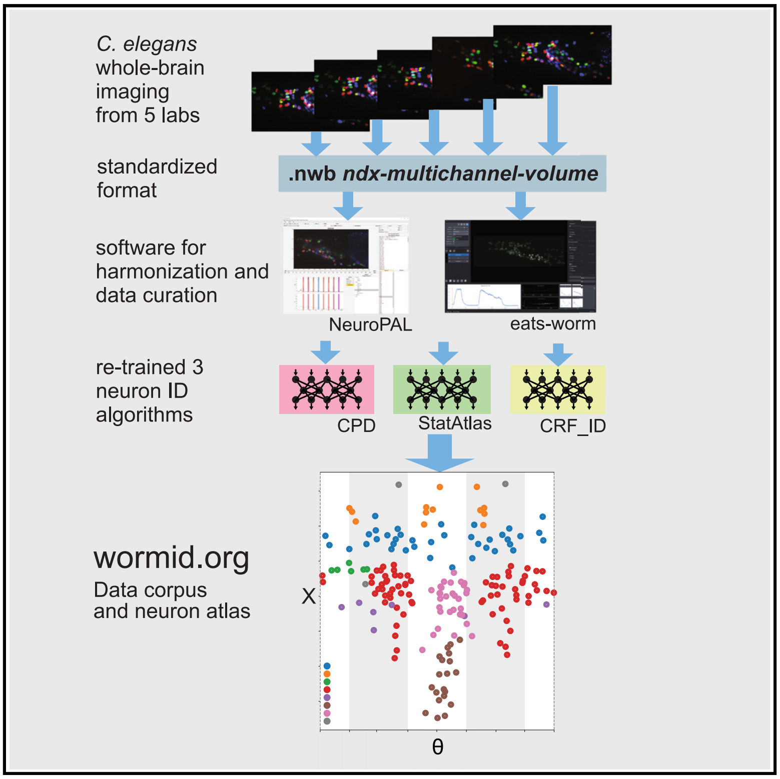
Cell Reports
Methods

Resource

Unifying community whole-brain imaging datasets enables robust neuron identification and reveals determinants of neuron position in *C. elegans*

Graphical abstract


C. elegans whole-brain imaging from 5 labs
standardized format
.nwb *ndx-multichannel-volume*
software for harmonization and data curation
NeuroPAL
eats-worm
re-trained 3 neuron ID algorithms
CPD
StatAtlas
CRF_ID
wormid.org
Data corpus and neuron atlas
X
θ

Authors

Daniel Y. Sprague, Kevin Rusch, Raymond L. Dunn, ..., Koutarou D. Kimura, Eviatar Yemini, Saul Kato

Correspondence

eviatar.yemini@umassmed.edu (E.Y.),
saul.kato@ucsf.edu (S.K.)

In brief

Training machine-learning algorithms on diverse data improves performance. Sprague et al. develop a standardized format and harmonization tools to unify a corpus of *C. elegans* whole-brain imaging data from five labs. Neuron ID algorithms show improved performance and generalization, obviating individual lab retraining, and the harmonized atlas yields neurobiological insights.

Highlights

- NWB file format extended to support *C. elegans* whole-brain imaging data
- Production of a diverse data corpus using harmonization tools at wormid.org
- Neuron ID algorithms improve in performance and do not need lab-specific retraining
- Neuronal atlas reveals biological determinants of neuron positioning

Sprague et al., 2025, Cell Reports Methods *5*, 100964
January 27, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.crmeth.2024.100964

Cell Reports Methods

CellPress
OPEN ACCESS

Resource

Unifying community whole-brain imaging datasets enables robust neuron identification and reveals determinants of neuron position in *C. elegans*

Daniel Y. Sprague,[1] Kevin Rusch,[2] Raymond L. Dunn,[1] Jackson M. Borchardt,[1] Steven Ban,[1] Greg Bubnis,[1] Grace C. Chiu,[1] Chentao Wen,[3] Ryoga Suzuki,[4] Shivesh Chaudhary,[5] Hyun Jee Lee,[5] Zikai Yu,[5] Benjamin Dichter,[6] Ryan Ly,[7] Shuichi Onami,[3] Hang Lu,[5] Koutarou D. Kimura,[4] Eviatar Yemini,[2,*] and Saul Kato[1,8,*]

[1]Department of Neurology, University of California San Francisco, San Francisco, CA 94158, USA
[2]Department of Neurobiology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA
[3]RIKEN Center for Biosystems Dynamics Research, Kobe, Hyogo 650-0047, Japan
[4]Graduate School of Science, Nagoya City University, Nagoya, Aichi 467-8501, Japan
[5]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[6]CatalystNeuro, LLC, Casper, WY 82601, USA
[7]Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[8]Lead contact
*Correspondence: eviatar.yemini@umassmed.edu (E.Y.), saul.kato@ucsf.edu (S.K.)
https://doi.org/10.1016/j.crmeth.2024.100964

MOTIVATION The *C. elegans* community continually produces whole-brain imaging datasets that could in theory be reused for new scientific inquiry. Nevertheless, because the data are produced using disparate equipment and techniques and stored in ad hoc formats, it is challenging to assimilate and reuse this growing collection. We developed a standard format and data harmonization methods to assimilate over 100 animals from five labs into a single data corpus. Using this corpus, we trained algorithms to produce a robust, lab-agnostic neural identification atlas system that does not require re-training for new labs and discovered biological factors that influence neural positioning.

SUMMARY

We develop a data harmonization approach for *C. elegans* volumetric microscopy data, consisting of a standardized format, pre-processing techniques, and human-in-the-loop machine-learning-based analysis tools. Using this approach, we unify a diverse collection of 118 whole-brain neural activity imaging datasets from five labs, storing these and accompanying tools in an online repository WormID (wormid.org). With this repository, we train three existing automated cell-identification algorithms, CPD, StatAtlas, and CRF_ID, to enable accuracy that generalizes across labs, recovering all human-labeled neurons in some cases. We mine this repository to identify factors that influence the developmental positioning of neurons. This growing resource of data, code, apps, and tutorials enables users to (1) study neuroanatomical organization and neural activity across diverse experimental paradigms, (2) develop and benchmark algorithms for automated neuron detection, segmentation, cell identification, tracking, and activity extraction, and (3) share data with the community and comply with data-sharing policies.

INTRODUCTION

Whole-brain imaging experiments with single-neuron resolution (herein shortened to simply "whole-brain imaging") have undergone explosive growth since first demonstrated in the millimeter-long nematode worm *Caenorhabditis elegans* and the zebrafish *Danio rerio* in 2013.[1,2] Since then, these methods have been widely adopted and advanced in the worm,[3–7] zebrafish,[8–13] and larval[14] and adult[15,16] fly communities. Moreover, there have been significant efforts and advances in neuron-resolution imaging of multiple and/or large brain regions in mammals, rapidly approaching whole-brain imaging, especially in mice.[17–21]

In *C. elegans*, whole-brain imaging datasets have enabled characterization of neural network dynamics,[3,6] functional connectivity,[22–24,25] and the roles of individual neurons during behavior.[7] These studies leverage the property of eutely in this organism: each cell has a unique and stereotyped identity, consistent across every animal, that allows for data from individual neurons to be pooled and compared across multiple trials and animals. However, analyses of these experiments are bottlenecked by the need to determine the unique identities of each neuron in 3D volumetric recordings. Manual cell identification from fluorescent microscope images is a notoriously difficult skill, requiring substantial expertise and labor. This task is particularly difficult for neurons labeled with nuclear localized fluorophores, which is typical for whole-brain recordings. We recently developed NeuroPAL,[6] the first method where the unique identity of every single neuron can be distinguished by an invariant fluorescent color barcode in living animals at all developmental stages of both sexes.[26] NeuroPAL has greatly simplified the task of cell identification and has thus seen rapid adoption, with at least six labs[6,7,22,24,25] publishing whole-brain imaging datasets using these animals and many more labs incorporating the system into their experimental protocols since its release in 2021.

Despite this innovation, neural identification remains a challenging task that requires expertise and many hours of manual work. In the past few years, researchers have proposed various algorithmic auto-identification approaches to attack this problem.[25,27–31] However, none of them have achieved widespread adoption, due at least in part to their incompatibility with different microscopy data formats and low performance on data acquired from different labs. Automatic approaches to the complementary problem of tracking neurons across video frames have achieved some generalized performance across various datasets,[32,33] but so far, there have not been efforts to perform similar training and benchmarking for automatic cell identification. To build automatic approaches that are robust, accurate, and generalizable, there is a critical need for a standardized format and compatible tools trained and benchmarked on a consolidated corpus of data that reflects the heterogeneity of microscopy equipment, experimental conditions, and protocols across labs.

To address this need, we take a data harmonization approach: a process of combining datasets from different sources and homogenizing them to produce a substantially larger data corpus that, in our case, minimizes non-biological inconsistencies across individual datasets while increasing the overall biological diversity of training and benchmarking data. Harmonization includes (1) aggregating the data, (2) converting it to a standardized format, (3) normalizing it, (4) handling duplicate and missing data, and (5) pre-processing data to register it to a common space and coordinate system. Data harmonization is standard in many data science fields but has seen slower adoption in the life sciences.[34] Similar efforts to standardize data formats and build large corpuses of data have been essential in the development and benchmarking of many modern machine-learning algorithms.[35–37]

We introduce WormID (wormid.org). This resource consists of (1) data harmonization tools including a standardized file format for both raw and processed data alongside related metadata that extends the existing Neurodata Without Borders (NWB) format, (2) pre-processing to align the color and coordinate space of new datasets, and (3) open-source software to analyze whole-brain activity images. We also provide tutorials and documentation that enable researchers to easily incorporate these tools into their data pipelines. Finally, we provide a large online corpus of harmonized *C. elegans* whole-brain activity imaging and structural data that can be used for large-scale experimental analysis, neurobiological modeling, and algorithmic development. This corpus is stored in a popular community archive called the Distributed Archives for Neurophysiology Data Integration (DANDI), which serves as a free, persistent, open access repository for experimental neuroscience data from a variety of model organisms.[38] The addition of new datasets to this corpus are encouraged, facilitated by tutorials and software, and provide a simple means for submitters to comply with data-sharing policies of federal and private funders.

By aggregating a diversity of datasets from multiple labs into a large data corpus, we achieve a substantial boost in the performance of three existing neural auto-identification algorithms, arguably moving into the regime of practical utility for the broader community of users. Furthermore, we mine this corpus to investigate the relationship between neural lineage, synaptic connectivity, and somatic positioning of *C. elegans* neurons to better understand the factors that drive the positioning of neurons in the adult worm.

This corpus and set of tools should be of wide utility to *C. elegans* researchers. We hope it will serve as a seed for continued community aggregation of brain imaging datasets and further the development and improvement of community data-analysis tools applicable across many model organisms.

RESULTS

A standardized format for whole-brain *C. elegans* recordings enables data aggregation and algorithm interoperability

Current state-of-the-art whole-brain recordings of *C. elegans* typically consist of a combination of structural images that often use the NeuroPAL multi-channel fluorescent system to determine neuron identities (Figures 1A and 1B) and time-series images of neural activity acquired by using genetically encoded activity sensors (e.g., GCaMP6s[39]) (Figure 1C). This imaging is performed either on immobilized worms (often constrained within a microfluidic chip to maximize image quality[1,3,40]) or on freely moving worms.[4,5] To aid interpretation, herein we visualize whole-brain structural NeuroPAL images via (1) an unrolled "butterfly" plot of neuron positions that projects the 3D worm structure into a 2D plane (Figure 1A), (2) a 2D projection plot of the NeuroPAL color space (Figure 1B), and (3) 2D dorsal-ventral and lateral projection plots of the neurons (Figure 1B). These visualizations facilitate quick comparisons of neuron color and position from different samples and fine-tuning of their global alignment.

