2024). The model is pre-trained on 2d microscopy images, applied to all xy, xz, and yz slices independently, and merged into a 3d segmentation mask during inference. We also compare against transformer-based foundation models in computer vision, such as microSAM Archit et al. (2023), which fine-tuned Segment Anything Model Kirillov et al. (2023) on microscopy images.

**Challenges.** Although there have been several advances in automated detection and segmentation of cells and neurons in biological datasets Ulman et al. (2017); Maška et al. (2023), detecting neurons in *C. elegans* poses unique challenges due to the presence of non-neural tissue, cell-like structures that occupy the field-of-view in many of the microscopy datasets that are captured in the worm (Fig.4). Also, since neurons are typically tagged with fluorescent proteins that encode the presence of calcium, the brightness and contrast of cells across may vary, posing an additional challenge in detecting them. It is not uncommon that a certain percentage of cells that should be present are undetectable by the imaging features alone. Furthermore, the heterogeneity of imaging setups across different labs introduces non-biological confounds such as non-linear deformations, varying fluorescence bleaching effects, brightness inhomogeneities, lowered resolution along the depth axis, and other optical effects. Therefore, it is imperative to develop robust methods that can be invariant to such effects, to improve the neural detection throughput across a diverse set of input images.

## C. Task 2: 3D Neuron Identification in 3D Volumes

The 302 neuron nervous system of the worm is eutelic, meaning that these 302 neurons are the same across individuals and each comes with a distinct identity label. The incorporation of four colored fluorophores (neuron-specific combinations of mTagBFP2, CyOFP1, and mNeptune2.5, and pan-neuronal TagRFP-T) within NeuroPAL volumes offers a distinctive fluorescent color barcoding scheme to aid the identification of neuron subtypes. As such, the computational objective, when presented with a 3D NeuroPAL image volume as input and detected locations of neuron centers, is to predict their class label out of a possible 302 classes. Generally, experiments will focus on either the head or tail of the worm. The datasets presented here focus on the head of the worm, which contains a subset of 120-195 neurons depending on how far down the ventral nerve chord the image goes.

**Baseline methods.** The 3D neuron identification problem is generally formulated as a point cloud alignment problem. The test point cloud of neuron center positions is roughly aligned with an initial template and then assignments are made using the Hungarian algorithm Kuhn (1955) based on a chosen distance metric between the test and template. As an initial baseline, we use the non-rigid Coherent Point Drift Myronenko and Song (2010) algorithm which is a standard algorithm for rough point cloud alignment. We then compare three of the main existing approaches to the *C. elegans* 3D neuron classification problem: a Statistical Atlas approach Varol et al. (2020), a graph-based conditional random fields model Chaudhary et al. (2021), and a transformer based neural network model Yu et al. (2021). For each of these methods, we report both the accuracy for a base model as well as the model retrained using the corpus of data presented here and in Sprague et al. 2024 Sprague et al. (2024). For the Statistical Atlas and CRF ID, the base model is trained only on the 10 worms in the NP dataset, which serves as the gold-standard reference dataset for NeuroPAL imaging. For fDNC, the base model is the pre-trained model presented in the paper.

**Task specific pre-processing steps.** Each approach to the Neuron ID problem evaluated here assumes that the input to the model is a point cloud of neuron center xyz coordinates and rgb colors. To obtain these point clouds, the first neuron center positions are detected in the 3D volume using a combination of algorithmic outputs and manual curation as outlined in the detection task section of this paper. Next, we perform color corrections on the raw RGB image to roughly align the color space of the image to the images in the NeuroPAL reference set Yemini et al. (2021) as outlined in Sprague *et al.*Sprague et al. (2024). After correction, the color values in each channel are Z-scored. Color values are then extracted by taking a median sample of a small area of voxels surrounding each neuron center. Once these point clouds have been extracted from each dataset in the corpus, we roughly align the principal axes of each worm using a learned affine transformation as described in Varol et al. (2020). The accuracy of each approach is highly sensitive to the initial alignment, so we chose to start with a roughly aligned input to more directly compare

how each model captures the inherent biological and experimental diversity across datasets rather than just how well they can learn this rough alignment.

**Challenges.** While the stereotyped nature of the *C. elegans* nervous system theoretically makes this task fairly straightforward, the inherent variability caused by natural variation in cell position, deformities and movement of the worm body, and variation in imaging conditions makes this problem deceptively difficult. Neurons are not always in the same position relative to one another, the detection algorithm may miss some neurons and falsely identify non-neurons, and many neurons are very difficult to distinguish with 100% confidence even for an expert annotator. NeuroPAL simplifies this problem greatly by adding a distinctive fluorescent color to each neuron subtype, but annotation still requires many hours of work by an expert annotator and it is still not always possible to unambiguously resolve all neuron identities even for a human annotator.

### D. Task 3: 3D Neuron Tracking in 3D Calcium Videos

Recent strides in calcium imaging technology have ushered in a new era wherein the simultaneous recording of neural activity from every neuron within the *C. elegans* nervous system has become achievable. This technological advancement affords unprecedented insights into the dynamics of the entire nervous system's activity. The process of extracting such activities entails the meticulous tracking of each localized neuron across all frames in the video, coupled with the estimation of its activity based on image intensity.

**Baseline methods.** There are two popular paradigms for video object tracking Feichtenhofer et al. (2017): detect first and then track or propagate based on image appearance. We here only investigate the first approach, as the neurons have a less distinctive texture and can have low image intensity value when non-excited in the video. Thus, it is more challenging for the propagation approach. For the detect-to-track approach, Ultrack is a recent method that leverages ultrametric contour maps and formulates the cell tracking problem as an integer linear program (ILP) Bragantini et al. (2023). By modeling the segmentation selection process as an ILP, Ultrack maximizes the overlap between adjacent frames to ensure accurate cell tracking and achieves competitive results on various benchmarks, including the 2D epithelial cell benchmark and 3D datasets from the CTC. 3DeeCellTracker Wen et al. (2021) uses a segment-then-track pipeline, using StarDist Weigert et al. (2020) as a backbone for segmentation and a Feed Forward Network (FFN) to generate a matching matrix to align detections between frames. We use the models pre-trained on *C. elegans* data: "worm1" for StarDist and "worm3" for the FFN, running inference in "single"-mode.

**Challenges.** This computational undertaking is inherently intricate, owing to a multitude of factors, including the presence of intrinsic biological noise, deformations, sporadic absence of neurons within specific datasets, and the utilization of varied acquisition conditions across different laboratories. For instance, previous methods Nejatbakhsh et al. (2020b); Venkatachalam et al. (2016) have utilized pose registration techniques to obtain canonical neuronal positions from differing worm deformations. Moreover, due to the nature of calcium imaging, non-excited neurons are often unobservable in the calcium indicator channel, although they may be identifiable in the RGB fluorescence channels that are invariant to excitation.

## 4. Benchmark Results

All inference and model fine-tuning are run on 4 NVIDIA A100 GPUs. The implementations use the recommended versions of PyTorch and other libraries used in the official repositories. We will release our repository to allow the reproducibility of all the experiments below. To alleviate the bias from the data split, we report 5-fold cross-validation results with the standard deviation.

### A. Task 1: 3D Neuron Detection

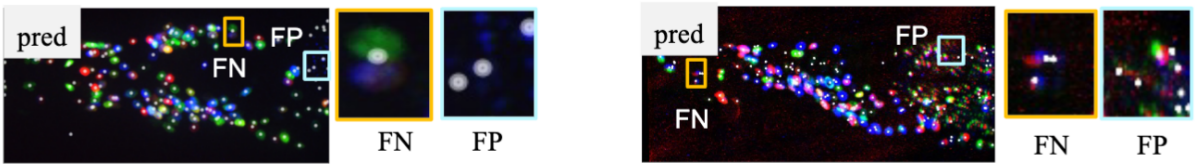
**Evaluation metrics.** We adopted the evaluation metrics from the Cell Detection from Cell-Tissue Interaction (OCELOT) challenge Ryu et al. (2023) in MICCAI 2023. We use mean F-1 score as primary metrics, along with precision and recall for cell detection. As our neuron annotations are composed of points instead of segments, it becomes imperative to establish a distance threshold for assessing detection accuracy as an analog for IoU matching. In the evaluation stage, we first compute the deviation distances from each ground truth point to its corresponding predicted point, taking into account the resolutions along different dimensions, $\alpha_{xy}$ and $\alpha_z$. This is expressed as

$$d_i = \sqrt{\alpha_{xy}^2(x_i - \hat{x}_i)^2 + \alpha_{xy}^2(y_i - \hat{y}_i)^2 + \alpha_z^2(z_i - \hat{z}_i)^2}. \quad \textbf{(1)}$$

Such a calculated deviation distance is then compared with the pre-defined distance threshold $d_{\text{th}}$. A predicted detection is considered correct if it falls within $d_{\text{th}}$ from the GT point, thus accurately reflecting the spatial precision required by point-based annotations. For each predicted cell, if there is no ground truth cell within the valid distance, it is considered a false positive. Conversely, if there is one or more GT cells within the threshold, it is considered a

| Method | $d_{\text{th}}$=3$\mu m$ | | | $d_{\text{th}}$=6$\mu m$ | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| CellPose Stringer and Pachitariu (2024) | **0.45±0.02** | **0.58±0.01** | **0.47±0.01** | **0.50±0.03** | **0.65±0.01** | **0.53±0.02** |
| Micro-SAM Archit et al. (2023) | 0.41±0.01 | 0.43±0.02 | 0.39±0.01 | 0.49±0.01 | 0.58±0.03 | 0.48±0.01 |

**Table 2.** Neuron detection results (Task 1). We report metrics using two different distance thresholds.




(a) Example (F1=0.72) (b) Example 2 (F1=0.52)

**Figure 4.** Qualitative neuron detection results (Task 1). We zoom into the false positive and false negative detections for two representative datasets.

true positive, with the nearest ground truth cell matched and excluded from further matching. Analogous to different IoU thresholds in CTC, here we employ distance thresholds of $d_{\text{th}} \in \{3\mu m, 6\mu m\}$ to evaluate the model performance under varying proximity criteria.

