nature methods

Resource

https://doi.org/10.1038/s41592-023-01848-5

BigNeuron: a resource to benchmark and predict performance of algorithms for automated tracing of neurons in light microscopy datasets

Received: 10 May 2022

Accepted: 14 March 2023

[redacted]

Check for updates

A list of authors and their affiliations appears at the end of the paper

BigNeuron is an open community bench-testing platform with the goal of setting open standards for accurate and fast automatic neuron tracing. We gathered a diverse set of image volumes across several species that is representative of the data obtained in many neuroscience laboratories interested in neuron tracing. Here, we report generated gold standard manual annotations for a subset of the available imaging datasets and quantified tracing quality for 35 automatic tracing algorithms. The goal of generating such a hand-curated diverse dataset is to advance the development of tracing algorithms and enable generalizable benchmarking. Together with image quality features, we pooled the data in an interactive web application that enables users and developers to perform principal component analysis, *t*-distributed stochastic neighbor embedding, correlation and clustering, visualization of imaging and tracing data, and benchmarking of automatic tracing algorithms in user-defined data subsets. The image quality metrics explain most of the variance in the data, followed by neuromorphological features related to neuron size. We observed that diverse algorithms can provide complementary information to obtain accurate results and developed a method to iteratively combine methods and generate consensus reconstructions. The consensus trees obtained provide estimates of the neuron structure ground truth that typically outperform single algorithms in noisy datasets. However, specific algorithms may outperform the consensus tree strategy in specific imaging conditions. Finally, to aid users in predicting the most accurate automatic tracing results without manual annotations for comparison, we used support vector machine regression to predict reconstruction quality given an image volume and a set of automatic tracings.

Q1 Quantification of neuron morphology is an essential process in defining neuron type, assessing neuronal changes in development and aging, determining effects of brain disorders and treatments, and providing important parameters for neuronal computations. However, quantifying the three-dimensional structure of neuronal trees has remained a challenge[1,2], even though researchers have been developing methods for fully automated neuron reconstruction for nearly four decades[3,4]. While automatic reconstruction of neuron tree structures

e-mail: meijering@imagescience.org; ascoli@gmu.edu; h@braintell.org

based on three-dimensional (3D) microscopy imaging datasets was expected to be a feasible task for computers, experience during the Q2 last decades has underlined the difficulty of this challenge. Diversity in Q3 Q4 animal species, developmental stages, brain location and image quality Q5 Q6 of microscopy datasets implies that algorithms with an impressive Q7 performance in small sets of images do not generalize well when applied Q8 to image volumes obtained under different conditions.

Q9 Advances in labeling[5–7], tissue preparation[8,9] and imaging techniques Q10 [10–12] enable both individual laboratories and large-scale brain Q11 science projects[13–17] to generate increasingly large fluorescence microscopy Q12 datasets for the reconstruction of single neurons. Several automatic tracing algorithms have been developed[18–23], and individual groups have tackled the challenge of applying automatic neuron tracing by mainly focusing on their own datasets[24,25]. Improving labeling and imaging quality is key for simplifying the task of automatic reconstruction[26], but bottlenecks remain given that manual correction and fine-tuning by experts are still needed. A faithful annotation of neuron morphology is relevant for estimating potential connectivity between brain regions. This is especially important when whole brain modeling techniques rely on synthetic generation of neuron populations based on annotation data[27]. Artifacts in the tracing process can result in altered tree topology, leading to unreliable simulation of signal integration and transmission when modeling neuronal networks. Understanding the performance of available algorithms and how they match with specific characteristics of different imaging datasets is crucial for achieving fully automatic neuron tracing[28].

Two common problems in the use of tracing algorithms are that imaging quality varies between labeling and imaging techniques, and that the growing list of available algorithms complicates the testing of their suitability for specific tasks in a systematic and fast manner. Similarly, algorithm developers lack a standard set of images for benchmarking. The DIADEM challenge (https://diadem.janelia.org/history.html[29]) is an example of successful standardized benchmarking. However, the diversity of datasets tested is limited when studying the relevance of image quality features, and since that challenge several algorithms have been developed[28].

We devised the BigNeuron project to address these challenges and advance toward a consensus on how to use and improve automatic neuron tracing tools[30]. The results presented here summarize the goals reached through its completion, including the gathering and sharing of a community-contributed, diverse and extensive set of 3D neuron imaging datasets, the provision of gold standard annotations for a selected subset of images to be used as reference for bench-testing, the organization of collaborative events for the development of automatic tracing algorithms, the provision of a platform for benchmarking algorithms against gold standard reconstructions, the integration of the obtained knowledge to improve the accessibility, accuracy and efficiency of automatic reconstruction methods, and last, the provision of a tool to suggest the most suitable automatic tracing algorithm in external datasets based on our results.