All associated raw data and metadata is stored in the standardized NWB[41] file format with an additional extension that we developed, *ndx-multichannel-volume* (ndx = neurodata extension), to provide support for multi-channel volumetric recordings and *C. elegans*-specific metadata (Figure 2A). This extension is available in the NWB Extensions Catalog and is now the official NWB standard for data sharing of *C. elegans*

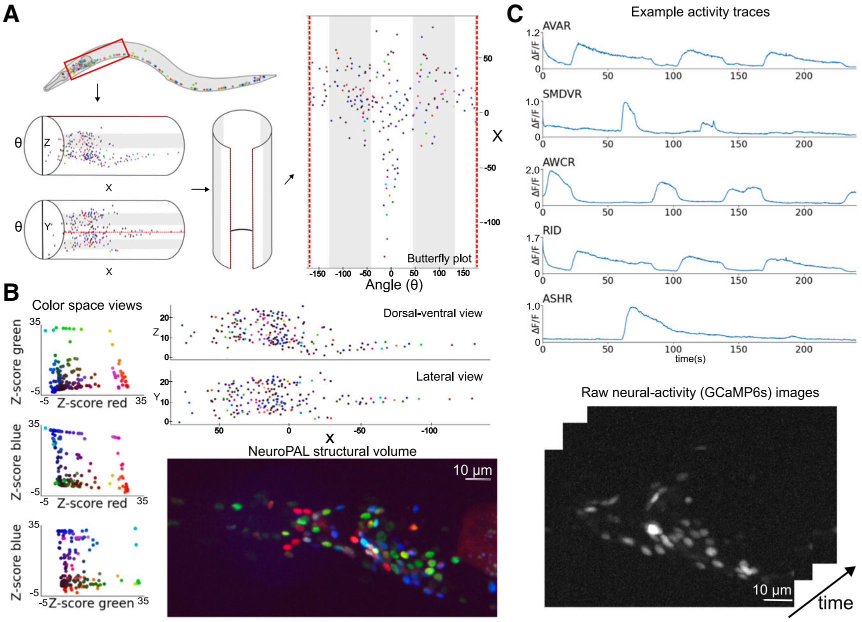


Figure 1. Example NWB file contents

(A) Illustration of a NeuroPAL worm. The head is highlighted by a red box. A butterfly plot visualizes the full 2D representation of the worm brain by projecting neurons onto the surface of the cylindrical body and then unrolling the cylinder. Neuron centers are colored using their composite NeuroPAL expression.

(B) Visualizations of the raw NeuroPAL structural image and 2D projections of its RG, RB, and GB color subspaces and XZ and XY projections of its neuron positions. Neuron centers are colored using their composite NeuroPAL expression.

(C) Example activity traces for five neurons contained in the NWB file and of the raw neural-activity (GCaMP6s) images. See also Figure S2.

whole-brain neural-activity imaging. NWB data are hierarchically organized with basic metadata stored at the file's root level, raw data stored in the "acquisition module," and various processed experimental data stored in "processing modules" (Figure 2B). Individual NWB files contain a single experimental run for a single animal. These NWB files are then stored and accessed from the DANDI archive, where they receive a unique persistent digital object identifier (DOI) in accordance with the International Organization for Standardization (ISO). Datasets can then be downloaded or streamed from the DANDI archive using the DANDI interface or API (Figure 2C).

We incorporated NWB *ndx-multichannel-volume* read and write functionality into two software tools. These independent software implementations both offer both user-friendly GUIs that allow the visualization, analysis, and curation of *C. elegans* NeuroPAL structural images and neural activity recordings in immobilized worms (NeuroPAL software https://github.com/Yemini-Lab/NeuroPAL_ID, Figure 3A; eats-worm software https://github.com/focolab/eats-worm, Figure 3B). This functionality can be straightforwardly incorporated into other data-analysis pipelines and software.

In Table 1, we present a summary of the data we aggregated and harmonized into a corpus: 108 worms from six datasets acquired by five different labs, each with neuron positions and human-labeled neuron identities. This corpus can be mined for biological insights, training and benchmarking of machine-vision approaches, and neurobiological studies of structural and neural-activity time-series data. Each of these datasets is stored on DANDI and range from a few hundred megabytes to several terabytes (see STAR Methods for dataset references). DANDI supports streaming from the cloud and allows users to selectively load data objects and data chunks, substantially reducing the local data storage and RAM requirements necessary to work with these data on a personal computer.

An updated atlas of neuron positions in the *C. elegans* hermaphrodite head

Our multi-lab data corpus allows data scientists to train and benchmark the performance of algorithms for automated neuron identification using datasets that reflect real-world diversity. In this section, we focus on the statistical atlas approach presented in Varol et al.[27] This approach was the first to take advantage of the color information provided by NeuroPAL and was presented alongside the original NeuroPAL work.[6] This neuron identification assignment algorithm was framed as a bipartite graph matching problem, with the goal of minimizing the total assignment cost using the well-known Hungarian algorithm.[42] Cost is calculated by comparing neuron positions and colors in the animal sample with the mean and covariance of neuron positions and colors in a reference statistical atlas (see STAR Methods). The original atlas presented in the paper was trained on 10 worms from the original NeuroPAL work. We retrained this atlas using the full multi-lab corpus that we present in this work, increasing the training set by over 10-fold. To train the statistical atlas, we first performed an affine transformation to roughly register each training dataset to one of the 10 original NeuroPAL worms based on ground-truth labels. This roughly aligns the principal axes of the worm as well as the scaling to a common space. The algorithm then calculates the means and covariances of neuron positions and colors in this roughly aligned space for individual recordings. See STAR Methods for further details on atlas training. The statistical atlas generated by this approach serves the additional purpose of characterizing the mean and covariance of neuron positions and colors across the whole corpus of data. To the best of our knowledge, this newly generated atlas is the most comprehensive atlas of neuron positions and NeuroPAL coloring. We find that training on the multi-lab corpus significantly decreases the alignment cost for the ground truth labeled neurons across datasets, indicating that the true distribution of neuron positions and variances is more accurately represented in the more comprehensive atlas compared to the original one (Figure S1).

In Figures 4 and S2, we present visualizations of the statistical atlas of neuron positions and colors trained on 104 of the 118 worms in our consolidated NWB/DANDI dataset as well as on

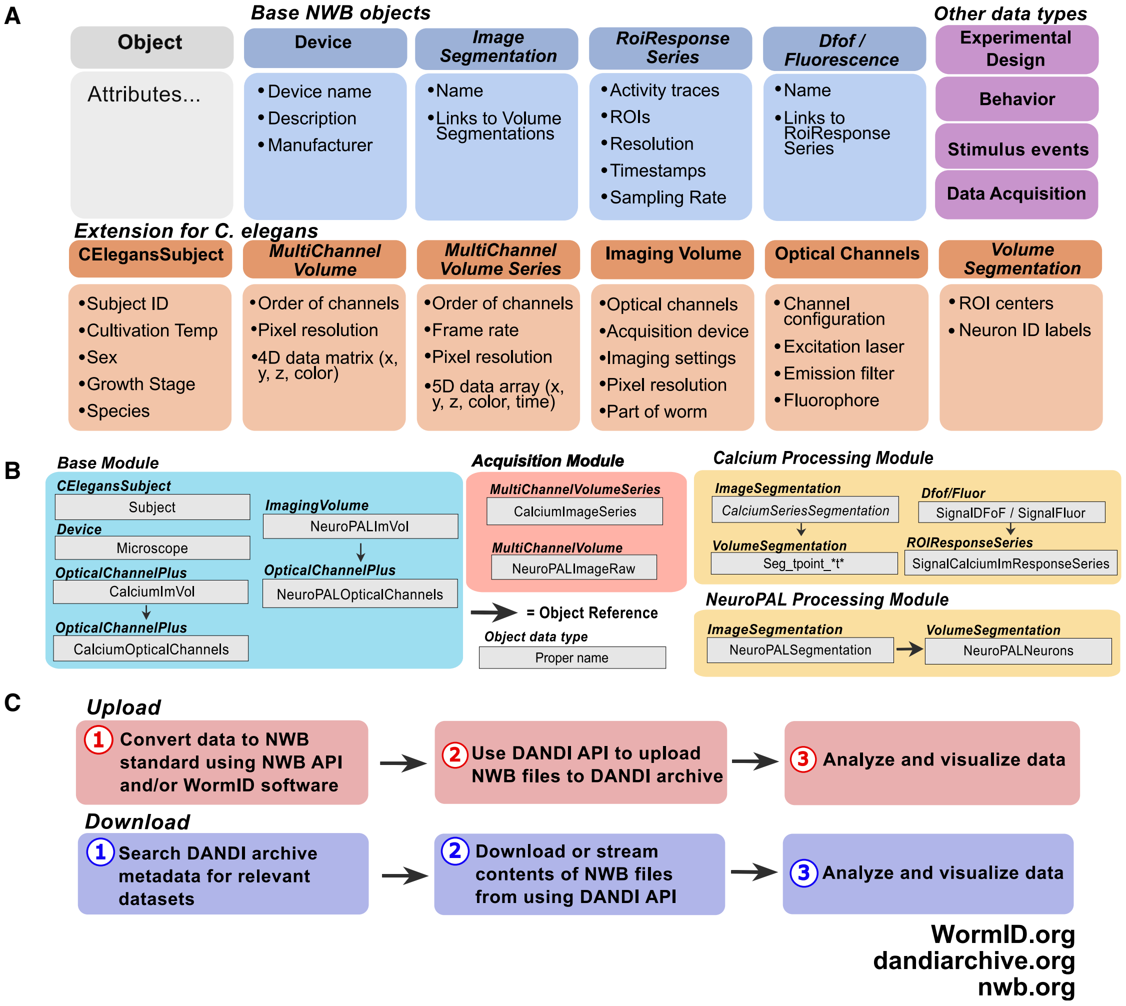


Figure 2. Schema for *C. elegans* extension to the NWB architecture

(A) Names and content for objects used in *C. elegans* optophysiology NWB files.

(B) File organization hierarchy of NWB files for *C. elegans* optophysiology. Modules are structured like folders within the root file in an HDF5-based hierarchy.

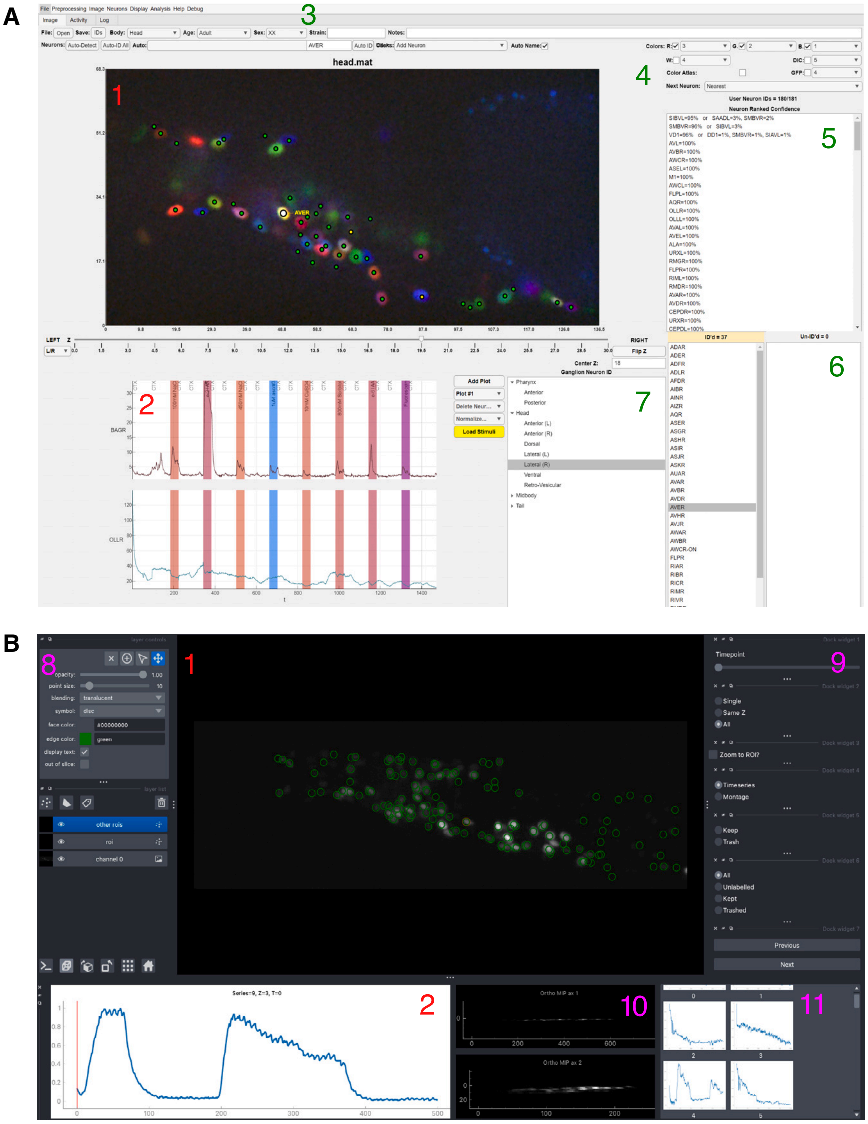
(C) Flowchart of steps for converting new data to NWB and uploading to the DANDI archive and for using NWB datasets that are already up on DANDI.