**Quantitative results.** We evaluated the performance of two pre-trained models and established a baseline for comparison with other models in Tab. 2. For Cellpose, we use the grayscale image as input since the pre-trained model was trained on two-channel images where the first channel is for nuclei segmentation (check again here). The Cellpose model achieves better out-of-the-box performance.

**Qualitative results.** As shown in Fig. 4, the false positive errors are often caused by non-cellular structures in the image, particularly in the green and blue channels. These structures can confuse the segmentation model into misclassifying them as cells, leading to incorrect prediction. For the false negative errors, ground truth neurons are often missed due to their low intensity or low contrast, making distinguishing from the noisy background difficult. We provide more examples in Fig. 5.

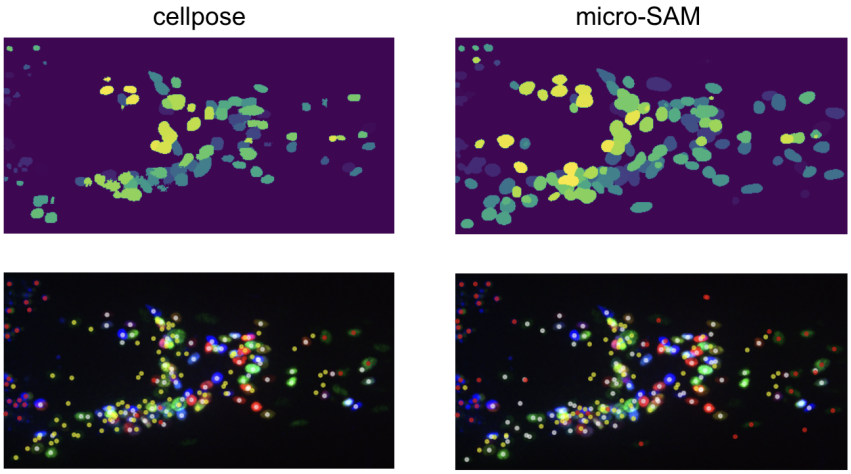



**Figure 5.** Comparison of segmentation mask predicted by cellpose and micro-SAM. The bottom two figures display the detected points (the little points in the center of neurons): white points indicate true positives, red points indicate false positives, and yellow points indicate false negatives.

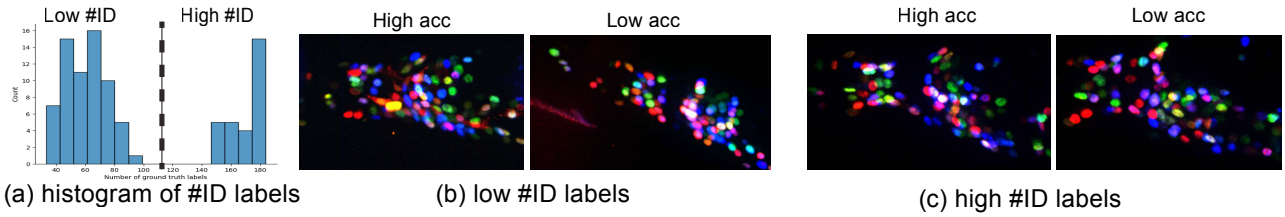
Additional fine-tuning of CellPose and Micro-SAM did not converge on a proper outcome due to these methods not being natively designed for 3D volumetric images. Nevertheless, the pre-trained models seem to yield reasonable results due to post-hoc stitching of 2D segmentation maps.

## B. Task 2: 3D Neuron Identification

**Evaluation metrics.** The evaluation metric for the 3D neuron identification task is a simple accuracy metric for the proportion of neuron centers with ground truth labels in the test image that is correctly identified by the algorithm.

| Method | Low #ID labels (easy) | | High #ID labels (hard) | | Overall | |
|---|---|---|---|---|---|---|
| | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| CPD Myronenko and Song (2010) | 0.39±0.15 | 0.65±0.14 | 0.39±0.09 | 0.67±0.10 | 0.39±0.14 | 0.65±0.13 |
| fDNCYu et al. (2021) | 0.52±0.04 | 0.73±0.04 | 0.29±0.07 | 0.50±0.09 | 0.45±0.03 | 0.66±0.03 |
| fDNC † Yu et al. (2021) | 0.60±0.14 | 0.79±0.11 | 0.38±0.16 | 0.61±0.20 | 0.53±0.18 | 0.74±0.17 |
| Statistical Atlas Varol et al. (2020) | 0.40±0.12 | 0.64±0.12 | 0.41±0.08 | 0.72±0.05 | 0.41±0.11 | 0.66±0.11 |
| Statistical Atlas† Varol et al. (2020) | 0.66±0.14 | 0.89±0.07 | 0.54±0.06 | **0.79±0.06** | 0.62±0.13 | 0.86±0.08 |
| CRF ID Chaudhary et al. (2021) | 0.59±0.20 | 0.84±0.19 | 0.47±0.07 | 0.73±0.05 | 0.55±0.18 | 0.81±0.17 |
| CRF ID† Chaudhary et al. (2021) | **0.81±0.09** | **0.95±0.05** | **0.57±0.06** | 0.78±0.04 | **0.74 ±0.14** | **0.89±0.09** |

**Table 3.** Neuron identification results (Task 2). Mean top-1 and top-5 accuracies are reported for both pre-trained and re-trained (†) models.




(a) histogram of #ID labels (b) low #ID labels (c) high #ID labels

**Figure 6.** Qualitative neuron identification results (Task 2). We divide the samples into low # and high # of ID labels and visualize results with high and low identification accuracies.

We report this accuracy metric for both the top-1 ranked assignment and the top-5 ranked assignments. Ranked assignments are obtained by setting the costs for each initial match in the Hungarian algorithm to infinity and then rerunning the Hungarian algorithm. We repeat this process $k$ times to get the top-k ranked assignments. We also separate datasets into 'low # of labels' and 'high # of labels' referring to the number of ground truth labels present in the dataset. Datasets with fewer ground truth labels will naturally have higher accuracy because, generally, the first labels assigned by an expert annotator are the easiest neurons to identify in the image. Datasets that are completely labeled or near completely labeled typically have a lower accuracy because of the inherent difficulty in labeling many of the more ambiguous neurons in the worm. For CPD and fDNC, inference is done by matching a test sample to a template worm, so the accuracy measure is sensitive to the choice of the template. Both of these models attempt to match every neuron in the test sample with a neuron in the template and provide a confidence probability to each of these matchings. For each test sample, we report the accuracy of the template with the highest average probability.

**Quantitative results.** As shown in Fig. 6, all methods have higher accuracy with a low number of labels compared to a high number of labels. Since identification is a classification task, performance improves when there are more candidate choices, which explains the higher accuracy in Top-5 compared to Top-1. The traditional methods (CPD, Statistical Atlas) have relatively lower accuracy than the deep learning methods (CRF ID, fDNC) when they are not trained on all datasets. However, there is a significant improvement in their performance after retraining on the full datasets, especially for the CRF ID, which has the best performance among all methods.

**Qualitative results.** The CPD method does not use additional information when performing the identification task, which might serve as a baseline for all methods. The other three methods show an increase in accuracy due to the use of additional information, including color and probability distribution. The fDNC uses a transformer network to decode the matching information, but it is sensitive to pre-processing for the orientation of the worm body, which might be one of the reasons for the low accuracy in our test. The Statistical Atlas method uses a statistical model, while the CRF ID method utilizes a graphical model. The high accuracy of both methods after re-training demonstrates the importance of training on diverse datasets. We show some examples of right prediction (true positive) and wrong prediction (false positive and false negative) for the neuron detection task in Fig. 7 and Fig. 8.

## C. Task 3: 3D Neuron Tracking

**Evaluation metrics.** We adopted the evaluation metrics from the Cell Tracking Challenge Maška et al. (2023), specifically Detection Accuracy (DET) and Tracking Accuracy (TRA). As in Task 1, setting a distance threshold is crucial to assessing detection and tracking accuracy. Therefore, we utilize two distance thresholds, $d_{\text{th}} \in \{3\mu m, 6\mu m\}$, to evaluate the model's performance according to the deviation distance formula outlined in Eq Eq. (1).