Results

An open bench-testing platform for neuron tracing

The BigNeuron project utilizes neuron image stacks from different species (including fruitfly and other insects, fish, turtle, chicken, mouse, rat and human) and nervous system regions such as cortical and subcortical areas, the retina and peripheral nervous system. The data include multiple light microscopy modalities, especially laser scanning microscopy (confocal or 2-photon) and brightfield or epi-fluorescent imaging. The neurons are labeled using different methods, such as genetic labeling and virus, dye or biocytin injection, and span a broad range of types (for example, unipolar, multipolar, release of different neurotransmitters, and with a wide variety of electrophysiological properties). Many of these image volumes were generated by large-scale neuroinformatics projects such as the Allen Mouse and Human Cell Types projects (http://celltypes.brain-map.org/), Taiwan FlyCircuit (15,921 image volumes; http://www.flycircuit.tw/ ref. [13]) and Janelia FlyLight (13,449 image volumes; https://www.janelia.org/project-team/flylight), but several datasets are also contributed directly by neuroscientists worldwide. In total, we gathered approximately 30,000 single-neuron 3D image volumes and used them for bench-testing of automated tracing algorithms, and generated 1.4 million tracing results. To generate a representative dataset of various organisms, cell types and imaging conditions, we randomly selected a small subset from the datasets of the large-scale projects. Practically, it was not feasible to provide gold standard annotations for all 30,000 volumes. Thus, we selected 166 neurons for the generation of a diverse set of manually curated gold standard reconstructions in annotation workshops and for posterior benchmarking of automatic tracing algorithms, named as the Gold166 dataset (Supplementary Table 1). The resulting dataset exceeds previous benchmarking studies in both number and diversity.

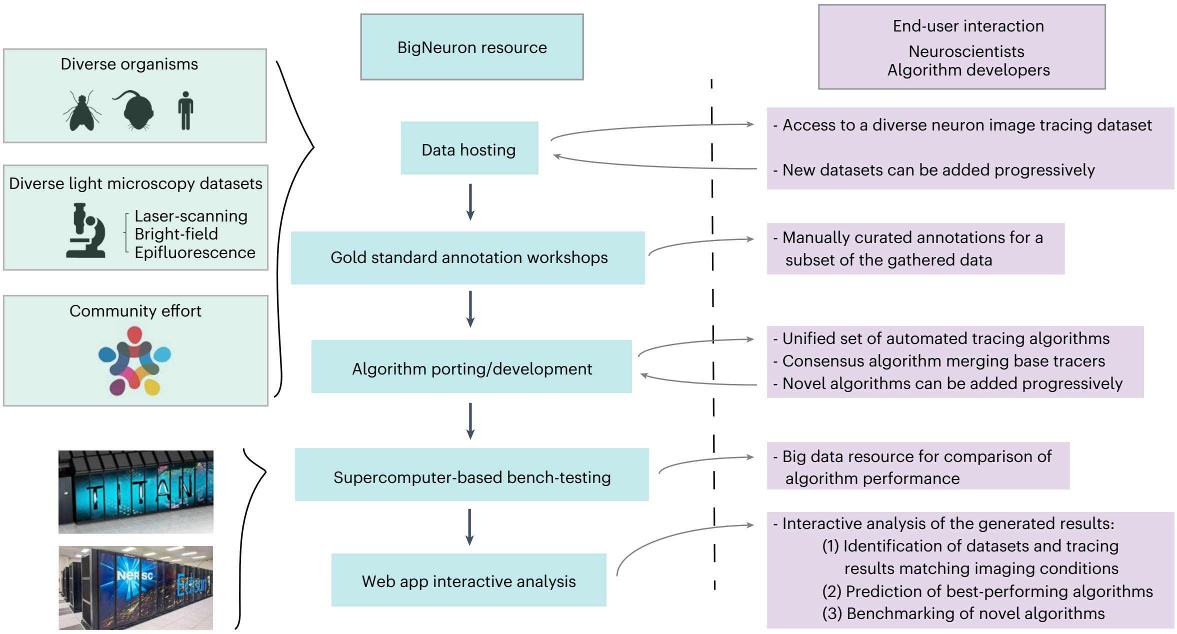
We aimed to bench-test automated single-neuron tracing algorithms on a common open platform: Vaa3D (RRID:SCR_002609). We held a total of 11 community hackathons and events in the first phase. As a result, 16 automatic tracing algorithms were developed (Supplementary Table 2). In the bench-testing phase, all ported algorithms were tested on the Gold166 dataset using TITAN at Oak Ridge National Laboratory (United States), as well as supercomputers at Lawrence Berkeley National Laboratory (United States) and Human Brain Project (Europe).