the smaller dataset of 10 worms used in the original NeuroPAL paper (employing the StatAtlas algorithm in Yemini et al.[6] and Varol et al.[27]). We found that substantial deviations from a linear pose and occluding image artifacts significantly impacted algorithmic alignment and thus accuracy, therefore, 14 worms were omitted from the atlas due to large nonlinear deformities or pronounced imaging artifacts. With 104 worms, this represents, to the best of our knowledge, the most broadly trained statistical atlas for *C. elegans* neuron positions and NeuroPAL coloring available. By leveraging the diversity of the multi-lab corpus this atlas captures variability between individual worms, strains, and lab-specific experimental conditions. This atlas can be used as a basis for training and testing automatic labeling algorithms as well as biological investigations of neuron positions and structural brain organization. This statistical atlas further complements detailed electron microscopy (EM)-based anatomical atlases by adding cellular structural detail and, notably, it provides nearly 100 more animals in its corpus than the approximately 10 EM ones available now.[43–45] Although our corpus lacks the synaptic connectivity found in the EM datasets, it provides the complementary functional activity that is not available and cannot be obtained from EM imaging of fixed animals.

WormID.org supplies links to the software, visualization tools, and datasets discussed previously in this paper. Furthermore, wormid.org provides links to the data corpus and related tools to work with whole-brain structural images and activity recordings, convert datasets to NWB, and supplies tutorials and instructions for using these tools. Our aim is that these data standard, data corpus, and atlas of cell positions will be a continually evolving resource for the *C. elegans* neuroscience community and eventually for the communities of other model organisms.

Analysis of biological factors influencing neuron positions

We statistically analyzed the spatial positions of *C. elegans* neuronal somas across individuals, strains, and lab conditions



1 Z-stacks from the 3D image volume
2 Neural activity trace (time series)
3 Experiment information
4 Color channels to display
5 Automated neural ID predictions
6 Neurons that have been identified / remain unidentified
7 Body regions and ganglia containing each neuron
8 Display options for 3D image and ROIs
9 Options to move through time points
10 XY and XZ maximum intensity projections
11 Select trace to display

Figure 3. Two software GUI applications with NWB I/O

(A and B) GUIs with NWB I/O support for visualization and annotation of NeuroPAL structural images, neural detection, and automated identification and time-series of neural activity with stimulus presentation in immobilized worms. Colored numerical callouts highlight specific modules and features in each software.

(A) NeuroPAL ID software from Yemini Lab.

(B) EATS-worm software from Kato Lab.

based on their means and covariances in the statistical atlas. We focused on relative pairwise displacements rather than absolute positions because the absolute position of cells is dependent on the pose and deformation of the animal's body during imaging. Interpreting absolute cell positions would require the complex step of computationally aligning all animals; furthermore, aligning multiple animals into identical positions is an imperfect task. In contrast, pairwise cell displacements are alignment independent and relatively robust to animal deformation.

Before analyzing statistical properties on the neuron positions, we assessed the percentage of neurons that were labeled by humans in each dataset. We found that neurons in the ventral ganglion and retrovesicular ganglion were less commonly labeled than neurons in other ganglia. As is shown here and previously in Yemini et al.,[6] neurons in the ventral and retrovesicular ganglia exhibit high relative positional variability, which may explain why fewer of them were confidently labeled by researchers (Figure 5A). For this reason, we explored several different factors hypothesized to contribute to the organization and variability of relative cell positions: (1) gangliar boundaries (e.g., basal lamina and abutting tissue), which may restrict cell movement within the coelem; (2) synaptic connectivity, which may impose energetic costs dependent on neuronal proximity; and (3) developmental-time and cell-lineage effects whereby recently divided cells (i.e., sister cells) remain close together and more distant relatives (e.g., mother and grandmother cells) end up further apart.

To test the first hypothesis, that gangliar boundaries regulate positional organization and variability, we measured the positional variability of neurons that are spatially close and compared pairs within the same ganglion to pairs straddling each other in different ganglia (see STAR Methods). We found that neurons in the anterior pharyngeal bulb and neurons in the dorsal, lateral, and retrovesicular ganglia all exhibit significantly lower variability for pairs within the same ganglion compared to pairs in different ganglia. Conversely, for neurons in the anterior and ventral ganglia, we observed no significant difference in positional variability between pairs in the same ganglion and pairs in different ganglia (Figure 5B). We used an independent-sample t test to compare pairs within the same ganglion with those in different ganglia. To determine whether disparities in the distributions of pairwise distance between intra- and inter-ganglion groups could trivially account for this difference in variability, we performed distance-matched

Table 1. Summary of aggregated dataset characteristics

Dataset #	Dandiset ID	# of worms in the dataset	Lab code	NeuroPAL, GCaMP, or both	# of neurons marked (avg.)	# of ID labels (avg.)
NP	000715	10	NP	NeuroPAL	189–196 (193)	186–193 (190)
1	000541	21	EY	Both	166–188 (177)	164–184 (175)
2	000714	9	HL	NeuroPAL	113–125 (119)	58–69 (64)
3	000692	9	KK	Both	149–163 (154)	149–163 (154)
4	000776	38	SF	Both	29–96 (70)	29–96 (70)
5	000565	21	SK1	Both	78–139 (111)	30–82 (48)
6	000472	10	SK2	NeuroPAL	166–180 (173)	38–63 (49)
Summary		118	5 labs		29–196 (126)	29–193 (99)

The NP dataset comes from the original NeuroPAL paper.[6] Datasets can be accessed programmatically using the DANDI API and Dandiset ID or by searching the Dandiset ID on dandiarchive.org.

comparisons of the variability of inter-ganglia pairs to intra-ganglia pairs in 1 μm bins. We observed a consistently higher variability of inter-ganglia pairs to intra-ganglia pairs regardless of mean distance (Figure S3). We also found no correlation between mean distance and variability for both inter- and intra-ganglia pairs (Figure S3C). Known anatomical features of the worm are consistent with our findings and hypothesis of ganglial influence on the spatial relationships of neurons: the pharynx is a muscular epithelial tube[46] that rigidly encases neurons, the remaining ganglia are separated by basal lamina that loosely restricts their boundaries[43] and, finally, the anterior and ventral ganglia (and comparatively smaller retrovesicular ganglion) are completely bounded whereas all other ganglia are open at least at one end, and in White et al.,[43] it had been noted that tight cellular packing in these regions led to "slop," "uncertainty," and, in live animals, even "flipping" from side to side of the cells

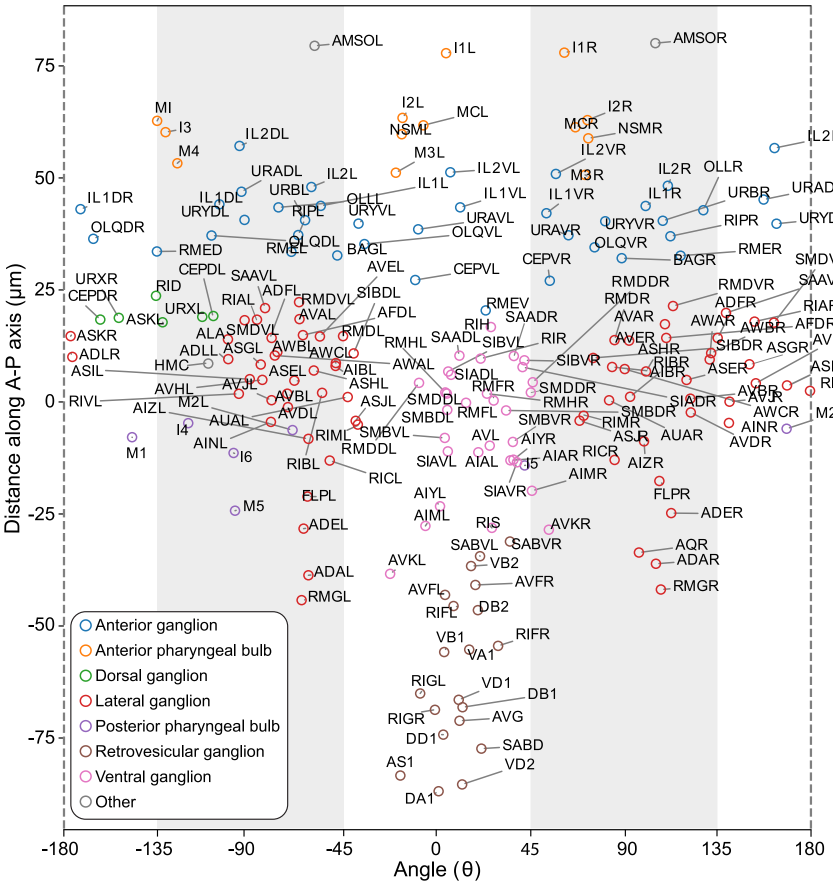


Figure 4. Harmonized multi-lab atlas of *C. elegans* neurons

Butterfly plot shows the mean locations of neurons in the atlas, colored by ganglion. The scaling of the anterior-posterior axis is slightly different from the original 3D representation of the worm due to coordinate conversion. See also Figures S1 and S2.

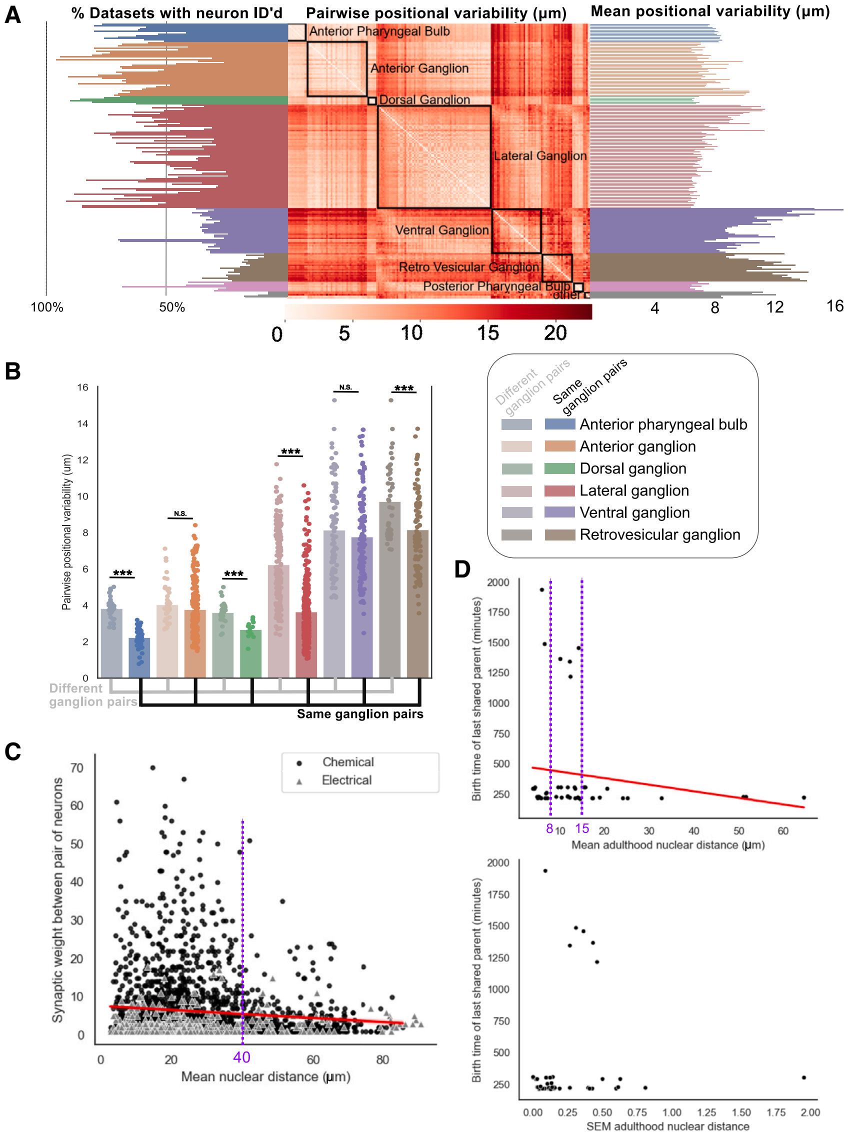


Figure 5. Analyses of neuron positions, distances, and positional variability

(A) Left: percentage of datasets containing each labeled neuron, organized anterior-to-posterior within each ganglion. Middle: heatmap of the standard deviation of pairwise positional distances between each pair of neurons across datasets. Right: averaged sums of heatmap rows. Neurons with higher mean positional variability have less stereotyped positions within the worm body.