**Quantitative results.** We report our metrics in Tab. 4. Ultrack consistently achieves higher tracking and detection accuracies on both distance thresholds compared to 3DeeCellTracker across the datasets, due to a relatively high number of detections in each time frame in line with the number of neurons in the volume. We hypothesize that using the "ensemble"-mode provided by 3DeeCellTracker can increase its performance.

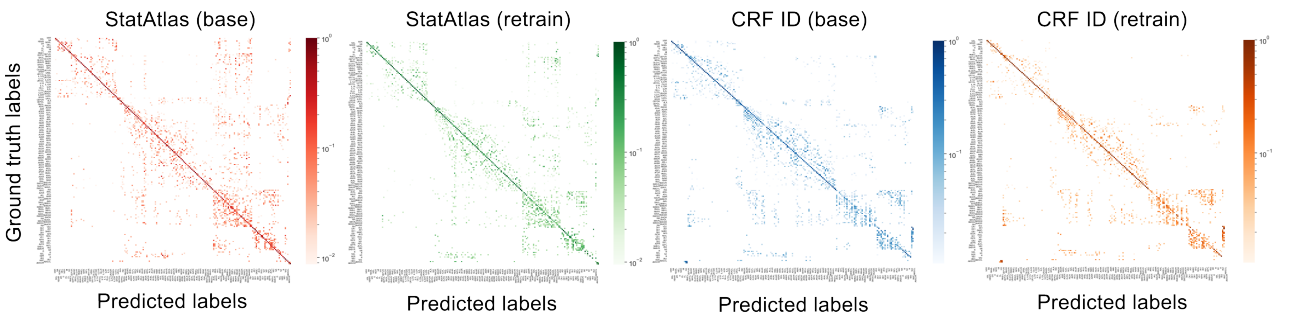



**Figure 7.** Confusion matrices of neuron misidentification for statistical atlas and CRF ID. Neurons are ordered first by ganglion and then from anterior to posterior within each ganglion. Values along the diagonal of the confusion matrix represent correct labels. Most errors occur with nearby neurons in the same ganglion across each model.

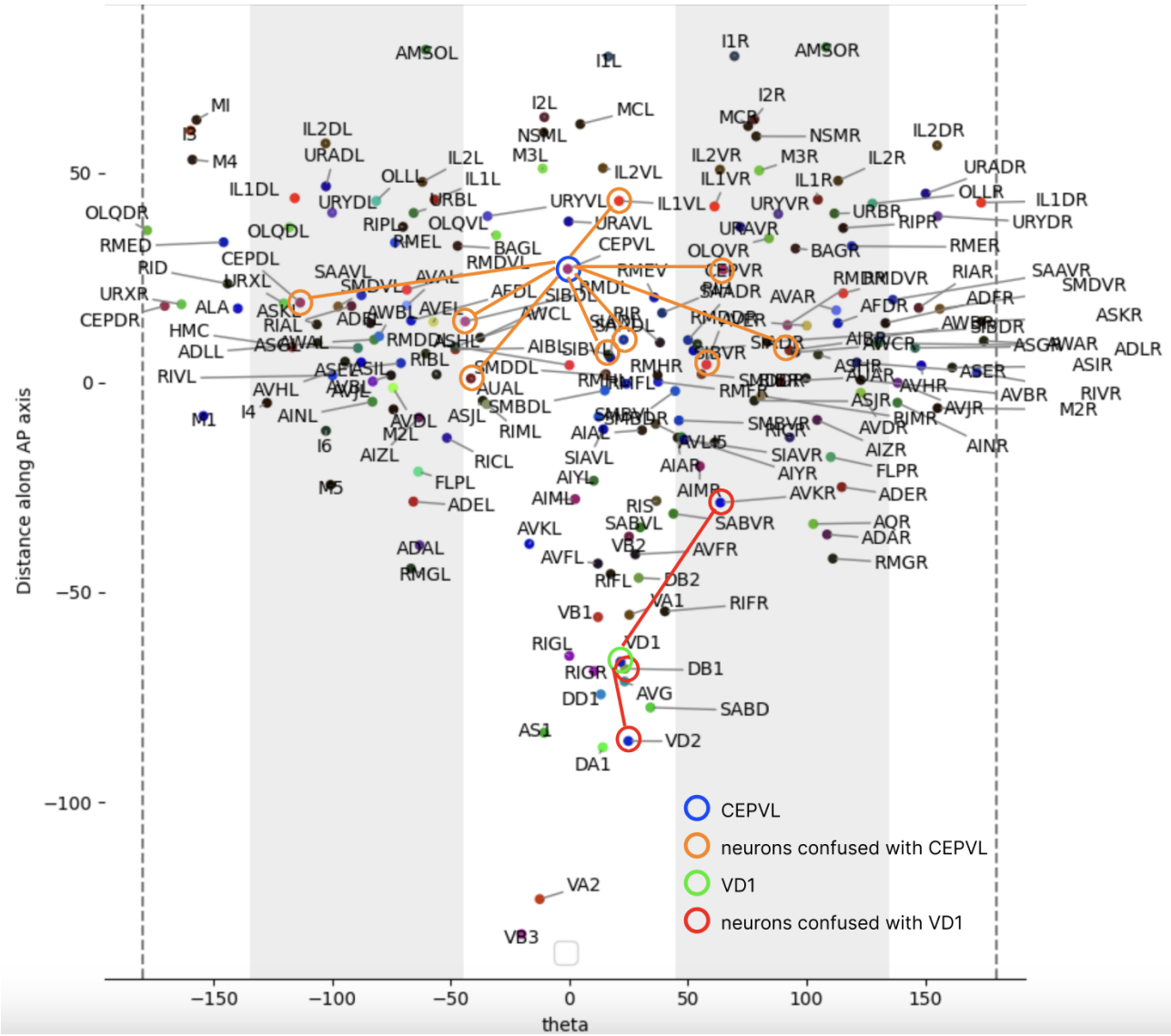


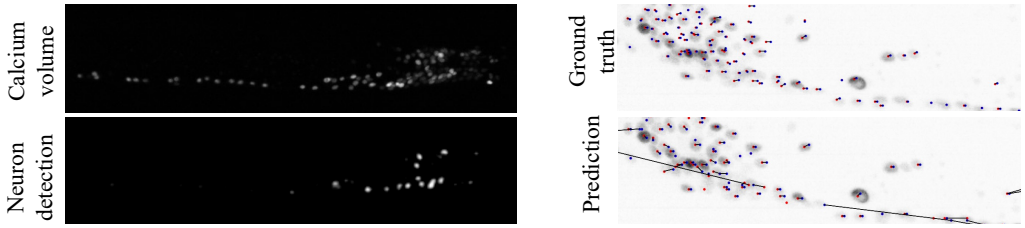

**Figure 8.** Visualization of atlas from statistical atlas model with select neurons CEPVL (blue) and VD1 (green) overlaid on top. Neurons that are commonly confused for these two select neurons are circled in orange and red respectively

Moreover, we retrain the StarDist backbone Weigert et al. (2020) of 3DeeCellTracker Wen et al. (2021) for 1000 epochs to obtain significantly better performance compared to the pre-trained model.

**Qualitative results.** As shown in Fig. 9a, The poor performance of 3DeeCellTracker can be attributed to the low number of detections in each frame; on average 70 segments are generated in each time step, but 150 are expected. This happens despite setting a low probability threshold of 0.01 for the StarDist. Moreover, the inconsistent number of detections in each frame leads to extraneous links between time steps as seen through the long trajectories generated

| Method | $d_{\text{th}}$=3$\mu m$ | | $d_{\text{th}}$=6$\mu m$ | |
|---|---|---|---|---|
| | DET | TRA | DET | TRA |
| Ultrack Bragantini et al. (2023) | **0.70±0.04** | **0.66±0.05** | **0.81±0.05** | **0.78±0.05** |
| 3DeeCellTracker Wen et al. (2021) | 0.28±0.04 | 0.26±0.03 | 0.34±0.05 | 0.32±0.04 |
| 3DeeCellTracker† Wen et al. (2021) | 0.43±0.07 | 0.40±0.07 | 0.51±0.09 | 0.47±0.08 |

**Table 4.** Neuron tracking results (Task 3). We report metrics using two different distance thresholds on both pre-trained and re-trained (†) models using 5-fold cross-validation.




(a) Missing detections (b) Linking errors

**Figure 9.** Qualitative neuron tracking results (Task 3) for 3DeeCellTracker Wen et al. (2021). The errors arise from (a) the detection module probability maps being insensitive to less bright neurons and (b) the linking module due to inconsistent detections between time frames forcing unfavorable matches. In (b) red and blue points denote detections at time $t$ and $t+1$ respectively.

by the matching as seen in Fig. 9b.