This community effort (Fig. 1) yielded several notable outcomes:

- In a series of hackathons worldwide, developers learned (from each other) the relative pros and cons of various methods and how to leverage existing resources to refine or develop tracing algorithms.
- The project has served as a practical guideline for neurobiologists in determining the suitability of specific tracing methods for a variety of image datasets and providing feedback regarding the utility of various sample preparation and imaging protocols.
- Bringing neuron tracing methods and results together encouraged method developers to collaborate, share and reuse each other's software modules.
- We developed a community-derived phenotype databases for single neurons, cataloging neuron shape and projection patterns from different species and different brain regions, and offering an opportunity to mine and query the patterns of neurons with distinct shapes.
- We provide end-users with a tool to predict the accuracy of automated tracing algorithms using as input only the image quality features, and a set of automated tracing results, without the need for manually annotated data.

A web app to navigate heterogeneous bench-testing results

To enable user-defined interactive exploration of the data, we organized gold standard annotations, automatic reconstructions, their imaging datasets and associated metadata in R data frames. Both gold standard annotations and automated reconstructions were stored and analyzed using the Stockley–Wheal–Cole (SWC)[31] format. For bench-testing, we measured the distance between all of the automatic reconstructions and the respective gold standard annotations, and computed image quality metrics for each dataset. We pooled all of the data in an interactive web app, Shiny (https://linusmg.shinyapps.io/BigNeuron_Gold166 and https://neuroxiv.net/bigneuron/, Extended Data Fig. 1; the two links are mirrors of the same application) and the documentation can be found at https://github.com/lmanubens/BigNeuron. The app enables users to perform principal component analysis, *t*-distributed stochastic neighbor embedding (t-SNE), correlation and clustering, visualization of imaging data and reconstruction in 2D projections, and benchmarking of automatic tracing algorithms in user-defined data subsets.


BigNeuron resource
End-user interaction
Neuroscientists
Algorithm developers
Diverse organisms
Diverse light microscopy datasets
Laser-scanning
Bright-field
Epifluorescence
Community effort
Data hosting
- Access to a diverse neuron image tracing dataset
- New datasets can be added progressively
Gold standard annotation workshops
- Manually curated annotations for a subset of the gathered data
Algorithm porting/development
- Unified set of automated tracing algorithms
- Consensus algorithm merging base tracers
- Novel algorithms can be added progressively
Supercomputer-based bench-testing
- Big data resource for comparison of algorithm performance
Web app interactive analysis
- Interactive analysis of the generated results:
(1) Identification of datasets and tracing results matching imaging conditions
(2) Prediction of best-performing algorithms
(3) Benchmarking of novel algorithms
TITAN
NERSC

Q13

Fig. 1 | Overview of the BigNeuron project and how the community can interact with it. BigNeuron has gathered tens of thousands of 3D neuronal images from a diverse set of experimental preparations. Through the effort of the open-source community, and using powerful supercomputers, we unified a broad range of tracing algorithms in a common platform and bench-tested them on a manually annotated subset of the gathered data. We developed an interactive web app that enables any user to find similarities between their data and the BigNeuron dataset; it provides a method for obtaining consensus reconstructions and a tool for neuroscientists to predict the best automatic tracing algorithm in any single-neuron imaging dataset. This resource not only provides access to and interactive analysis of the obtained datasets and results, but it also aims to incorporate datasets and algorithms provided by the community in the future.

Q14

To encourage developers to develop algorithms and to simplify the benchmarking of algorithms developed in the future, we added the possibility of uploading and interactively bench-testing reconstruction results of user-defined algorithms. Developers can test their algorithms with the gold standard preprocessed imaging datasets, which can be downloaded from https://github.com/BigNeuron/Data/releases/tag/Gold166_v1. After generating single-cell reconstructions for any subset of the data, users can upload the obtained automatic reconstructions by specifying the dataset identity (ID) of each reconstruction in the filename (see the ID lookup table https://github.com/lmanubens/BigNeuron/blob/main/lookup_gold166.csv). Once uploaded, users can include the tracings in the interactive analysis and benchmarking. We invite developers and users to send algorithm implementations and imaging datasets for inclusion in the platform. All submissions will be assessed once per year and introduced in the platform. Those should conform to the guidelines described in the Open Data Agreement (Supplementary Information) of the BigNeuron project.