(B) Pairwise positional variability by ganglia for 10 closest neighbors of each neuron, separating neuron pairs in the same ganglion from pairs in different ganglia. Pairs where the atlas distance was over 20 µm were removed from this analysis. Anterior pharynx: 95% CI effect size = $[-1.79, -1.38]$ µm, $p = 2.5 * 10^{-21}$, $N_{same} = 52$, $N_{diff} = 37$. Dorsal: 95% CI effect size = $[-1.15, -0.71]$ µm, $p = 7.1 * 10^{-6}$, $N_{same} = 13$, $N_{diff} = 27$. Lateral: effect size 95% CI $[-2.80, -2.38]$ µm, $p = 5.3 * 10^{-31}$, $N_{same} = 317$, $N_{diff} = 161$. Retrovesicular: 95% CI effect size = $[-1.94, -1.17]$ µm, $p = 9.3 * 10^{-5}$, $N_{same} = 89$, $N_{diff} = 44$. Anterior: $p = 0.161$, $N_{same} = 176$, $N_{diff} = 50$. Ventral: $p = 0.252$, $N_{same} = 135$, $N_{diff} = 94$.

(C) Pairwise relationship between neuron synaptic weights and their mean positional distance for chemical and electrical synapses. Chemical synapses: KendallTau $\tau = -0.036$, $p = 0.021$, Pearson $R = -0.098$, $p = 6.4 * 10^{-6,}$ $N = 2119$. Electrical synapses: KendallTau $\tau = 0.009$, $p = 0.80$, Pearson $R = 0.031$, $p = 0.51$, $N = 444$.

(D) Relationship between cell birth times and the mean and SEM of their nuclear positional distance in adulthood for sister cells. Mean: KendallTau $\tau = -0.144$, $p = 0.052$, Pearson $R = -0.161$, $p = 0.129$. SEM: KendallTau $\tau = 0.014$, $p = 0.845$, Pearson $R = 0.074$, $p = 0.487$. Most sisters are within 15 µm of each other in adulthood. More sisters that divide embryonically remain close together (<8 µm) than sisters that divide >16 h later at postembryonic larval stages.

See also Figure S3.

contained therein. This finding suggests that neural identification algorithms could be improved by a hierarchical approach, such as first predicting ganglion membership, then predicting neuron identities within this ganglion.

Next, we explored the relationship between somatic distance and synaptic connectivity. Overall, there was a very weak but statistically significant correlation between nuclear distance and synaptic weight for chemical synapses and no significance or detectable correlation for electrical synapses. However, we found that nearby neurons (mean distance <40 µm) exhibit a wide range of chemical synaptic weights ranging anywhere from 0 to 70 synapses (with a median synaptic count of 3), whereas distant neurons (mean distance >40 µm) have a maximum synaptic count of ~25 synapses (with a median count of 2) (Figure 5C). This choice of distance cutoff was chosen based on the observation of a distinct elbow at 40 µm in a 2D kernel density estimate plot of the scatter data (Figure S3D). Our findings suggest that neurons that are strongly wired together tend to be close to each other, although somatic proximity alone is not sufficient to imply strong connectivity. Recent findings in *C. elegans* have substantiated Peter's rule: neurons with larger colocalized axodendritic regions are more likely to form connections.[47] Our findings thus lend further support to this principle and suggest that close somatic or nuclear proximity also plays a role in determining neural connectivity.

Lastly, we explored the hypothesis that cell lineage is a determinant of adult cell positioning. Embryonic *C. elegans* are confined to a fixed volume within an eggshell approximately 50 µm in length and 30 µm in diameter.[48] After hatching, they grow over 4 times in length from birth (~250 µm) to adulthood (over 1 mm), with an exponential expansion in their volume.[49,50] Sister cells are those whose lineage differs only at the very last division. We hypothesized that animal growth should lead to both larger distances and higher variability between older sister

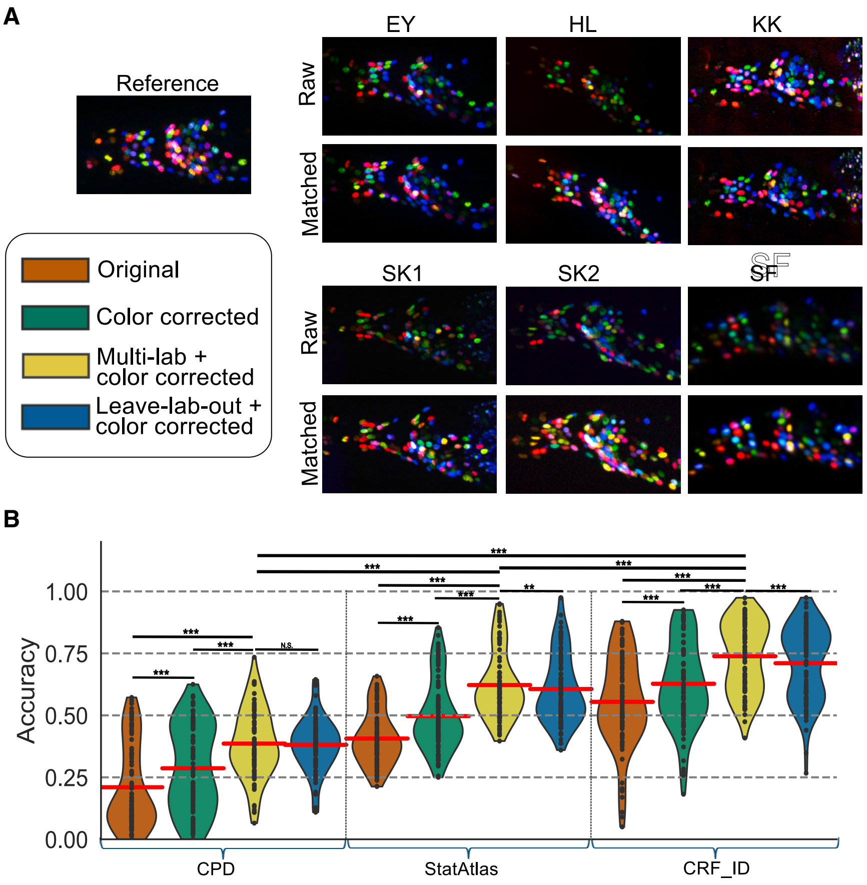


Figure 6. Improvements in neural identification accuracy

(A) Examples of raw and color-corrected (histogram matched) images from each lab and dataset.

(B) Top ranked test accuracy for training set of original 10 reference worms with no color correction (orange), 10 reference worms with color correction (green), the multi-lab corpus with color correction (yellow), and training on all but the test dataset with color correction (blue) for coherent point drift (CPD, left), the statistical atlas model (StatAtlas, middle), and the conditional random field model (CRF_ID, right). All models were tested on the same 94 datasets in the corpus. For multi-lab and leave-lab-out atlases, test datasets were held out during training and then tested using a *k*-fold cross validation strategy. Algorithmic performance was evaluated using paired t tests (where $N = 94$ for each test set) to compare the performance of different atlases. Significance is reported using a Bonferroni correction with the convention of * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. Summary statistics and p values can be found in Table S1. See also Figures S4–S7.

cells that divided in the embryo versus younger sister cells born much later at postembryonic larval stages of development. Surprisingly, we found no statistically significant correlation between the time of cell division and nuclear distance or between the time of cell division and distance variability as measured by the scanning electron microscopy (SEM) (Figure 5D). In fact, most sisters remained within 15 μm of each other (~3 nuclei apart) at adulthood, regardless of when they were born. Strikingly, a substantial cohort of embryonic sisters ended up closer together at adulthood (<8 μm) than those dividing at larval stages that occur more than 16 h later (Figure 5D). Our data rule out exponential postembryonic growth spurts as a major determinant of divergence and variability in neuron positions.

Neural identification performance increases for all laboratories and all tested algorithms when trained on a harmonized multi-lab corpus

Our previously published statistical atlas algorithm (“StatAtlas”) for automated neural identification was trained on a homogeneous dataset of 10 NeuroPAL worms.[6,27] Formerly, this 10-worm training set achieved average accuracies of 86% overall in head neurons that ranged from 50% for the ventral ganglion to 100% for the anterior pharyngeal ganglion. These accuracies facilitate neural identification, but in practice, they require substantial verification and manual correction, necessitating significant time expenditure. Moreover, the algorithm fails to generalize to datasets produced by other labs (Figure 6). We tested the performance of our previously published algorithm, StatAtlas, on each of the six aggregated datasets. Initial performance on these datasets ranged from ~21% to ~65% with an average of 41% (Figure S4). This substantial decrease in accuracy on datasets from different labs exposes the limitation of using single-lab training sets to produce tools intended for use by different labs with different instrumentation, experimental methods, and data acquisition pipelines.

To assess the performance benefits of using a large, harmonized corpus to train commonly used automated neural identification methods, we tested two more popular algorithms: coherent point drift (CPD)[51] and CRF_ID.[25] Coherent point drift is an untrained and unsupervised algorithm that (1) globally aligns a sample point cloud of neurons to a reference atlas, then (2) locally matches points from the sample to their nearest neighbors in the atlas, and finally (3) identifies sample neurons (points) by their corresponding matches in the atlas. CRF_ID is a newer graph-based approach that identifies neurons using a combination of statistics describing their individual features (e.g., absolute position and color) and pairwise spatial relationships (e.g., displacement and angle relative to each other). There are no currently published benchmarks for neural identification using CPD. Formerly, CRF_ID demonstrated a high accuracy of 83% when originally trained and tested solely on the HL dataset. Like StatAtlas, when testing the generalizability of the CPD and CRF_ID base models on the full WormID corpus, we observed poor performance with an average overall accuracy of 39% and 59%, respectively.

CPD, StatAtlas, and CRF_ID all take a conceptually similar approach of matching a sample point cloud to a reference atlas, but they each differ in what they use as the reference. For CPD, the reference is simply a single static reference point cloud; in this case, that is a single labeled NeuroPAL image. For StatAtlas, the reference is a statistical atlas of the mean and variance of neuron positions and colors across a corpus of data. CRF_ID extends the StatAtlas approach by including pairwise statistical spatial features such as the likelihood that one neuron is anterior to another or the relative distances between neuron pairs. These three approaches leverage increasingly more statistical information derived from our comprehensive corpus of data. We thus expected that CRF_ID would perform better than StatAtlas, which would in turn perform better than CPD, and indeed this was the case (Figure 6B).