# 5. Limitation and Discussion

This benchmark, WormID-Benchmark, establishes a structured framework for addressing the key bottlenecks towards obtaining high-throughput capture of annotated neural dynamics of the *C. elegans* nervous system. Namely, we have split the overall problem of annotation of functional activity data of the worm into the sub-problems of neural detection, identification, and tracking and measuring the performance of state-of-the-art methods across these tasks. Although existing tools are capable of performing moderate levels of detection, identification, and tracking, their performance varies significantly across different imaging setups and conditions, highlighting the need for either more advanced domain-specific methods, or foundation models that are trained on large corpora, potentially for other tasks, that are fine-tuned for this model organism.

However, the introduction of standardized evaluation metrics and clearly defined tasks within WormID-Benchmark encourages transparency and comparability in the future development of tools. By providing a common ground for testing and comparison, we aim to foster a collaborative environment, where researchers can iteratively improve existing methods and innovate new solutions that will advance our understanding of neural dynamics at an unprecedented scale, resolution, and completeness level.

Moreover, WormID-Benchmarkhas several limitations. Although we curate a large and diverse dataset of *C. elegans* microscopy videos, these may not span the entire space of experimental conditions and optical hardware that can be observed across the entire *C. elegans* community. However, we envision that the community development of better tools will be complemented by more rapid annotation of novel datasets from new labs and experimental conditions that can persistently be incorporated into newer versions of this benchmark.

**Potential negative societal impacts.** We do not envision any direct negative societal impacts arising from our research with the WormID-Benchmark, as it is focused exclusively on the nematode *C. elegans*. This organism is phylogenetically distant from humans, with a simple and entirely mapped neural circuitry that does not parallel the complexity of human neurobiology. Consequently, the techniques and findings from this research are unlikely to be directly applicable or transferrable to higher organisms, including humans, thereby minimizing concerns related to ethical issues or misuse. This research is intended to advance basic scientific understanding and improve methodologies within the scope of invertebrate neuroscience.

# 6. Bibliography

Alwes, F., Enjolras, C., and Averof, M. (2016). Live imaging reveals the progenitors and cell dynamics of limb regeneration. *Elife*, 5: e19766.

Archit, A., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Gupta, S., Dengel, A., Ahmed, S., and Pape, C. (2023). Segment anything for microscopy. *bioRxiv*. doi: 10.1101/2023.08.21.554208.

Atanas, A. A., Kim, J., Wang, Z., Bueno, E., Becker, M., Kang, D., Park, J., Kramer, T. S., Wan, F. K., Baskoylu, S., et al. (2023). Brain-wide representations of behavior spanning multiple timescales and states in c. elegans. *Cell*, 186(19):4134–4151.

Bragantini, J., Lange, M., and Royer, L. (2023). Large-scale multi-hypotheses cell tracking using ultrametric contours maps. *arXiv preprint arXiv:2308.04526*.

Chaudhary, S., Lee, S. A., Li, Y., Patel, D. S., and Lu, H. (2021). Graphical-model framework for automated annotation of cell identities in dense cellular images. *Elife*, 10:e60321.

Feichtenhofer, C., Pinz, A., and Zisserman, A. Detect to track and track to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, (2017).

Halchenko, Y., II, J. T. W., Christian, H., Ghosh, S., Sharda, S., Jarecka, D., Baker, C., Chiquito, D., Dichter, B., Nesbitt, J., Gunalan, K., Dedicke, M. E., Macdonald, A., VanDenburgh, M., Meyer, K., LaManna, D., To, I., Grauer, M., Saunders, J., Flynn, G. M., Sprenger, J., Migrator, L., Górny, M., Lo, N., and Ly, R. dandi/dandi-cli: 0.62.1, (2024). URL https://doi.org/10.5281/zenodo.11268706.

Haspel, G., Boyden, E. S., Brown, J., Church, G., Cohen, N., Fang-Yen, C., Flavell, S., Goodman, M. B., Hart, A. C., Hobert, O., et al. (2023). To reverse engineer an entire nervous system. *arXiv preprint arXiv:2308.06578*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, (2023).

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research logistics Quarterly*, 2.

Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al. (2024). The multimodality cell segmentation challenge: toward universal solutions. *Nature methods*, pages 1–11.

Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez-de Mariscal, E., Nečasová, T., Guerrero Peña, F. A., Ren, T. I., Meyerowitz, E. M., Scherr, T., Löffler, K., et al. (2023). The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20 (7):1010–1020.

Myronenko, A. and Song, X. (2010). Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275. doi: 10.1109/tpami.2010.46.

Nejatbakhsh, A., Varol, E., Yemini, E., Hobert, O., and Paninski, L. Probabilistic joint segmentation and labeling of c. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 130–140. Springer, (2020).

Nejatbakhsh, A., Varol, E., Yemini, E., Venkatachalam, V., Lin, A., Samuel, A. D., and Paninski, L. (2020). Extracting neural signals from semi-immobilized animals with deformable non-negative matrix factorization. *bioRxiv*, pages 2020–07.

Nguyen, J. P., Shipley, F. B., Linder, A. N., Plummer, G. S., Liu, M., Setru, S. U., Shaevitz, J. W., and Leifer, A. M. (2016). Whole-brain calcium imaging with cellular resolution in freely behaving caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1074–E1081.

Ruszczycki, B., Pels, K. K., Walczak, A., Zamłyńska, K., Such, M., Szczepankiewicz, A. A., Hall, M. H., Magalska, A., Magnowska, M., Wolny, A., et al. (2019). Three-dimensional segmentation and reconstruction of neuronal nuclei in confocal microscopic images. *Frontiers in Neuroanatomy*, 13:81.

Ryu, J., Puche, A. V., Shin, J., Park, S., Brattoli, B., Lee, J., Jung, W., Cho, S. I., Paeng, K., Ock, C.-Y., et al. Ocelot: Overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23902–23912, (2023).

Rübel, O., Tritt, A., Ly, R., Dichter, B. K., Ghosh, S., Niu, L., Baker, P., Soltesz, I., Ng, L., Svoboda, K., Frank, L., and Bouchard, K. E. (2022). The neurodata without borders ecosystem for neurophysiological data science. *eLife*, 11:e78362. doi: 10.7554/eLife.78362.

Schrödel, T., Prevedel, R., Aumayr, K., Zimmer, M., and Vaziri, A. (2013). Brain-wide 3d imaging of neuronal activity in caenorhabditis elegans with sculpted light. *Nature methods*, 10(10):1013–1020.

Sprague, D. Y., Rusch, K., Dunn, R. L., Borchardt, J., Bubnis, G., Chiu, G., Wen, C., Suzuki, R., Chaudhary, S., Dichter, B., et al. (2024). Unifying community-wide whole-brain imaging datasets enables robust automated neuron identification and reveals determinants of neuron positioning in c. elegans. *bioRxiv*, pages 2024–04.

Strange, K. (2006). An overview of c. elegans biology. *C. elegans: Methods and Applications*, pages 1–11.

Stringer, C. and Pachitariu, M. (2024). Cellpose3: one-click image restoration for improved cellular segmentation. *bioRxiv*. doi: 10.1101/2024.02.10.579780.

Tokuoka, Y., Yamada, T. G., Mashiko, D., Ikeda, Z., Hiroi, N. F., Kobayashi, T. J., Yamagata, K., and Funahashi, A. (2020). 3d convolutional neural networks-based segmentation to acquire quantitative criteria of the nucleus during mouse embryogenesis. *NPJ systems biology and applications*, 6(1):32.

Toyoshima, Y., Tokunaga, T., Hirose, O., Kanamori, M., Teramoto, T., Jang, M. S., Kuge, S., Ishihara, T., Yoshida, R., and Iino, Y. (2016). Accurate automatic detection of densely distributed cell nuclei in 3d space. *PLoS computational biology*, 12(6): e1004970.

Toyoshima, Y., Wu, S., Kanamori, M., Sato, H., Jang, M. S., Oe, S., Murakami, Y., Teramoto, T., Park, C., Iwasaki, Y., et al. (2020). Neuron id dataset facilitates neuronal annotation for whole-brain activity imaging of c. elegans. *BMC biology*, 18:1–20.

Ulman, V., Maška, M., Magnusson, K. E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al. (2017). An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141–1152.

Varol, E., Nejatbakhsh, A., Sun, R., Mena, G., Yemini, E., Hobert, O., and Paninski, L. (2020). Statistical atlas of c. elegans neurons. *Medical Image Computing and Computer Assisted Intervention*, 12265:119–1295. doi: 10.1007/978-3-030-59722-1_12.

Venkatachalam, V., Ji, N., Wang, X., Clark, C., Mitchell, J. K., Klein, M., Tabone, C. J., Florman, J., Ji, H., Greenwood, J., et al. (2016). Pan-neuronal imaging in roaming caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8): E1082–E1088.

Weigert, M., Schmidt, U., Haase, R., Sugawara, K., and Myers, G. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3666–3673, (2020).