Variance in the Gold166 dataset

An overview of the morphological features (Supplementary Table 3) of the analyzed trees shows that there is high morphological diversity in the analyzed neurons. To quantify the heterogeneity of the data, we computed the coefficient of variation for a set of morphological features shared by the Gold166 dataset and a published neuron tracing data mining work[32] based on data from NeuroMorpho.org[33]. For seven out of nine of the neuromorphological features, the coefficient of variation of the Gold166 dataset was similar to or higher than that of the NeuroMorpho.org dataset (Fig. 2a), indicating that the diversity of the Gold166 dataset is sufficient to sample the performance of automatic tracing algorithms in heterogeneous neuron types.

Q15

To identify the features that account for variance in the datasets, we performed dimensionality reduction using principal component analysis (PCA). The first two principal components explain 43.3% of the variance in the data (Fig. 2b). Measures that account for image quality mainly contribute to principal component 1 (PC1, 24.8% explained variance): that is, the focus score computed on the SWC nodes, the percentage of minimal intensity voxels in the SWC nodes, and the standard deviation of the intensity in the image volume (Fig. 2c). Following those metrics, the median intensity in the image volume and the median intensity on the SWC nodes make considerable contributions to PC1. Neuromorphological metrics mainly contribute to PC2 (18.5% explained variance): that is, the total length, followed by the maximum path distance, the maximum branch order, and the number of tips of the trees (Fig. 2d). The focus score and contrast-to-noise ratio in SWC nodes contribute to PC2 to a lower extent.

A 2D projection overview of those principal components shows that the datasets are clustered by the laboratories that provided them. Nevertheless, neurons from the same organism obtained by different laboratories tend to cluster together (Fig. 2b).

Even though a systematic comparison between species should be done for similar cell types, we nevertheless present here a comparison of broad neuromorphological differences to provide context information for the analyzed neuronal tracings. It is worth noting that neuron types of different organisms have not been matched, and that we did not choose the sample size based on this aim. Consequently, we do not claim either significant or non-significant differences between neurons from different species. That being said, human and silkmoth neurons are big and complex, and have branches that extend more than 200 µm from the soma (Fig. 2e). In the case of silkmoth neurons, this is explained by the fact that the reconstructions include long-range projections (Fig. 2e,f). Mouse neurons are similar to human cells: although they are smaller and less complex, they have a comparable ratio of branch points per unit of cable length (Fig. 2g) and cluster together with ex vivo human neurons in the PCA (Fig. 2b). By contrast, fruitfly, silkmoth, zebrafish and chicken neurons have ratios of branch points per unit of cable length higher than mammalian counterparts (Fig. 2g).

Frog neurons have branch point density values closer to human ex vivo and mouse counterparts. As shown in the PCA (Fig. 2b), fruitfly neurons differ from silkmoth neurons mainly in terms of maximum path distance and maximum branch order, while having increased average diameter and bifurcation angles compared with zebrafish and chicken Q17 neurons, while the silkmoth neurons have values closer to human, mouse and chicken neurons. Chicken neurons have high dendritic complexity at small radii (Fig. 2e), but they are closer to zebrafish and fruitfly larvae neurons in regard to the size and density of branches per unit of dendritic length (Fig. 2f,g).

To explore putative functional heterogeneity in the dataset, we quantified the centripetal bias (k, ref. 34). When the centripetal bias is 0, neurites are not preferentially radial to the soma, and their angles toward the root of the tree are distributed uniformly. As the centripetal bias k goes to infinity, all branch segments increasingly have radial directions from the soma. The distribution of branch angles to the tree root as a function of k can be expressed analytically as a modified von Mises distribution[35]. The Sholl intersection profile (SIP) of specific neuron types can be predicted by their span, total length and centripetal bias, each of which has a specific impact on neuron functionality[34]. Furthermore, the impact of centripetal bias on electrotonic compartmentalization has been demonstrated[34]. Our analysis shows that the centripetal bias of the Gold166 neurons is constrained to values lower than 2 (Fig. 2h), indicating that the analyzed neurons have low electrotonic compartmentalization and long conduction times compared with hippocampal neurons (with k ~ 7 and k ~ 12 for cornu Q18 ammonis (CA1) and dentate gyrus, respectively). All planar cells have a length-to-SIP scale ratio consistent with the theoretical 2D von Mises root angle distribution. The 3D silkmoth and fruitfly neurons have a higher dendritic occupancy, consistent with the theoretical 3D von Mises root angle distribution. In terms of cross-species comparison, we observed similar distributions in the centripetal bias of all species.