After inspecting recordings from multiple labs, we hypothesized that lab-to-lab differences in color space may negatively impact algorithmic performance. Potential sources of color space variability include differences in microscope hardware, software and image settings, and configuration of the optical path. Anecdotally, in addition to these known sources of variability, researchers also typically adjust exposure, contrast, and other channel display parameters to make the composite rendered colors appear more like the images in the NeuroPAL reference manual.[52] In aggregate, this suggested that harmonizing the color space might aid automated algorithms.

Therefore, we developed an approach to match the color histogram of a sample image to a reference histogram representing ideal coloring (see STAR Methods). Histogram-matching the original small training set improved the accuracy of all three tested algorithms by an average of 8%, 9%, and 7% for CPD, StatAtlas, and CRF_ID respectively. Qualitatively, it also made composite color renderings better match the NeuroPAL reference manual, aiding users in annotating and correcting algorithmic predictions (Figures 6A and 6B).

Given this success on the original small training set, we used the histogram-matched images to train the StatAtlas and CRF_ID algorithms on the full corpus of data. Test accuracy is reported using 5-fold cross-validation where each worm is tested against an atlas that excluded that worm from the training set. For CPD, we updated the algorithm to select the best template out of the full corpus (see STAR Methods for further details). This led to significant improvement in accuracy across algorithms (Figure 6B), with an average improvement of 17%, 22%, and 18% for CPD, StatAtlas, and CRF_ID, that further raised average predictive accuracy from 22% to 39%, 41% to 62%, and 55% to 74% for them respectively. This is equivalent to a ~1.3×, ~1.6×, and ~1.7× reduction in error rate. Accuracy reached as high as 95% for several individual datasets when tested with both StatAtlas and CRF_ID. Furthermore, when considering the top 5 neural identity assignments (rather than just the top 1), the multi-lab models showed average accuracies of 65%, 86%, and 89% for CPD, StatAtlas, and CRF_ID, respectively, with some datasets reaching 100% accuracy for both StatAtlas and CRF_ID (Figures 6B and S5). In addition, we saw similar improvements in accuracy across most datasets for StatAtlas and CRF_ID when training on all except one dataset and then testing on the left-out dataset (Figures 6B and S6). This indicates that most of the benefits from retraining come from achieving a better representation of the full diversity across datasets, rather than capturing the specific nuances of any one dataset. This generalizability should enable labs to use these retrained algorithms out of the box rather than needing to do additional fine-tuning on their own data.

Differences in accuracy between datasets may have been caused by a variety of factors including poor initial alignment, optical quality, non-neuronal artifacts in the images, and nonlinear deformations of the worm body. Additionally, datasets with fewer annotated neurons had better automatic labeling accuracy, presumably because experimenters only labeled the easiest neurons to identify and left the hardest ones unannotated (Figure S7).

DISCUSSION

Aggregation and harmonization of data from a variety of different sources is necessary to build a corpus for analytical methods and machine-learning tools that generalize across the diversity of real-world data. In this work, we present a data-harmonization pipeline for analyzing whole-brain structural and activity imaging in *C. elegans*. This pipeline includes data aggregation, conversion to a standardized file format, software for analyzing these standardized datasets, pre-processing approaches to align images and color spaces, and spatial registration of sample neuron point clouds to a common atlas.

We used this corpus to study potential biological factors that organize cell position in *C. elegans.* Specifically, we find that (1) restrictions in bounding tissue and gangliar space likely contribute to variability in neuron positions, (2) neurons with somatic distances less than ~40 μm of each other show higher synaptic connectivity, and (3) sister neurons that divide in the embryo can be found closer together at adulthood than ones dividing at larval stages more than 16 h later. The positive relationship between synaptic connectivity and neuron somatic proximity thus augments the previously observed correlation of synaptic connectivity to axodendritic adjacency, termed Peter's rule. Moreover, the close distances and low positional variability we measured for embryonically born sister neurons rules out exponential organismal growth as a major cause in driving neurons apart from each other during the establishment of the adult Bauplan.

We then used the corpus to train machine-learning algorithms to automate the intensive task of labeling cells in these datasets, producing an updated statistical atlas of neuron positions and colors. The harmonized data corpus substantially boosts generalized neural identification performance across datasets from contributing labs for each tested algorithm, despite the variability in data from these different groups. Full recovery of human labeling was achieved in the top 5 automatic label assignments for 12 out of 94 datasets using either the StatAtlas or CRF_ID algorithms. This major improvement in accuracy compared to using older atlases indicates that training on our much larger corpus of data allows these systems to learn a more accurate representation of the true distribution of neuron position and color across individuals. CRF_ID performs the best of the three algorithms we tested, which is in line with our expectation because CRF_ID leverages pairwise positional features rather than only the mean and variance of individual neurons. In the future, our corpus can be used to incorporate neuronal shape- and size-based descriptors as well as dynamical time-series features to further improve neural identification algorithms. As new automated labeling algorithms continue to be developed using the tools outlined in this paper, there is the additional possibility of using ensemble methods to improve neuron labeling by leveraging multiple algorithms.

The wormid.org tools and resources are readily applicable to new whole-brain structural and activity imaging datasets, and these new datasets can be easily added to the existing corpus. These tools streamline public data sharing to facilitate both open science and to satisfy data-sharing mandates. We hope this resource will continue to grow in size and breadth to enable the development and benchmarking of new machine-learning tools and algorithms. Our analyses of cell features based on the full corpus of data can immediately be used to inform better feature selection and algorithms to continually improve automated approaches for neuron-subtype identification in volumetric images. Additionally, the large corpus and trained statistical atlas can serve as a descriptive resource of the underlying neurophysiology of *C. elegans*. Moreover, our resources can be incorporated into computational neurobiology courses, such as the Neuromatch Academy (neuromatch.io)[53] to train the next global generation of neuroscientists on real-world datasets. As the community continues to develop new tools, this corpus will allow these new tools to be benchmarked for generalizable performance, spurring innovation.

As of January 2023, the United States' National Institute of Health (NIH) has instituted a Data Management and Sharing Policy, and other agencies such as the European Research Council (ERC) have instituted similar policies. Compliance with these policies can require substantial thought, effort, and cost on the part of investigators. The wormid.org tools and resources provide a free and simple mechanism to satisfy these new policies. Furthermore, the DANDI repository permits data embargoes, thus providing researchers with a means to manage timing for the public release of their data. Our standard and resource should significantly aid researchers in policy compliance and facilitate open science and data sharing.

As the community continues to scale up the generation of neural data and increasingly relies on machine learning analysis to tame this "big data," there is an ever-growing need to unify disparate datasets to produce verifiably robust, accurate, and generalizable analytical approaches. Harmonization efforts such as ours can significantly reduce the activation energy necessary for collaboration, data sharing, and the development of unified community-wide tools across labs. While some of the resources we created are specific to *C. elegans*, the framework and much of our toolkit can be applied to other model organism imaging communities.

Limitations of the study

At present, whole-brain activity imaging in *C. elegans* is performed using nuclear-localized calcium sensors (such as NLS-GCaMP6s) rather than cytosolic or membrane-bound sensors, due principally to the ease of cell segmentation afforded by nuclear localization of fluorophores. Voltage sensors in *C. elegans* are immature, at the time of this work. The NWB format can be easily extended to accommodate higher-resolution subcellular activity data as they become available.

Although we identified developmental and anatomical factors that correlate with neural position (gangliar boundaries, timing of cell divisions, and connectivity), it remains experimentally challenging to perturb these factors to probe their causal influence.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Saul Kato (saul.kato@ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- For convenience, wormid.org links to all data and code resources listed below.
- All datasets are available on the DANDI archive, with URLs provided in STAR Methods.
- All original code repositories are available on Zenodo and Github, with URLs provided in STAR Methods.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Erdem Varol and Amin Nejatbakhsh for providing code and troubleshooting support to train the statistical atlas and to label images of dense neurons. We thank Adam Atanas and Steve Flavell for contributing their datasets. Research funding was provided by the Weill Institute for Neurosciences, the Weill Neurohub, NIGMS NIH (R35GM124735 to S.K.), Chan Zuckerberg Initiative (napari Plugin Foundation grant), Esther A. & Joseph Klingenstein Fund (E.Y.), the Simons Foundation (E.Y.), the Hypothesis Fund (E.Y), JST CREST, (JPMJCR23B2 to K.D.K.), Japan Society for the Promotion of Science (JP21H02533 and JP22KK0100 to K.D.K), RIKEN (Special Postdoctoral Researchers Program to C.W.), NINDS NIH (5U24NS120057 to B.D.), NIMH NIH (2R24MH117295, DANDI, and R01MH130064 to H.L.) and the NSF (1764406 to H.L.).

AUTHOR CONTRIBUTIONS

Conceptualization, D.Y.S, E.Y., and S.K.; methodology, D.Y.S., E.Y., S.K., and G.B.; software, D.Y.S., J.M.B., S.B., R.L.D., K.R., B.D., and R.L.; validation, D.Y.S. and K.R.; formal analysis, D.Y.S., K.R., H.J.L., and Z.Y.; resources, D.Y.S, K.R., G.B., R.L.D., J.M.B., G.C.C., C.W., R.S., S.C., H.L., E.Y., S.O., and K.D.K.; data curation, D.Y.S, K.R., G.B., R.L.D., J.M.B., G.C.C., C.W., R.S., S.C., H.L., E.Y., S.O., and K.D.K.; writing – original draft, D.Y.S., E.Y., and S.K.; writing – review & editing, all authors; visualization, D.Y.S.; supervision, S.K. and E.Y.; funding acquisition, S.K. and E.Y.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2024.100964.

Received: August 1, 2024
Revised: November 12, 2024
Accepted: December 31, 2024
Published: January 17, 2025