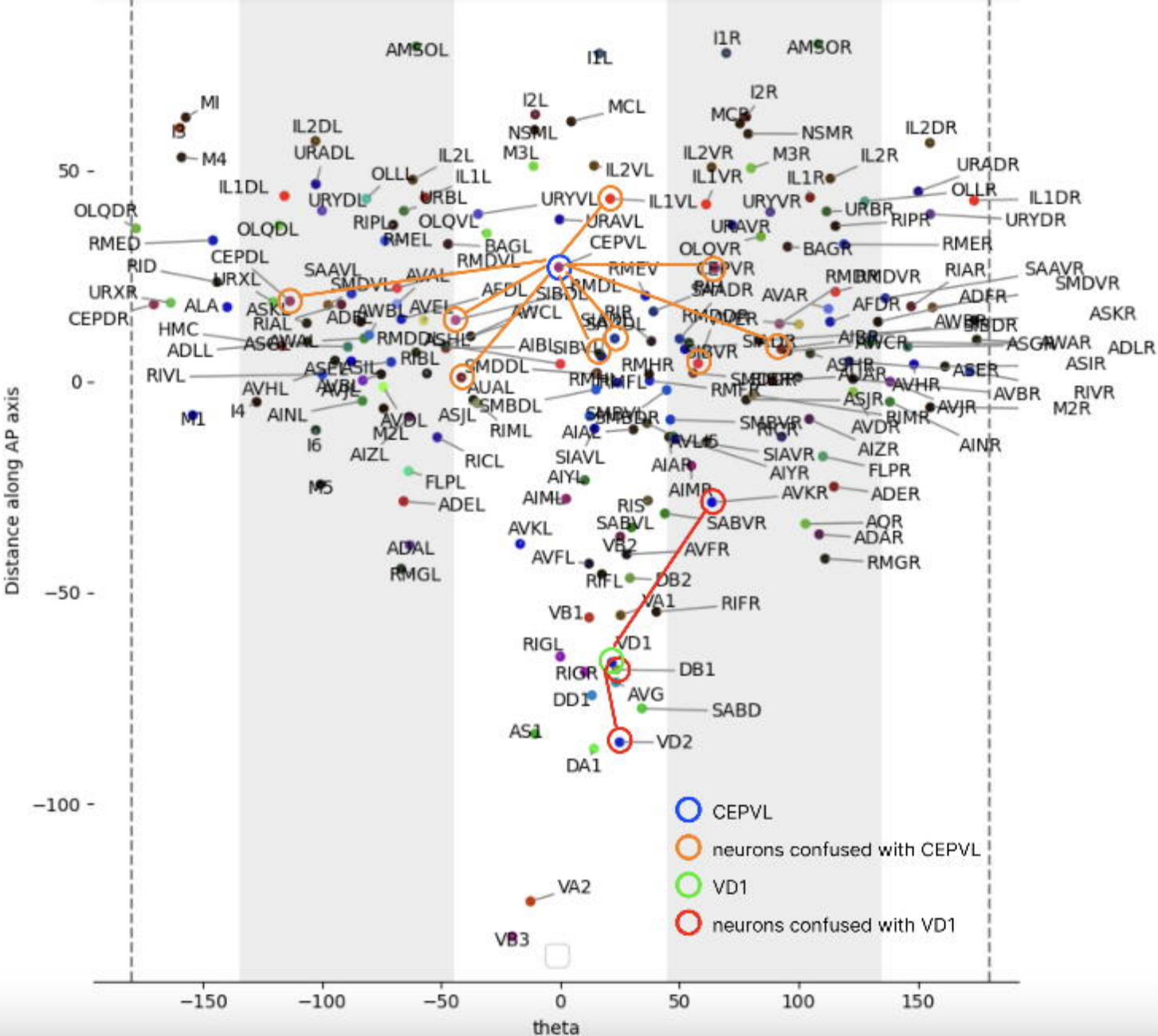
Wen, C., Miura, T., Voleti, V., Yamaguchi, K., Tsutsumi, M., Yamamoto, K., Otomo, K., Fujie, Y., Teramoto, T., Ishihara, T., et al. (2021). 3deecelltracker, a deep learning-based pipeline for segmenting and tracking cells in 3d time lapse images. *Elife*, 10: e59187.

Wu, Y., Wu, S., Wang, X., Lang, C., Zhang, Q., Wen, Q., and Xu, T. (2022). Rapid detection and recognition of whole brain activity

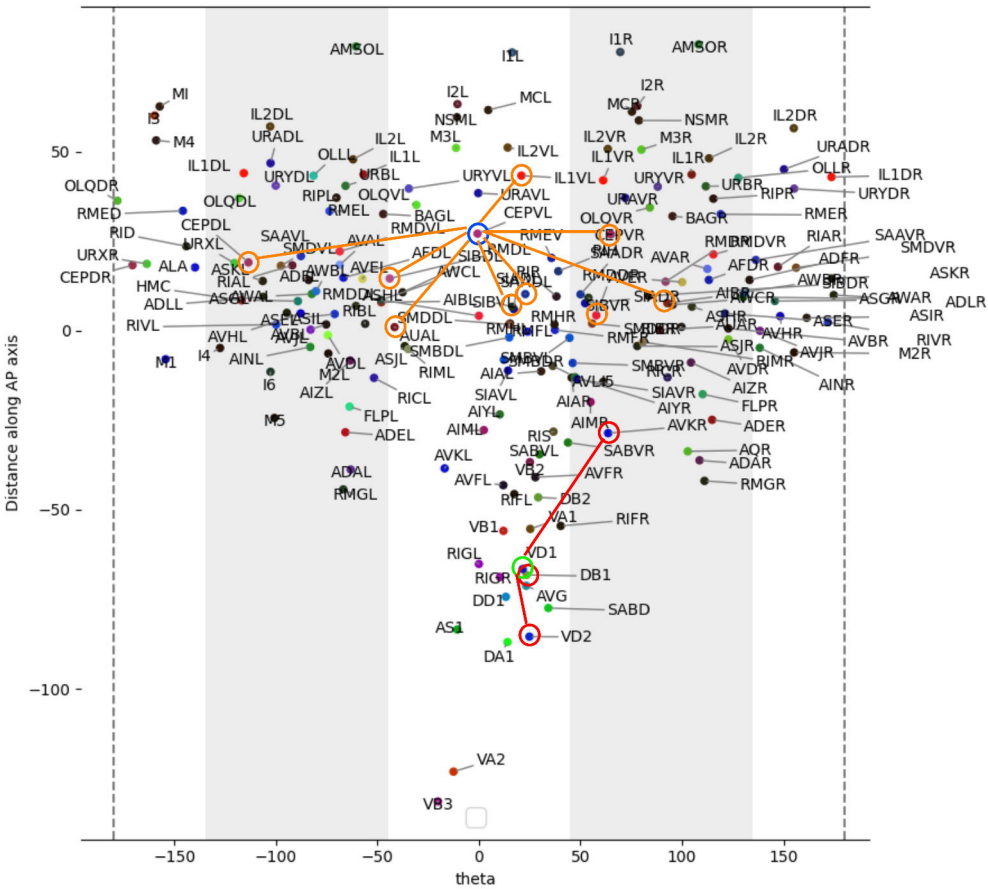
in a freely behaving caenorhabditis elegans. *PLoS computational biology*, 18(10):e1010594.

Yemini, E., Lin, A., Nejatbakhsh, A., Varol, E., Sun, R., Mena, G. E., Samuel, A. D., Paninski, L., Venkatachalam, V., and Hobert, O. (2021). Neuropal: a multicolor atlas for whole-brain neuronal identification in c. elegans. *Cell*, 184(1):272–288.

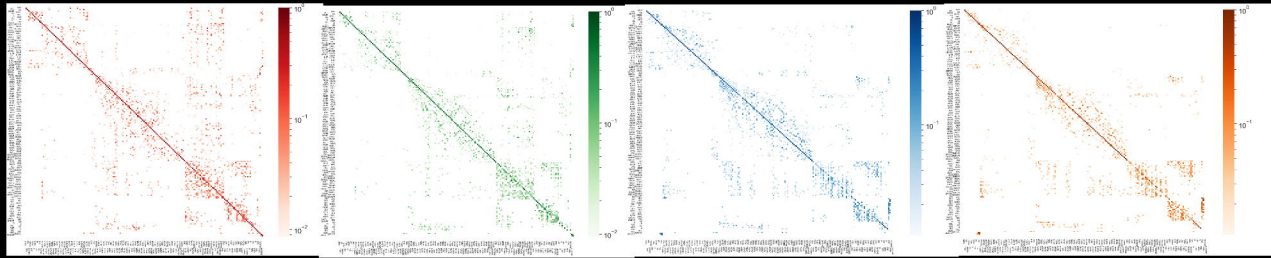
Yu, X., Creamer, M. S., Randi, F., Sharma, A. K., Linderman, S. W., and Leifer, A. M. (2021). Fast deep neural correspondence for tracking and identifying neurons in *C. elegans* using semi-synthetic training. *eLife*, 10:e66410. doi: 10.7554/eLife.66410.