We performed unsupervised hierarchical clustering using both neuromorphological and image quality features. We found that our data contains 16 clusters and is best approximated by an EEE (ellipsoidal, equal volume, shape and orientation) model (Extended Data Fig. 2, maximum Bayesian information criterion[36] of 2989.32). The data are clustered by both dataset and organism.

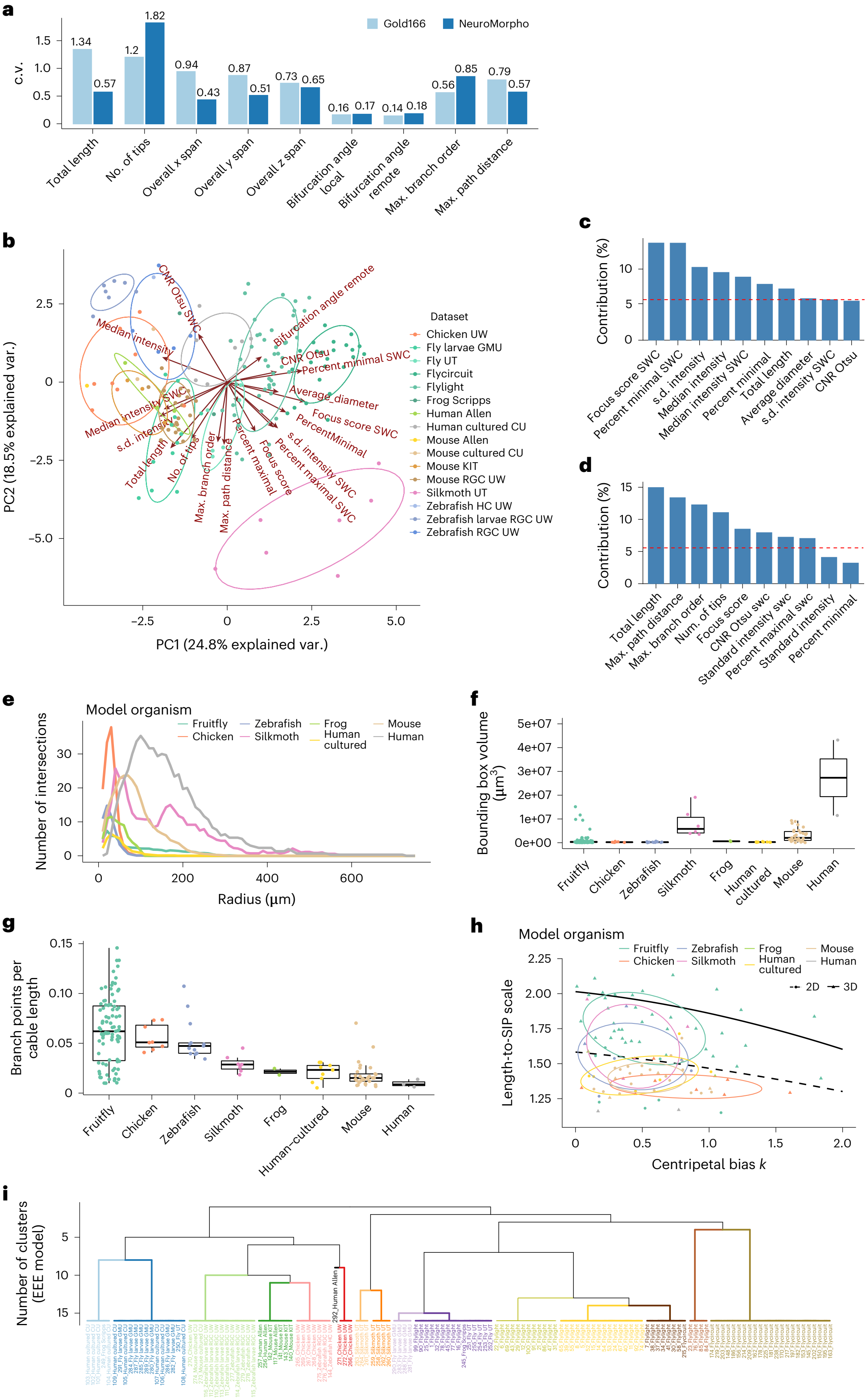
Consensus provides best estimates of neuronal structure