REFERENCES

1. 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. Nat. Methods *10*, 1013–1020. https://doi.org/10.1038/nmeth.2637.
2. Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., and Keller, P.J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods *10*, 413–420. https://doi.org/10.1038/nmeth.2434.
3. Kato, S., Kaplan, H.S., Schrödel, T., Skora, S., Lindsay, T.H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Global Brain Dynamics Embed the Motor Command Sequence of *Caenorhabditis elegans*. Cell *163*, 656–669. https://doi.org/10.1016/j.cell.2015.09.034.
4. 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. Proc. Natl. Acad. Sci. USA *113*, E1082–E1088. https://doi.org/10.1073/pnas.1507109113.
5. 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 *Caeonorhabditis elegans*. Proc. Natl. Acad. Sci. USA *113*, 1074–1081. https://doi.org/10.1073/pnas.1507110112.
6. Yemini, E., Lin, A., Nejatbakhsh, A., Varol, E., Sun, R., Mena, G.E., Samuel, A.D.T., Paninski, L., Venkatachalam, V., and Hobert, O. (2021). NeuroPAL: A multicolor atlas for whole-brain neuronal identification in *C. elegans*. Cell *184*, 272–288.e11. https://doi.org/10.1016/j.cell.2020.12.012.
7. 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*, 4134–4151.e31. https://doi.org/10.1016/j.cell.2023.07.035.
8. Portugues, R., Feierstein, C.E., Engert, F., and Orger, M.B. (2014). Whole-brain activity maps reveal stereotyped distributed networks for visuomotor behavior. Neuron *81*, 1328–1343. https://doi.org/10.1016/j.neuron.2014.01.019.
9. Prevedel, R., Yoon, Y.G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E.S., and Vaziri, A. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-fiel microscopy. Nat. Methods *11*, 727–730. https://doi.org/10.1038/nmeth.2964.
10. Royer, L.A., Lemon, W.C., Chhetri, R.K., Wan, Y., Coleman, M., Myers, E.W., and Keller, P.J. (2016). Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. Nat. Biotechnol. *34*, 1267–1278. https://doi.org/10.1038/nbt.3708.
11. Kim, D.H., Kim, J., Marques, J.C., Grama, A., Hildebrand, D.G.C., Gu, W., Li, J.M., and Robson, D.N. (2017). Pan-neuronal calcium imaging with cellular resolution in freely swimming zebrafish. Nat. Methods *14*, 1107–1114. https://doi.org/10.1038/nmeth.4429.
12. Vladimirov, N., Wang, C., Höckendorf, B., Pujala, A., Tanimoto, M., Mu, Y., Yang, C.-T., Wittenbach, J.D., Freeman, J., Preibisch, S., et al. (2018). Brain-wide circuit interrogation at the cellular level guided by online analysis of neuronal function. Nat. Methods *15*, 1117–1125. https://doi.org/10.1038/s41592-018-0221-x.
13. Chen, X., Mu, Y., Hu, Y., Kuan, A.T., Nikitchenko, M., Randlett, O., Chen, A.B., Gavornik, J.P., Sompolinsky, H., Engert, F., and Ahrens, M.B. (2018). Brain-wide organization of neuronal activity and convergent sensorimotor transformations in larval zebrafish. Neuron *100*, 876–890.e5. https://doi.org/10.1016/j.neuron.2018.09.042.
14. Lemon, W.C., Pulver, S.R., Höckendorf, B., McDole, K., Branson, K., Freeman, J., and Keller, P.J. (2015). Whole-central nervous system functional imaging in larval *Drosophila*. Nat. Commun. *6*, 7924. https://doi.org/10.1038/ncomms8924.
15. Mann, K., Gallen, C.L., and Clandinin, T.R. (2017). Whole-brain calcium imaging reveals an intrinsic functional network in *Drosophila*. Curr. Biol. *27*, 2389–2396.e4. https://doi.org/10.1016/j.cub.2017.06.076.
16. Aimon, S., Katsuki, T., Jia, T., Grosenick, L., Broxton, M., Deisseroth, K., Sejnowski, T.J., and Greenspan, R.J. (2019). Fast near-whole-brain imaging in adult Drosophila during responses to stimuli and behavior. PLoS Biol. *17*, e2006732. https://doi.org/10.1371/journal.pbio.2006732.
17. Stirman, J.N., Smith, I.T., Kudenov, M.W., and Smith, S.L. (2016). Wide field-of-view, multi-region, two-photon imaging of neuronal activity in the mammalian brain. Nat. Biotechnol. *34*, 857–862. https://doi.org/10.1038/nbt.3594.
18. Skocek, O., Nöbauer, T., Weilguny, L., Martínez Traub, F., Xia, C.N., Molodtsov, M.I., Grama, A., Yamagata, M., Aharoni, D., Cox, D.D., et al. (2018). High-speed volumetric imaging of neuronal activity in freely moving rodents. Nat. Methods *15*, 429–432. https://doi.org/10.1038/s41592-018-0008-0.
19. Klioutchnikov, A., Wallace, D.J., Frosz, M.H., Zeltner, R., Sawinski, J., Pawlak, V., Voit, K.-M., Russell, P.S.J., and Kerr, J.N.D. (2020). Three-photon head-mounted microscope for imaging deep cortical layers in freely moving rats. Nat. Methods *17*, 509–513. https://doi.org/10.1038/s41592-020-0817-9.
20. Zong, W., Obenhaus, H.A., Skytøen, E.R., Eneqvist, H., de Jong, N.L., Vale, R., Jorge, M.R., Moser, M.-B., and Moser, E.I. (2022). Large-scale two-photon calcium imaging in freely moving mice. Cell *185*, 1240–1256.e30. https://doi.org/10.1016/j.cell.2022.02.017.
21. Manley, J., Lu, S., Barber, K., Demas, J., Kim, H., Meyer, D., Traub, F.M., and Vaziri, A. (2024). Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. Neuron *112*, 1694–1709.e5. https://doi.org/10.1016/j.neuron.2024.02.011.
22. Uzel, K., Kato, S., and Zimmer, M. (2022). A set of hub neurons and non-local connectivity features support global brain dynamics in *C. elegans*. Curr. Biol. *32*, 3443–3459.e8. https://doi.org/10.1016/j.cub.2022.06.039.
23. Flavell, S.W., and Gordus, A. (2022). Dynamic functional connectivity in the static connectome of *Caenorhabditis elegans*. Curr. Opin. Neurobiol. *73*, 102515. https://doi.org/10.1016/j.conb.2021.12.002.
24. Randi, F., Sharma, A.K., Dvali, S., and Leifer, A.M. (2023). Neural signal propagation atlas of *Caenorhabditis elegans*. Nature *623*, 406–414. https://doi.org/10.1038/s41586-023-06683-4.

25. 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. https://doi.org/10.7554/eLife.60321.

26. Tekieli, T., Yemini, E., Nejatbakhsh, A., Wang, C., Varol, E., Fernandez, R.W., Masoudi, N., Paninski, L., and Hobert, O. (2021). Visualizing the organization and differentiation of the male-specific nervous system of C. elegans. Development *148*, dev199687. https://doi.org/10.1242/dev.199687.

27. 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 – MICCAI 2020. https://doi.org/10.1007/978-3-030-59722-1_12.

28. Nejatbakhsh, A., Varol, E., Yemini, E., Hobert, O., and Paninski, L. (2020). Probabilistic Joint Segmentation and labeling of C. elegans neurons. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. https://doi.org/10.1007/978-3-030-59722-1_13.

29. Skuhersky, M., Wu, T., Yemini, E., Nejatbakhsh, A., Boyden, E., and Tegmark, M. (2022). Toward a more accurate 3D atlas of *C. elegans* neurons. BMC Bioinf. *23*, 195. https://doi.org/10.1186/s12859-022-04738-3.

30. 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 Biol. *18*, 30. https://doi.org/10.1186/s12915-020-0745-2.

31. Bubnis, G., Ban, S., DiFranco, M.D., and Kato, S. (2019). A probabilistic atlas for cell identification. Preprint at arXiv. https://arxiv.org/abs/1903.09227.

32. 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 Comput. Biol. *18*, e1010594. https://doi.org/10.1371/journal.pcbi.1010594.

33. 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. https://doi.org/10.7554/eLife.66410.

34. Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C., Grujic, V., Model, T., Kubinec, R., and Barceló, J. (2024). A general primer for data harmonization. Sci. Data *11*, 152. https://doi.org/10.1038/s41597-024-02956-3.

35. Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. IEEE Sig. Proc. Mag. *29*, 141–142. https://doi.org/10.1109/MSP.2012.2211477.

36. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

37. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. (2021). Datasets: A Community Library for Natural Language Processing. Preprint at arXiv. https://arxiv.org/abs/2109.02846.

38. Halchenko, Y., Wodder II, J.T., Christian, H., Ghosh, S., Sharda, S., Jarecka, D., Baker, C., Chiquito, D., Dichter, B., Gunalan, K., et al. (2024). dandi/dandi-cli: 0.61.2. Zenodo. https://doi.org/10.5281/zenodo.3692138.

39. Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295–300. https://doi.org/10.1038/nature12354.

40. Chronis, N., Zimmer, M., and Bargmann, C.I. (2007). Microfluidics for in vivo imaging of neuronal and behavioral activity in Caenorhabditis elegans. Nat. Methods *4*, 727–731. https://doi.org/10.1038/nmeth1075.

41. Rübel, O., Tritt, A., Ly, R., Dichter, B.K., Ghosh, S., Niu, L., Baker, P., Soltesz, I., Ng, L., Svoboda, K., et al. (2022). The Neurodata Without Borders ecosystem for neurophysiological data science. Elife *11*, e78362. https://doi.org/10.7554/eLife.78362.

42. Kuhn, H.W. (1955). The Hungarian method for the assignment problem. Nav. Res. Logist. *2*, 83–97. https://doi.org/10.1002/nav.3800020109.

43. White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. Phil. Trans. R. Soc. Lond. B *314*, 1–340. https://doi.org/10.1098/rstb.1986.0056.

44. Cook, S.J., Jarrell, T.A., Brittin, C.A., Wang, Y., Bloniarz, A.E., Yakovlev, M.A., Nguyen, K.C.Q., Tang, L.T.-H., Bayer, E.A., Duerr, J.S., et al. (2019). Whole-animal connectomes of both *Caenorhabditis elegans* sexes. Nature *571*, 63–71. https://doi.org/10.1038/s41586-019-1352-7.

45. Witvliet, D., Mulcahy, B., Mitchell, J.K., Meirovitch, Y., Berger, D.R., Wu, Y., Liu, Y., Koh, W.X., Parvathala, R., Holmyard, D., et al. (2021). Connectomes across development reveal principles of brain maturation. Nature *596*, 257–261. https://doi.org/10.1038/s41586-021-03778-8.

46. Mango, S.E. (2007). The *C. elegans* pharynx: a model for organogenesis. In WormBook. https://doi.org/10.1895/wormbook.1.129.1.

47. Cook, S.J., Kalinski, C.A., and Hobert, O. (2023). Neuronal contact predicts connectivity in the *C. elegans* brain. Curr. Biol. *33*, 2315–2320.e2. https://doi.org/10.1016/j.cub.2023.04.071.

48. D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess, eds. (1997). *C. elegans* II, 2nd Edition (Cold Spring Harbor Laboratory Press).

49. Byerly, L., Cassada, R.C., and Russell, R.L. (1976). The life cycle of the nematode *Caenorhabditis elegans*: I. Wild-type growth and reproduction. Dev. Biol. *51*, 23–33. https://doi.org/10.1016/0012-1606(76)90119-6.

50. Stojanovski, K., Großhans, H., and Towbin, B.D. (2022). Coupling of growth rate and developmental tempo reduces body size heterogeneity in *C. elegans*. Nat. Commun. *13*, 3132. https://doi.org/10.1038/s41467-022-29720-8.

51. Myronenko, A., and Song, X. (2009). Point-set registration: coherent point drift. Preprint at arXiv. https://doi.org/10.48550/arXiv.0905.2635.

52. Yemini, E. (2022). NeuroPAL annotations manual. https://www.yeminilab.com/neuropal.

53. Neuromatch. neuromatch.io

54. Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev. Biol. *100*, 64–119. https://doi.org/10.1016/0012-1606(83)90201-4.

55. Yemini, E., Venkatachalam, V., Lin, A., Varol, E., Nejatbakhsh, A., Sprague, D., Samuel, A., Paninski, L., and Hobert, O. (2023). NeuroPAL: Atlas of *C. elegans* neuron locations and colors in NeuroPAL worm (Version 0.240614.1942). DANDI archive. https://doi.org/10.48324/dandi.000715/0.241009.1514.

56. Yemini, E., Venkatachalam, V., Lin, A., Varol, E., Nejatbakhsh, A., Sprague, D., Samuels, A., Paninski, L., and Hobert, O. (2023). NeuroPAL microfluidic chip images and GCaMP activity (Version 0.241009.1457). DANDI archive. https://doi.org/10.48324/dandi.000541/0.241009.1457.

57. Chaudhary, S., Sprague, D., Lee, S.A., Li, Y., Patel, D.S., and Lu, H. (2023). Segmented and labeled NeuroPAL structural images (Version 0.240611.1954). DANDI archive. https://doi.org/10.48324/dandi.000714/0.241009.1516.

58. Suzuki, R., Wen, C., Sprague, D., Onami, S., and Kimura, K.D. (2023). Whole-brain spontaneous GCaMP activity with NeuroPAL cell ID information of semi-restricted worms (Version 0.240402.2118). DANDI archive. https://doi.org/10.48324/dandi.000692/0.240402.2118.

59. Atanas, A., Kim, J., Wang, Z., Bueno, E., Becker, M., Kang, D., Park, J., Kramer, T., Wan, F., Baskoylu, S., et al. (2023). Brain-wide representations of behavior spanning multiple timescales and states in C. elegans (Version 0.240625.0022). DANDI archive. https://doi.org/10.48324/dandi.000776/0.241009.1509.

60. Dunn, R., Sprague, D., and Kato, S. (2023). C. elegans whole-brain neuroPAL and immobilized calcium imaging (Version 0.240625.0439). DANDI archive. https://doi.org/10.48324/dandi.000565/0.241009.1504.

61. Sprague, D., Borchardt, J., Dunn, R., Bubnis, G., and Kato, S. (2023). NeuroPAL volumetric images (Version 0.240625.0454). DANDI archive. https://doi.org/10.48324/dandi.000472/0.241009.1502.