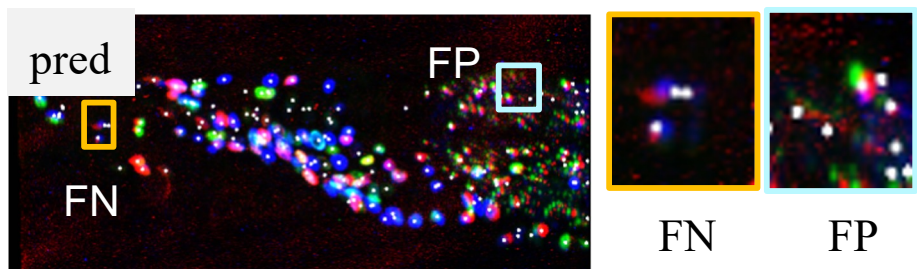
Distance along AP axis
theta
50
0
−50
−100
−150
−100
−50
0
50
100
150
CEPVL
neurons confused with CEPVL
VD1
neurons confused with VD1
AMSOL
I1L
I1R
AMSOR
MI
I3
M4
I2L
MCL
I2R
MCR
NSML
NSMR
IL2DL
URADL
IL2L
M3L
IL2VR
M3R
IL2R
IL2DR
URADR
IL1DL
OLLL
IL1L
IL2VL
IL1VR
IL1R
OLLR
IL1DR
URYDL
URBL
URYVL
IL1VL
URYVR
URBR
RIPR
URYDR
OLQDR
OLQDL
RIPL
OLQVL
URAVL
URAVR
RMED
RMEL
BAGL
CEPVL
OLQVR
BAGR
RMER
CEPDL
RMDVL
RMEV
CEPVR
RID
URXL
SAAVL
SMDVL
AVAL
RMDL
RMDVR
RIAR
SAAVR
SMDVR
URXR
ALA
ASKL
AFDL
SIBDL
AVAR
AFDR
ADFR
ASKR
CEPDR
RIAL
AWBL
AVEL
AWCL
HMC
ADLL
AIBL
SIBVL
SIBVR
AWAR
ADLR
RIVL
RIBL
SMDDL
RMHR
ASIR
AVHL
AUAL
AVHR
AVBR
RIVR
M1
I4
AINL
SMBDL
ASJR
AVJR
M2R
AVDL
ASJL
RIML
AVDR
I6
M2L
AIAL
AIZR
AINR
AIZL
RICL
SIAVL
SIAVR
AIYR
FLPR
M5
FLPL
AIYL
AIAR
AVKR
ADER
ADEL
AIML
AIMR
RIS
SABVL
SABVR
AQR
AVKL
ADAR
ADAL
VB2
AVFR
AVFL
RMGR
RMGL
RIFL
DB2
VA1
RIFR
VB1
RIGL
VD1
RIGR
DB1
DD1
AVG
SABD
AS1
VD2
DA1
VA2
VB3



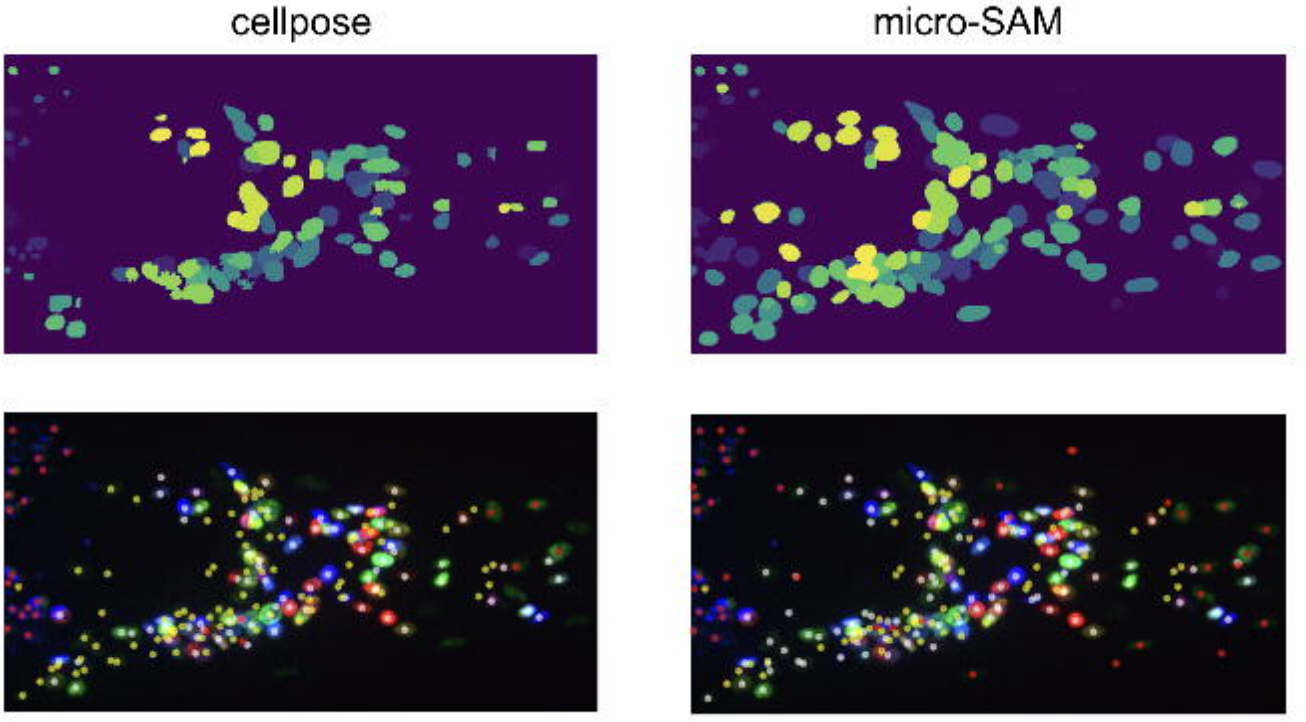
Distance along AP axis
theta
50
0
−50
−100
−150
−100
−50
0
50
100
150
AMSOL
I1L
I1R
AMSOR
I2L
MCL
I2R
MCR
NSML
NSMR
MI
I3
M4
IL2DL
URADL
IL2L
M3L
IL2VL
IL2VR
M3R
IL2R
IL2DR
URADR
OLLL
IL1L
IL1VR
IL1R
OLLR
IL1DL
URYDL
URBL
URYVL
IL1VL
URYVR
URBR
IL1DR
OLQDR
OLQDL
RIPL
OLQVL
URAVL
URAVR
RIPR
URYDR
RMED
RMEL
BAGL
CEPVL
OLQVR
BAGR
RMER
CEPDL
RMDVL
RMEV
CEPVR
RID
SAAVL
AVAL
RMDVR
RIAR
SAAVR
URXL
SMDVL
AFDL
SIBDL
RMDL
AVAR
ADFR
SMDVR
URXR
ALA
ASKL
AWBL
AVEL
AWCL
AFDR
ASKR
CEPDR
RIAL
ADEL
RIR
AWBR
SIBDR
HMC
AWAL
RMDDL
AIBL
SIBVL
AWCR
ASGR
AWAR
ADLR
ADLL
ASGL
RIBL
SIBVR
ASIR
RIVL
ASEL
ASIL
SMDDL
RMHR
ASHR
ASER
AVBL
AVJL
AUAL
RMHL
RMFL
AVHR
AVBR
RIVR
AVHL
AVJR
M2R
M1
I4
AINL
AVDL
ASJL
SMBDL
RMFR
ASJR
AVJR
RIML
AIAL
SMBVL
AVDR
RIMR
M2L
I6
AIZL
RICL
AVL
SIAVL
AIZR
AINR
SIAVR
AIAR
AIYR
FLPR
M5
FLPL
AIYL
AIML
AIMR
AVKR
ADER
ADEL
RIS
SABVL
SABVR
AQR
AVKL
ADAR
ADAL
VB2
AVFR
AVFL
RMGR
RMGL
RIFL
DB2
VA1
RIFR
VB1
RIGL
VD1
RIGR
DB1
DD1
AVG
SABD
AS1
VD2
DA1
VA2
VB3

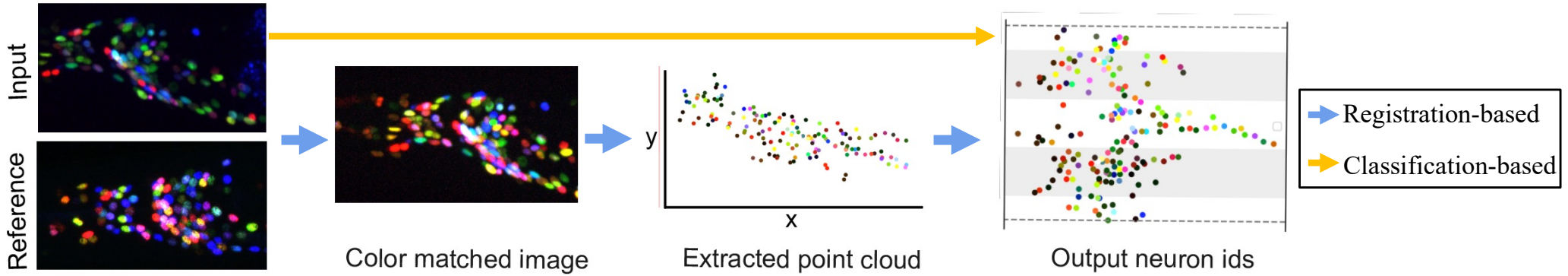


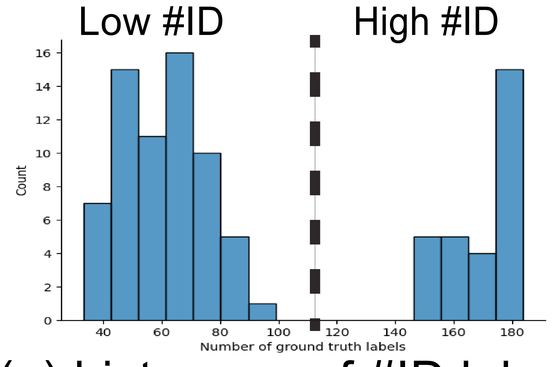
(a) Example 1 (F1=0.72)