Given the high diversity of existing automatic tracing algorithms, it is reasonable to assume that their performance in reconstructing specific tree morphology features may vary. We tested their accuracy by measuring the error between the morphological features in the automatic reconstructions and the gold standard trees. We found that each morphological feature was best estimated by different algorithms (Fig. 3a; Kruskal–Wallis; average contraction, H(5) = 83.17, $P = 1.8 \times 10^{-16}$; average fragmentation, H(5) = 143.93, $P = 2.6 \times 10^{-29}$; bifurcation angle remote, H(5) = 34.12, $P = 2.3 \times 10^{-6}$; maximum branch order, H(5) = 60.99, $P = 7.6 \times 10^{-12}$; maximum path distance, H(5) = 10.80, $P = 0.056$; number of tips, H(5) = 68.21, $P = 2.4 \times 10^{-13}$; overall x span, H(5) = 29.46, $P = 1.9 \times 10^{-5}$; total length, H(5) = 30.22, $P = 1.3 \times 10^{-5}$). Thus, a set of diverse algorithms can contribute complementary information to approximate the ground truth with increased precision. To build on this idea, we developed an algorithm for generating consensus trees based on a set of automatic reconstructions. By clustering the closest node positions of the set of reconstructions in space, the algorithm defines a set of consensus nodes (Fig. 3b). We assigned a confidence Q19 value to each consensus node using a voting strategy that depends on their existence in iterations of each individual automatic reconstruction. Thus, nodes highly prevalent in many reconstructions are kept in the consensus tree, while false-positive fragments in small numbers of reconstructions have low confidence and are discarded. Finally, the high-confidence set of consensus nodes obtained through this process is connected in a single tree using the maximum spanning tree algorithm (Fig. 3b,c). It is worth noting that one of the algorithms used for bench-testing is based on a similar idea. The ensemble neuron tracer[37] generates distinct models of neuron tracings based on a single base tracer at a time (for example, APP2[21]). In that case, what is perturbed is the data used for tracing, and the output tracings are then merged into an as complete as possible single result. The perturbations include thresholding, dilation and closing of foreground image pixels. In ensemble neuron tracer the final result is obtained by merging the diverse results into an as complete as possible single tree: there is no consensus node voting strategy nor are diverse base tracers combined. By contrast, the consensus algorithm combines the tracing results (SWC files) of diverse base tracer algorithms based on an iterative voting strategy to select and connect consensus nodes, without any perturbation of the input images. Based on the bench-testing of all algorithms, we ranked the best-performing algorithms. Consensus (best in 40 of the Gold166 images), SmartTracing (20), neuTube (19) and axis analyzer (15) are the algorithms that perform best in most images in the dataset (Extended Data Fig. 3a). We also obtained an overall benchmark using an aggregated distance metric (Extended Data Fig. 3b), which also showed that Consensus was the most accurate overall, followed by nctuTW_GD and neuTube. The bench-testing results also showed that, in 5 of 16 clusters identified in the data (Fig. 2i), the consensus tree algorithm provided the most accurate approximation to gold standard trees (Fig. 3c,d). However, in the other clusters, different algorithms outperformed the consensus tree strategy (Fig. 3e). This suggests that the consensus approach introduced here usually outperforms others when the datasets are noisy. However, when imaging conditions are not as challenging, neuTube may be optimal for intermediate contrast-to-noise ratios, and SmartTracing and Axis Analyzer seem to perform best in high contrast-to-noise ratio images (Extended Data Fig. 4).