62. Gonzales, R.C., and Fittes, B.A. (1977). Gray-level transformations for interactive image enhancement. Mech. Mach. Theory *12*, 111–122. https://doi.org/10.1016/0094-114X(77)90062-3.

63. Emmons, S.W., Yemini, E., and Zimmer, M. (2021). Methods for analyzing neuronal structure and activity in *Caenorhabditis elegans*. Genetics *218*, iyab072. https://doi.org/10.1093/genetics/iyab072.

64. Wen, C., Miura, T., Voleti, V., Yamaguchi, K., Tsutsumi, M., Yamamoto, K., Otomo, K., Fujie, Y., Teramoto, T., Ishihara, T., et al. (2021). 3DeeCell-Tracker, a deep learning-based pipeline for segmenting and tracking cells in 3D time lapse images. Elife *10*, e59187. https://doi.org/10.7554/eLife.59187.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Original NeuroPAL	Yemini et al.[6]	https://doi.org/10.48324/dandi.000715/0.241009.1514
EY dataset	Yemini et al.[6]	https://doi.org/10.48324/dandi.000541/0.241009.1457
KK dataset	This paper	https://doi.org/10.48324/dandi.000692/0.240402.2118
HL dataset	Chaudhary et al.[25]	https://doi.org/10.48324/dandi.000714/0.241009.1516
SF dataset	Atanas et al.[7]	https://doi.org/10.48324/dandi.000776/0.241009.1509
SK1 dataset	This paper	https://doi.org/10.48324/dandi.000565/0.241009.1504
SK2 dataset	This paper	https://doi.org/10.48324/dandi.000472/0.241009.1502
Synaptic connectivity dataset	Cook et al.[47]	https://doi.org/10.1016/j.cub.2023.04.071
Cell lineage dataset	Sulston et al.[54]	https://doi.org/10.1016/0012-1606(83)90201-4
Software and algorithms		
Code for conversion and analysis	This paper	https://doi.org/10.5281/zenodo.13910335
ndx-multichannel-volume (*C. elegans* extension for NWB)	This paper	https://github.com/focolab/ndx-multichannel-volume
NeuroPAL_ID software source code	Yemini et al.[6]	https://github.com/Yemini-Lab/NeuroPAL_ID
NeuroPAL_ID software for Mac OS	Yemini et al.[6]	https://doi.org/10.5281/zenodo.13906028
NeuroPAL_ID software for Windows OS	Yemini et al.[6]	https://doi.org/10.5281/zenodo.13905893
Eats-worm	This paper	https://doi.org/10.5281/zenodo.13910463
CPD	Yu et al.[33]	https://github.com/XinweiYu/fDNC_Neuron_ID
StatAtlas	Varol et al.[27]	https://github.com/amin-nejat/stat-atlas
CRF_ID	Chaudhary et al.[25]	https://github.com/lu-lab/CRF_Cell-ID
Other		
WormID.org	This paper	https://WormID.org
DANDI archive	Halchenko et al.[38]	https://dandiarchive.org/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Details on the specific strains and data acquisition systems for each dataset are presented below. All recorded worms are young adult hermaphrodites.

Dataset	Microscope	Length of recording	Sample rate	Resolution (um/pixel)	Strain	Setup
NP_og	Zeiss LSM880 spinning disk confocal	∼4 min	∼4Hz	0.208 × 0.208 × 1.02	OH16230	Microfluidic chip
SF	Andor spinning disk confocal w/Nikon ECLIPSE Ti microscope 40x water immersion	∼15 min	1.7 Hz	0.54 × 0.54 × 0.54	Various	Freely moving
SK1	Leica DMi8 inverted spinning disk confocal, 40x WI, 1.1 NA	∼25 min	1.04 Hz	0.1604 × 0.1604 × 1 (3 for calc images) OR 0.3208 × 0.3208 × 0.75 (2.5 for calc images)	FC121, FC128, OH16230	Microfluidic chip

(*Continued on next page*)

Continued

Dataset	Microscope	Length of recording	Sample rate	Resolution (um/pixel)	Strain	Setup
SK2	Leica DMi8 inverted spinning disk confocal, 40x WI, 1.1 NA	∼15 min	3.3 Hz	0.3208 × 0.3208 × 0.75 (1.5 for calc images)	OH16230	Microfluidic chip
KK	Nikon Eclipse Ti-U i nverted spinning disk confocal, 40x 1.3 NA	∼15–20	1.67	0.32 × 0.32 × 1.5 for both images	KDK92	Semi-restricted in microfluidic device
HL	Perkin Elmer spinning disk confocal 1.3 NA, 40x oil OR Brucker Opterra II swept field confocal 0.75 NA, 40x air	NA	NA	0.33 × 0.33 × 1	OH15495	Microfluidic device
EY	Spinning disk confocal	∼4 min	∼4 Hz	0.27 × 0.27 × 1.5	OH16230	

METHOD DETAILS

Standardized file format - Neurodata Without Borders

NWB is an HDF5-based format built specifically for neurophysiology data and has emerged as the *de facto* standard for storing neurophysiology datasets with associated metadata for reuse and sharing. NWB provides object types for data and metadata including acquisition parameters, segmentation of 3D image regions, fluorescent time series (e.g., for neural activity), experimental design information, multichannel electrophysiology time series data, 3D images, stimulus events during an experiment, and behavioral data.[41]

The base NWB schema supports two-dimensional structural and time-series multi-channel images but did not originally support the type of five-dimensional (multi-channel, volumetric, time-series) data that is used in *C. elegans* whole-brain activity imaging or other metadata associated with these types of experiments. To solve this problem, we developed ‘ndx-multichannel-volume’ as a novel extension to the existing Neurodata Without Borders (NWB) standardized file format. More information and resources about NWB can be found at nwb.org.

Our extension adds new objects built from existing ones in the schema to add and improve support for multi-channel, volumetric, time-series images and the metadata associated with those images as well as volumetric segmentation data and metadata fields specific to *C. elegans* (e.g., cultivation temperature and growth stage). This extension and the datasets presented in this work represent, to the best of our knowledge, the first applications of the NWB data format to *C. elegans* and have now been incorporated as the standard for this model organism. This extension is flexible, open-source, and can be continuously updated to incorporate new types of data for future experiments.

Storage on DANDI archive

Data and associated metadata were uploaded to the DANDI archive [RRID:SCR_017571] using the Python command line tool (https://doi.org/10.5281/zenodo.3692138). The data were first converted into the NWB format (https://doi.org/10.1101/2021.03.13.435173) and organized into a structure akin to the Brain Imaging Data Structure (BIDS) (https://doi.org/10.1038/sdata.2016.44).

All datasets can be streamed or downloaded from the DANDI archive, available on WormID.org as well as these individual URLs:[55–61]

Original NeuroPAL: https://doi.org/10.48324/dandi.000715/0.241009.1514.
EY: https://doi.org/10.48324/dandi.000541/0.241009.1457.
HL: https://doi.org/10.48324/dandi.000714/0.241009.1516.
KK: https://doi.org/10.48324/dandi.000692/0.240402.2118.
SF: https://doi.org/10.48324/dandi.000776/0.241009.1509.
SK1: https://doi.org/10.48324/dandi.000565/0.241009.1504.
SK2: https://doi.org/10.48324/dandi.000472/0.241009.1502.

We present two software packages with user-friendly GUIs to interface with NWB datasets and run standard analysis pipelines for neural detections/segmentation, identification, tracking, and extracting time-series of neural-activity traces annotated with any experimental stimuli presented. First, we present the NeuroPAL_ID software (Figure 3A) for visualization, annotation, neuronal segmentation and identification, neural tracking, activity trace extraction and stimulus presentation data of volumetric NeuroPAL images and whole-brain activity. This software is pre-compiled for use on MacOS and Windows. The software is open-source, available from https://github.com/Yemini-Lab/NeuroPAL_ID/releases, and is written in MATLAB and Python. It has now been updated to include functionality described in this paper to enable histogram matching, color-corrected image visualization, and automated neural identification using the new StatAtlas. This software is written and managed by the Yemini Lab; further information can be found at https://www.yeminilab.com/neuropal.

Second, we present the eats-worm software (Figure 3B) for visualization, segmentation, and activity extraction of neural-activity time series from immobilized worms. eats-worm similarly allows for manual verification and curation of the automatic segmentation and tracking algorithms. The tracking algorithm was optimized for tracking neurons across frames in immobilized worms, but there are currently efforts to extend this functionality to work for freely-moving worms as well. eats-worm is written in Python and is built as a plugin to napari, a popular 3D visualization tool. This software is written and managed by the Kato Lab; further information can be found at https://github.com/focolab/eats-worm.

Both software programs have embedded functionality to read and write NWB files. NWB I/O functionality enables a user to quickly run similar analyses on all the datasets presented in this work without the need to develop specific pipelines to read in data from each dataset. Furthermore, this functionality can be easily embedded into MATLAB or Python-based analysis software.

All other code used for data conversions and analysis can be found at https://github.com/focolab/NWBelegans and is publicly available.

Data acquisition

NeuroPAL structural volumes and neural activity time series volumes were acquired using the protocols outlined in Yemini et al. 2021.[6] After collection of these images, neurons were marked and annotated according to the guidelines in the NeuroPAL manual.[52] Specific immobilization methods, microscope setup, and experimental protocols differ slightly between datasets. All datasets were taken using spinning disk confocal microscopes with *xy* (lateral) resolution varying from 0.1604–0.54 μm/pixel and *z* (axial) resolutions varying from 0.54–1.5 μm/pixel. *xy* resolution was the same for NeuroPAL structural images and neural activity images (using GCaMP6s) for all datasets, but *z* resolution varied from 0.54 to 3 μm/pixel. *z* resolution is generally lower for neural activity images due to limitations in optical sectioning with confocal microscopes. Lower *z* resolution also reduces the number of frames needed to record a full volume for a single time-point, to allow imaging at a higher temporal resolution. Most images were taken with the worm immobilized in a microfluidic chip with the exception of the KK dataset (where worms were semi-restricted in a microfluidic device) and the SF dataset (where worms were freely moving). The NWB files and the DANDI datasets that hold them contain metadata for the specific setup and conditions in each dataset. For published datasets, additional information can be found in the associated publications.[6,7]

After acquisition of NeuroPAL structural volumes and whole-brain activity time-series, images were annotated using various automated detection/segmentation algorithms ranging from classical computer vision approaches (e.g., template matching[62]) to deep neural network approaches. These were then manually verified. Ground truth annotations were done using a combination of existing automated identification algorithms followed by manual corrections. Each neuron identity label was either explicitly annotated by experts or manually verified after algorithmic identification. Note that varying levels of completeness in labeling are due to the difficulty of this manual annotation task. For several datasets with lower image quality, even experts could only confidently label 30–50% of segmented neurons in the volume. For neural activity time-series, neuron centers were first tracked across images using various algorithms and then manually verified by experts.[63,64] Fluorescence activity is then extracted from these tracked ROIs to obtain time series of neural-activity traces. Neurons in the NeuroPAL structural volume were then matched to the ROIs in the neural activity time-series to get labeled activity traces.

Datasets from various labs were converted to the NWB standardized file format using the *ndx-multichannel-volume* extension presented in this work. These files were then uploaded to the DANDI archive where they are now publicly accessible for data streaming, download, or online visualization.

Butterfly plot

To produce the butterfly plot, we first manually found three orthogonal basis vectors to align neuron point clouds to a new cartesian coordinate space. To do so, we used human-guided affine transformation to roughly align these basis vectors to the anterior-posterior, dorsal-ventral, and left-right axes. The XYZ coordinates of each neuron were projected into this new cartesian coordinate space and then converted to cylindrical coordinates by the following equations.

$$x_{cylinder} = -x_{new}$$

$$r = \sqrt{{y_{new}}^2 + {z_{new}}^2}$$

$$\theta = arctan\,2(y_{new}, z_{new})$$

We plotted the new *x* and *θ* coordinates on a 2D plane to obtain the butterfly plots shown in (Figures 1A and 4). This projection is akin to flattening the positions of the neurons along the circumference of a cylinder of the worm body and then unrolling that cylinder into a flattened plane.