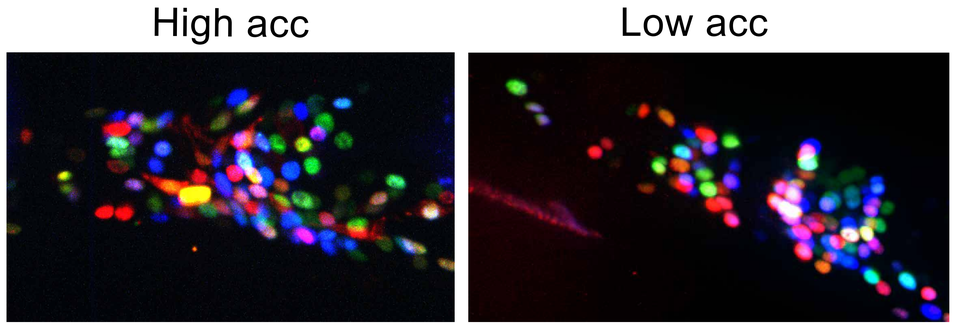

(b) Example 2 (F1=0.52)


cellpose
micro-SAM

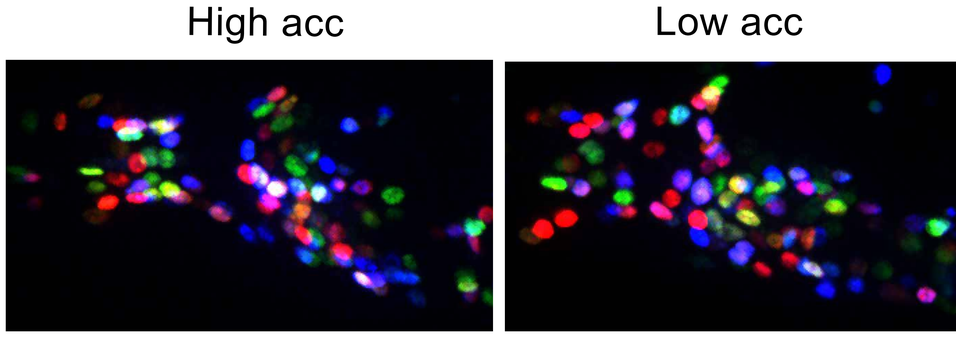

Input
Reference
Color matched image
y
x
Extracted point cloud
Output neuron ids
Registration-based
Classification-based



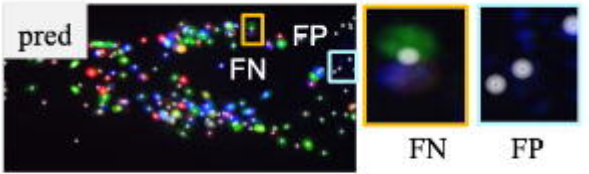

(a) histogram of #ID labels



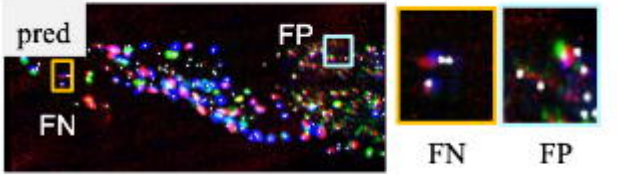

(b) low #ID labels



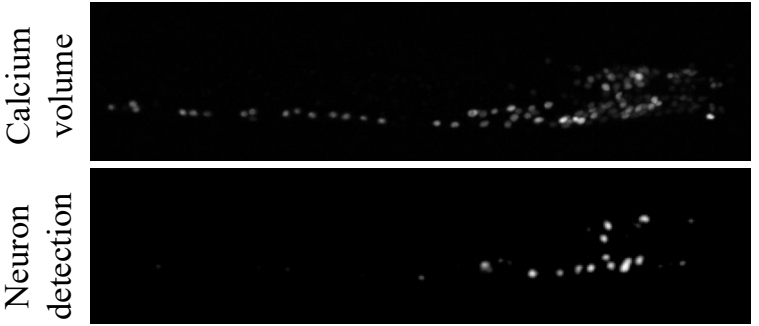

(c) high #ID labels



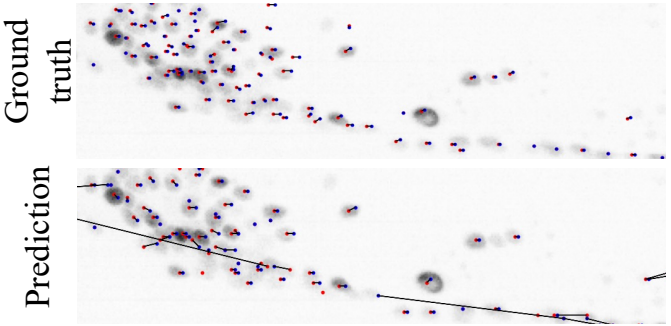

(a) Example (F1=0.72)



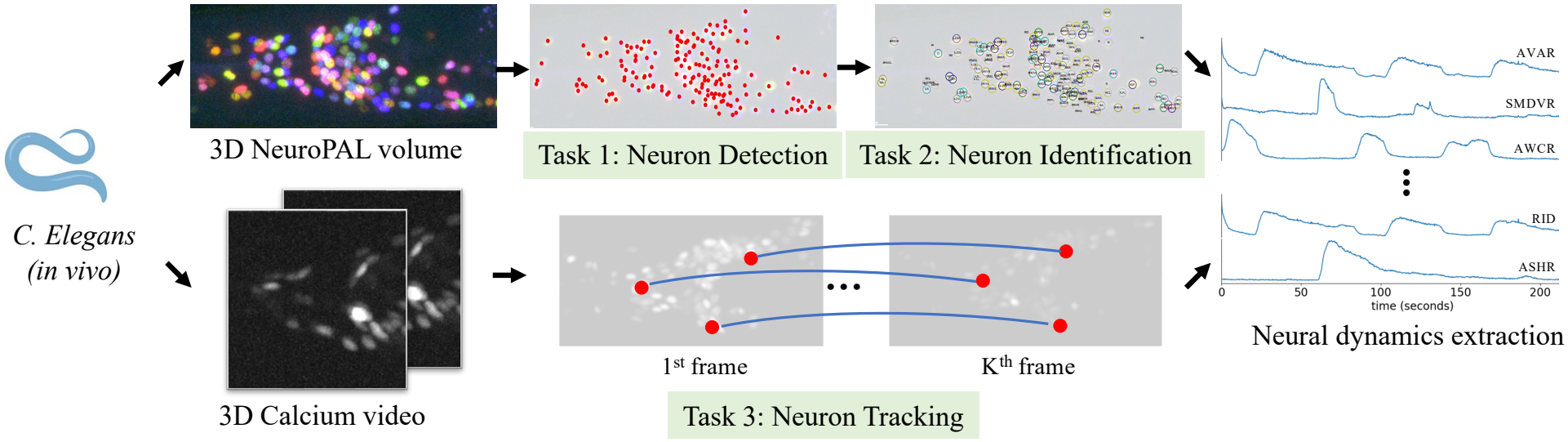

(b) Example 2 (F1=0.52)