Fig. 2 | Variance in the datasets is explained by both image quality and tree morphology features. **a**, Bar plot showing the coefficient of variation (c.v.) of various morphological features in the Gold166 dataset and in a dataset with 5,099 neurons from NeuroMorpho.org[32]. **b**, PCA of gold standard neuron reconstructions and their image stack quality metrics. Each point is one gold standard annotation, and the color indicates the dataset it comes from. Arrows represent the direction of each variable in the PCA space. Longer arrows belong to variables that are well represented by the two principal components. 68% confidence normal data ellipses for each group are drawn with solid lines. CNR, contrast-to-noise ratio; CU, XXX; GMU, XXX; HC, XXX; KIT, XXX; RGC, Q16 retinal ganglion cell; UT, XXX; UW, XXX. **c**,**d**, Bar plots showing the percentage of explained variance for PC1 (**c**) and PC2 (**d**). The red dashed lines indicate the expected average contribution. **e**, Sholl analysis of the neurons in the Gold166 dataset. Each line expresses the average number of intersections quantified at a given distance from the soma for the neurons of a given model organism (color-coded). **f**,**g**, Box plots and dot plots show the bounding box volume (number of neurons for each organism: fruitfly, $n = 91$; chicken, $n = 8$; zebrafish, $n = 13$; silkmoth, $n = 7$; frog, $n = 2$; human cultured, $n = 9$; mouse, $n = 29$; human=3) (**f**) and density of branch points per unit of neurite length (**g**) of the reconstructed cells for each model organism (color-coded). The center line is the median value, the ends of the box are Q1 and Q3, and the whiskers add 1.5-fold the interquartile range. Colored dot points are overlaid for all measures. **h**, Proportionality constant between the total length and the Sholl intersection profile (SIP) as a function of the centripetal bias k for planar 2D neurons (dashed black line) and 3D neurons (solid black line). Values for each of the cells in the Gold166 dataset are shown as solid dots color-coded by model organisms. Planar trees are shown as circles and 3D trees as triangles. **i**, Hierarchical clustering using both neuromorphological and image quality features. Colors indicate different clusters obtained with a Gaussian mixture ellipsoidal, equal volume, shape, and orientation (EEE) model. Labels indicate the neuron dataset ID and the organism they belong to.

Predicting best algorithm performance

Image quality metrics correlate with the accuracy of automated tracing (Extended Data Fig. 5), suggesting that specific tree morphology features, together with image quality metrics, can be informative in the choice of the most accurate algorithm. We used a support vector machine that, given an image volume and a set of automatic



reconstructions, predicts the accuracy of each tested algorithm based on a regression of the percentage of difference between the automatic reconstructions and gold standards (Fig. 4). By training on 85% of our data, we obtained a regression that enables prediction of the percentage of difference between the automatic reconstructions and gold standards (coefficient of determination of 0.637; Fig. 4a). Learning curves with an increasing percentage of data used for training (evaluation on a hold-out set of images only) show that the regression quality increases by the percentage of the training set with slight improvement once more than 30% of the data is used for training (Fig. 4b), suggesting that this strategy will generalize well when it is used in unknown neuron reconstruction datasets. We assessed the quality of the predictions by comparing the percentage of difference for all of the automatic reconstructions, the true best algorithms and the predicted best algorithms for each dataset. We found that both the known best algorithms and the predictions in most of the cases performed better than the algorithms chosen by chance (Fig. 4c,e; the percentage of difference distribution is positively skewed, medians of 27%, 35% and 54% for true best, predicted, and all algorithms, respectively; Wilcoxon test, all algorithms versus true best, $V = 7{,}319$, $P = 1.8 \times 10^{-6}$, n1 = 489, n2 = 18; all algorithms versus predicted best, $V = 6{,}389$, $P = 0.0011$, n1 = 489, n2 = 18; predicted best versus predicted worst, $P = 309$, $P = 3.2 \times 10^{-6}$, n1 = 18, n2 = 18; true best versus predicted best, $V = 111.5$, $P = 0.11$). Similarly, algorithms predicted to be worse had a negatively skewed distribution in the percentage of difference to gold standards (Fig. 4c; median of 99%). This analysis highlights that for a few neuron datasets, none of the automatic reconstruction algorithms was able to recapitulate the gold standard annotations. This was the case, for example, for chicken cells, which had a particularly low signal-to-noise ratio in their neurites and have specific distinctive morphological aspects (such as an increased soma size, high branch density and high centripetal bias).

Showcase of best algorithm prediction in fMOST data

Community efforts such as the BRAIN Initiative Cell Census Network (BICCN[38], https://biccn.org/) are using fluorescence micro-optical sectioning tomography (fMOST) to map neuron architecture in whole mouse brains. To illustrate the value of BigNeuron in this community, we used the support vector machine regression model trained with the Gold166 dataset to predict best-performing algorithms in fMOST image volumes. The support vector machine model provided predicted values for the percentage of difference to gold standard trees (not available for those images). The algorithms predicted to provide the best results were neuTube and Consensus (Fig. 5a,e). Visual inspection of the reconstructions suggests that the regression model provides a reasonable approximation of the best automatic algorithm selection (Fig. 5b–d,f–h). A comparison of image quality features between the Gold166 and fMOST images showed that the most similar dataset to fMOST was the zebrafish larvae retinal ganglion cell dataset (Extended Data Fig. 6). The benchmarking of automatic reconstruction algorithms in the images of this set is consistent with the predictions. NeuTube and Consensus algorithms are among the best approximations to gold standard reconstructions. Although these results suggest that the support vector machine-based predictor is a useful tool to select best-performing algorithms, they also highlight that the predicted values may underestimate algorithm accuracy.