Histogram matching

We modified the established approach of histogram matching to apply it to 3-D volumetric, multi-channel data.[62] We created a reference histogram using the 10 worms from the original NeuroPAL work. This data is stored as uint16 (65,536 possible values for each pixel). For each channel, we created a histogram counting the number of pixels within the bin edges, assigning each color value its own bin, and then averaged the values in each of these bins across the 10 images. Practically, these histograms were very similar across these 10 datasets, so the averaged histogram was similar to each of the individual histograms.

To color match a new animal sample, we calculated a histogram for each channel. The number of bins for each channel histogram was equal to the maximum intensity value present in that channel in the image: $bincount_{channel} = max(x_{channel})$. This meant that there was a different number of histogram bins for each channel in each image because images were collected at different bit depths and with varying levels of saturation.

We then calculated the cumulative density function (CDF) at each color value of both the sample and the reference. We created a lookup table M to associate each gray count value x in the sample to the color value in the reference with the closest CDF value. Next, we created a new matched image with each pixel transformed into the new color space using this lookup table as shown below.

$$M_{channel}(x) = \left(\left|cdf_{sample}(x) - cdf_{reference}(x')\right|\right)$$

$$MatchedImage_{channel}(i,j,k) = M_{channel}(A(i,j,k))$$

Color extraction

To extract the color values for the neurons in each image, we first calculated the mean and standard deviation of the pixel counts in each channel, and then converted each pixel value into its *Z* score based these channel values. We then sampled a 3 × 3 × 1 grid of pixel values around each segmented neuron center in each channel. We use the median values of this 3 × 3 × 1 grid as the RGB values for that neuron center. Color values were extracted post-histogram matching when training or testing using histogram-matched images. For non-histogram-matched images, there are no additional color pre-processing steps beyond Z-scoring.

Positional variability analysis

We calculated pairwise positional variability by measuring the Euclidean distance between every pair of canonical head neurons across each structural volume when both neurons in that pair had a ground truth label. We then took the average and standard deviation of these distances for each neuron pair to find mean nuclear distance and pairwise positional variability, respectively. For these analyses we ignored pairs that are not present in at least 5 datasets. We used pairwise positional variability instead of absolute positional variability because absolute position is extremely sensitive to point-cloud realignment, which would make it hard to disaggregate natural positional variability from alignment errors; furthermore, we are interested in how individual cells vary relative to each other, not in how they vary individually. To get the mean positional variability for a given neuron, we averaged the mean pairwise distance for all neuron pairs that contained that neuron.

We calculated Intra vs. inter ganglion measures as follows: for every neuron in the atlas, we found its n closest atlas neighbors and only performed measurements for these pairwise neighbors. We then separated these pairings based on whether the two neurons in the pair are within the same ganglion or in two different ganglions. Note that pairs in different ganglions will appear twice: e.g., if one neuron in the pair is in the anterior ganglion and the other is in the lateral ganglion, the pair will be counted in the analysis for both the anterior ganglion and the lateral ganglion (Figure 5B). Pairs within the same ganglion are only counted once. We compared this approach for n = 1–20 (Figure S4). For all numbers of neighbors there is higher positional variability for neighbors in different ganglia when compared to those in the same ganglion. The measurements stabilizes around *n* = 7 and holds steady through *n* = 20. Therefore, we selected 10 to use for n in our analysis.

Synaptic connection weights between neuron pairs are derived from the whole-brain connectome of the adult hermaphrodite in Cook et al.[44]

For lineal distance: the cell lineage tree and associated birth times were taken from Sulston et al.[64] The last shared parent cell between two neurons is the most recent shared parent node in the lineal tree. We used the birth time of the last shared parent cell between two neurons as the lineal distance and explored the relationship between this lineal distance and mean pairwise nuclear distance (Figure 5D). In this analysis, we focus only on sister cells: terminal cells that only divided from each other at the very last stage of their lineal tree.

Alignment of datasets

Before analysis, datasets were first roughly aligned to a common space by learning an affine transformation as described in Varol et al.[27] The algorithm chooses a reference dataset from the 10 original NeuroPAL worms to align every other dataset to. For each dataset, the algorithm then learns an affine transformation that minimizes the distance of each ground truth labeled neuron to the same neuron in the reference dataset. This roughly aligned the principal axes of the worm as well as the overall scaling across datasets. The accuracy of each automated identification algorithm is highly sensitive to pre-alignment and thus this step is critical for direct comparison of models and quantification of atlas performance. Pre-alignment is accomplished by labeling a small number of easily identifiable neurons.

Coherent point drift (CPD)

Coherent point drift has been a common algorithm for registering two similar point clouds to each other since its introduction in Myronenko and Song.[51] CPD allows for both rigid and non-rigid point set registration. CPD models one point set as a set of GMM centroids that are fit to the second point set by maximizing the likelihood. GMM centroids are set to move coherently to preserve the structure of the point clouds. In the rigid case, the algorithm learns an affine transformation of the GMM centroid locations while in the non-rigid case, the algorithm learns a displacement function on the original centroid positions with an enforced regularization term to enforce smoothness. The objective function is optimized using an iterative EM optimization approach and yields both the aligned point set as well as an *NxM* correspondence probability matrix that represents the likelihood that each point *n* in set 1 corresponds to each point *m* in point set 2.

In this paper, we use the specific implementation of CPD used in Yu et al.[33] First, rigid CPD is used to roughly align a test worm point cloud to a template point cloud. Then, non-rigid CPD is used to model non-linear deformations between the semi-aligned test and template. Neuron assignments are then determined by creating a matrix of pairwise Euclidean distances between every neuron's position and color in the test and every neuron in the template in the aligned space. We then use the Hungarian algorithm on this distance matrix to find the optimal label assignments. To get 2nd ranked assignments, we assigned an infinite cost to each label assignment from the first pass and reran the Hungarian algorithm. We repeat this for the 3rd-5th order assignments.

Accuracy was calculated by counting the number of neurons whose algorithmic assignment was the same as the ground truth label and then dividing by the total number of neurons that have a ground truth label. Note that neurons without a ground truth label were not included in the accuracy metric but are still part of the cost matrix and received neuron assignments. Since there is no ground truth for these neurons we did not determine the accuracy of their label assignments.

The difference between the use of the 'original (10) worms' versus 'multi-lab corpus' for CPD is what set is included in possible options for the template. For the 'original' group, we compare every test set to each of the original 10 NeuroPAL worms and report the accuracy for the template that has the highest average probability of correspondence after the rigid alignment step. Similarly, for the 'multi-lab corpus', each test worm is compared to each possible template worm in the whole multi-lab corpus and accuracy is reported similarly. The accuracy of CPD is highly sensitive to a good rough initial alignment and to similarity of the template and the test point cloud. The template with the highest average probability of correspondence is not necessarily the template that yields the highest accuracy, but it is the template that the algorithm has the highest confidence that it has found the 'correct' correspondence.

Statistical atlas training and inference

Statistical atlases used for testing performance were trained using the algorithm described in Varol et al.[27] This algorithm uses a training set of neuron point clouds with both XYZ and RGB values and takes a block-coordinate descent approach where it iteratively learns affine transformation parameters to align the neuron point clouds, then updates the means and covariances of the positions and colors of each neuron until reaching convergence. This process generates mean and covariance parameters for each neuron as well as an aligned coordinate space for all the worms in the training set. The trained atlas consists of a list of neuron names alongside their associated means and covariances in the aligned position and color space.

We trained three atlases: the original atlas trained on just the original 10 NeuroPAL worms from Yemini et al. 2021,[6] the color corrected atlas trained on these same 10 worms after histogram matching, and the multi-lab + color-corrected atlas which is trained on the full corpus of histogram-matched data.

For the original atlas, we tested every dataset in the full corpus without histogram matching. For the color-corrected atlas, we similarly tested every dataset on the full corpus of data with histogram matching. For the atlas trained on the full corpus, we use k-fold cross-validation. The corpus was split into five equally sized groups. For each group, an atlas was trained on all datasets in the other four groups and performance was reported for the out-of-training set group. The 10 worms used to train the original and color-corrected atlas were included in the training for each of these five groups. These 10 worms were not used to report testing accuracy for any of the atlases (Figure 6). The fully trained atlas presented in Figure 4 was trained using the 10 original worms and the full corpus of data presented in this work, without splitting it into groups. This full atlas was embedded into the "Auto ID" functionality of the NeuroPAL ID software shown in Figure 3A.

Neuron point clouds used for testing were pre-aligned by learning an affine transformation from each sample dataset to the aligned coordinates of the atlas based on a subset of the ground truth labeled neurons in the sample. Briefly, assuming *N* neurons in the test sample and *M* neurons in the atlas, we calculated an *NxM* cost matrix using the Mahalanobis distance between each neuron center in the sample and each neuron distribution in the atlas. xi represents the XYZRGB values of neuron *i*, while μ_j and Σ_j represent the XYZRGB mean and covariance respectively for neuron *j* in the atlas:

$$Cos\ t_{i,j} = \left(\underline{x_i} - \underline{\mu_j}\right)^T \Sigma_j^{-1}\left(\underline{x_i} - \underline{\mu_j}\right)$$

We then treated this cost matrix as a linear sum assignment problem. Label assignments (for neural identification) were calculated using the Hungarian algorithm.[42] 2nd-5th order ranked assignments and accuracy are calculated in the same way as described for CPD.

CRF_ID training and inference

CRF_ID atlases and inference are conducted using the algorithm described in Chaudhary et al.[25] This approach follows a probabilistic-graphical-model framework based on conditional random fields. The graph is defined by node features corresponding to unary measures for each neuron center such as position and color and edge features corresponding to pairwise measures for each pair of neurons such as distance, relative angle, or probability that one neuron is anterior to the other. After features are selected, a data-driven atlas is trained on a corpus of data to determine the average values for each of the measured features; then for a test worm, node and edge potentials are calculated based on comparison of each feature in the test worm to the atlas and infer the best global assignment of labels by maximizing an energy function using an approximate inference method. For the analysis in this work, we used the color information solely to define the node potentials, and the pairwise angle relationships only to define the edge potentials. Optimizing the weights of the node and edge features may result in a higher prediction accuracy.

We trained three atlases: the original atlas trained on just the original 10 NeuroPAL worms from Yemini et al.,[6] the color corrected atlas trained on these same 10 worms after histogram matching, and the multi-lab + color-corrected atlas which is trained on the full corpus of histogram-matched data. This training approach follows the same k-fold cross validation approach used for the StatAtlas method.

We use the roughly pre-aligned point clouds used in the StatAtlas algorithm as input to the CRF_ID algorithm to eliminate possible differences in the initial alignment step, which can dramatically change accuracy.

In practice, there are nearly always fewer detected neuron centers in a given image than total cells in the atlas. CRF_ID handles this by modeling a hidden variable $\boldsymbol{h} \in \{0,1\}^N$ where N is the number of neurons in the atlas. This variable specifies the probability that a given cell is missing in the image. Based on the number of cells in the test image, P cells are uniformly selected across different regions of the head and removed from the atlas. This process is repeated ~1000 times to sample multiple possible combinations of $\boldsymbol{h}$. The top 1–5 predicted assignments are generated by compiling a list of the most frequent labels for each cell in the test image across all runs. Accuracy is reported in the same way as CPD and CRF_ID.

Optimizing the aforementioned energy function using an approximate inference method produces marginal distributions of label assignments for each cell. The top 1–5 predicted label assignments for each cell were generated by sorting the marginal probability of labels in a descending order. The label that resulted in the highest marginal probability was assigned as top 1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Descriptions of statistical tests can be found in the text or figure captions where the analysis was conducted. All analyses were done in Python 3.12. Specific packages and versions used in analyses can be found in the setup.py file in the NWBelegans Github repository https://github.com/focolab/NWBelegans.

ADDITIONAL RESOURCES

We present WormID.org as an additional resource to guide users through the full process of collecting data, converting to the NWB format, uploading to DANDI, and replicating the analyses in this work. The website contains tutorials, links to code and data, and step-by-step instructions for data conversion.