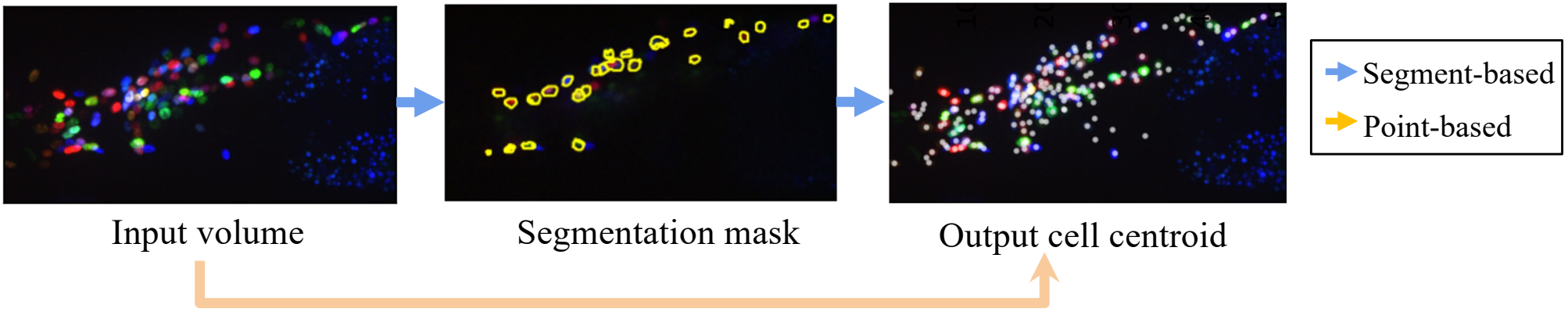


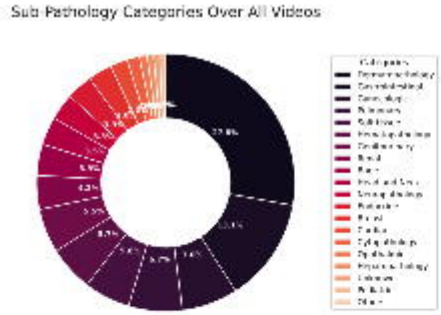

(a) Missing detections



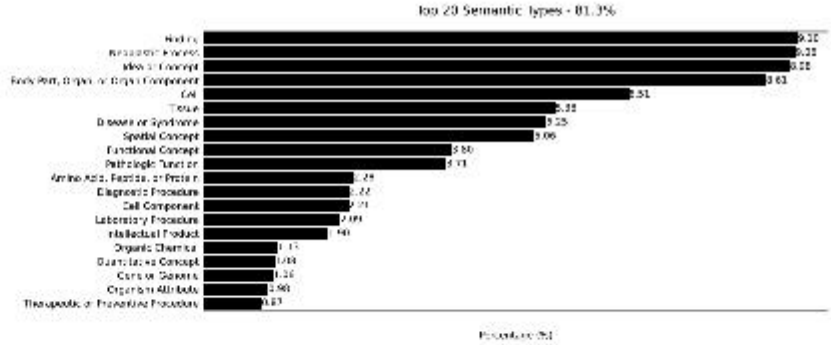

(b) Linking errors


C. Elegans
(in vivo)
3D NeuroPAL volume
Task 1: Neuron Detection
Task 2: Neuron Identification
AVAR
SMDVR
AWCR
RID
ASHR
0
50
100
150
200
time (seconds)
Neural dynamics extraction
3D Calcium video
1st frame
Kth frame
Task 3: Neuron Tracking

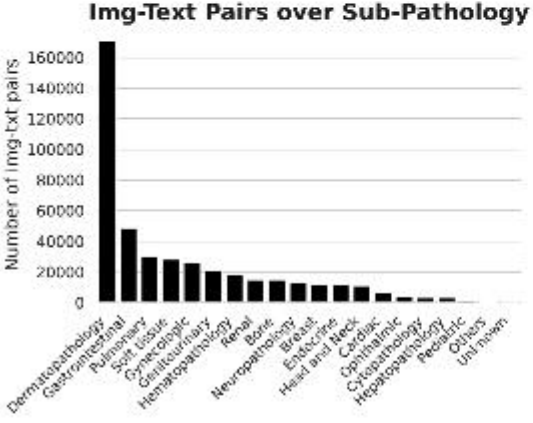

Segment-based
Point-based
Input volume
Segmentation mask
Output cell centroid



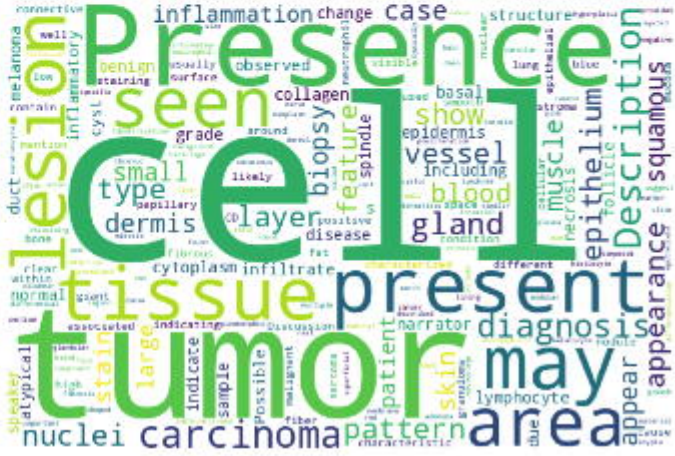

(a)

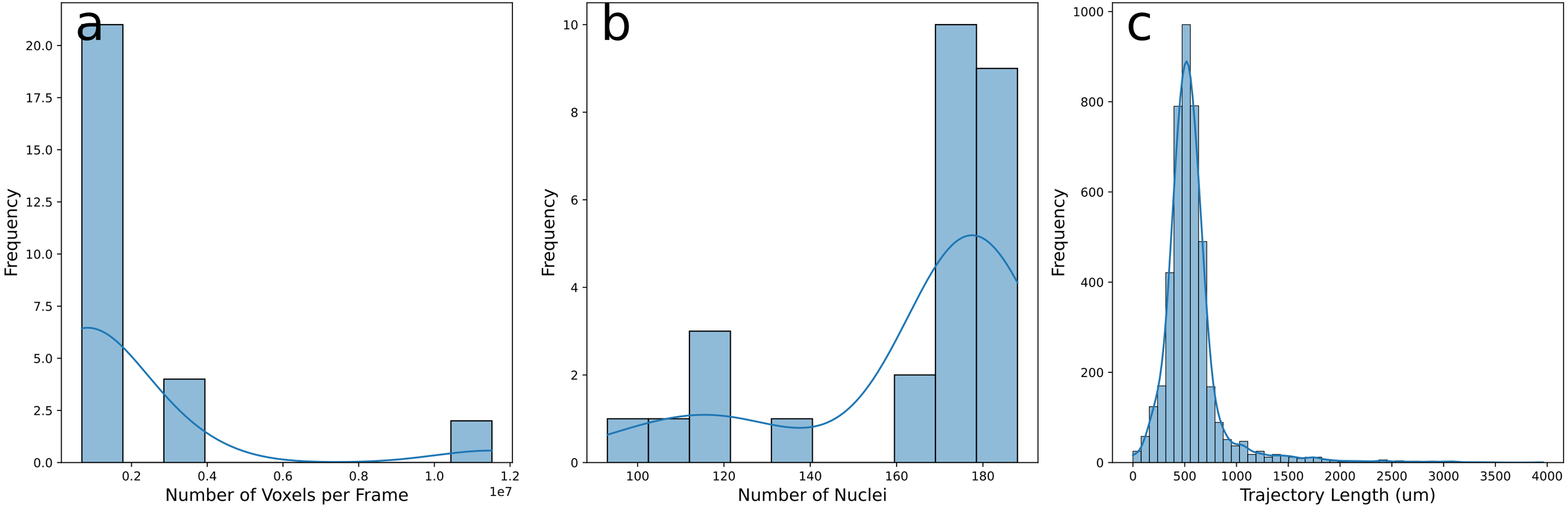



(b)




(c)


cell
tumor
Presence
tissue
present
area
may
seen
lesion
Description
carcinoma
nuclei
pattern
diagnosis
gland
vessel
epithelium
squamous
appearance
inflammation
case
change
structure
observed
collagen
basal
show
biopsy
feature
small
type
dermis
layer
blood
muscle
skin
patient
indicate
sample
possible
large
stain
atypical
appear
lymphocyte
narrator
infiltrate
cytoplasm
disease
epidermis
grade
including
spindle
melanoma
inflammatory
cyst
duct
benign
usually
surface
staining
low
well
connective
lung
blue
nuclear
epithelial
stroma
follicle
necrosis
different
condition
space
positive
CD
fibrous
bone
fat
clear
within
normal
giant
associated
indicating
discussion
malignant
sarcoma
fiber
characteristic
due
speckle
high
likely
around
contain
mention
papillary
Low
surface
visible
neutrophil
nuclei
cell
lymphocyte
appear
appearance
squamous
epithelium
muscle
blood
gland
diagnosis
may
area
pattern
carcinoma
nuclei
tissue
tumor
present
Presence
Description
lesion
seen
inflammation
case
structure
observed
collagen
basal
show
biopsy
feature
small
type
dermis
layer
vessel
including
epidermis
grade
spindle
melanoma
inflammatory
cyst
duct
benign
usually
surface
staining
low
well
connective
lung
blue
nuclear
epithelial
stroma
follicle
necrosis
different
condition
space
positive
CD
fibrous
bone
fat
clear
within
normal
giant
associated
indicating
discussion
malignant
sarcoma
fiber
characteristic
due
speckle
high
likely
around
contain
mention
papillary
Low
surface
visible
neutrophil
disease
infiltrate
cytoplasm
narrator
patient
skin


a
Frequency
Number of Voxels per Frame
1e7
b
Frequency
Number of Nuclei
c
Frequency
Trajectory Length (um)