Discussion

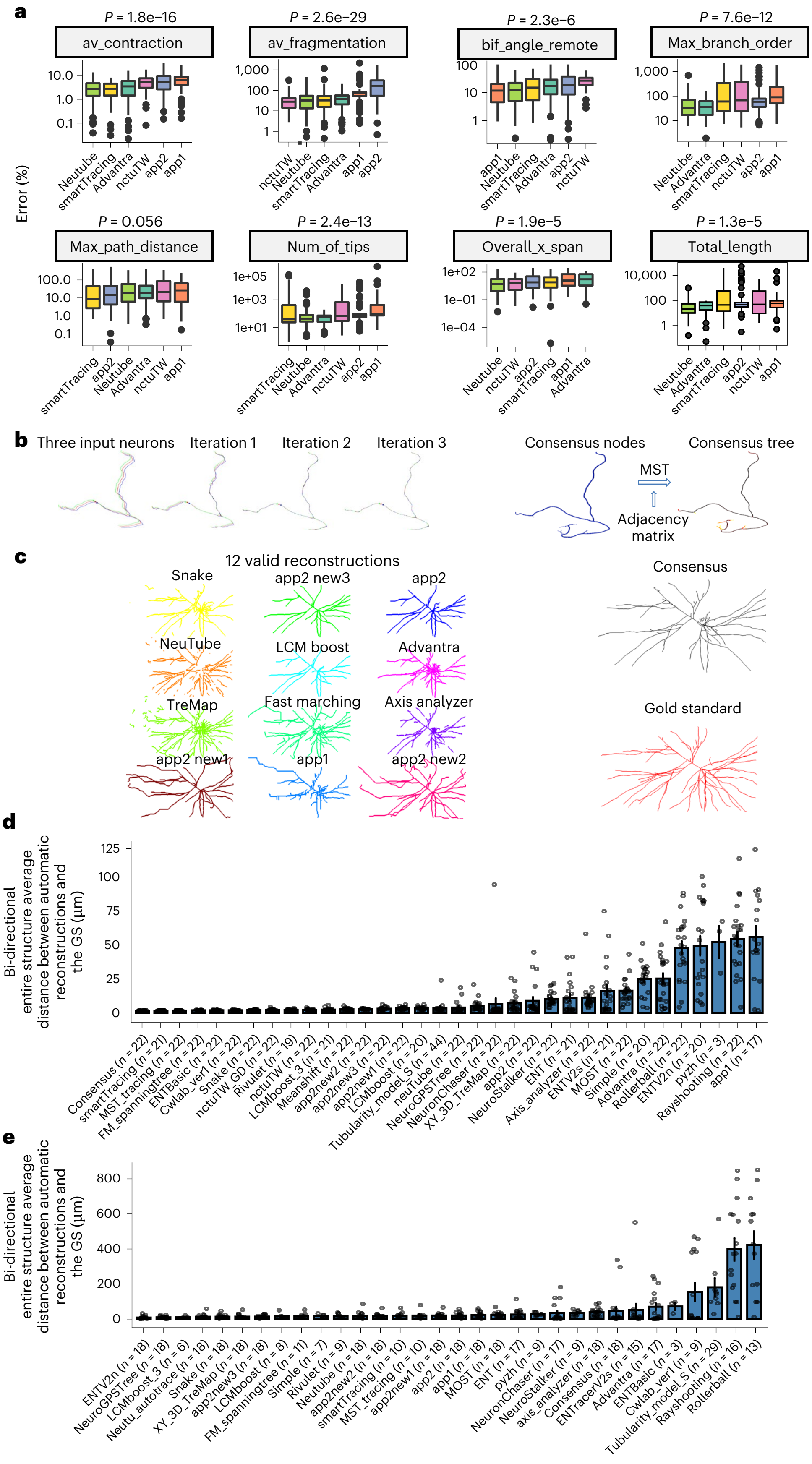
How to accurately and efficiently perform quantification of neuronal morphology across diverse types of neuronal images in an unbiased fashion has been a long-standing challenge, which has become even more acute with the introduction of exciting technologies that generate 3D, complete neuronal morphology datasets at high speed[8,11,39,40]. Moreover, large-scale brain science projects such as the BRAIN Initiative in the United Statess[41], Europe's Human Brain Project[42] and the Allen Cell Types Database (http://celltypes.brain-map.org/) integrate reliable automatic reconstruction into their production pipelines to enable fast quantification of the structure of hundreds of thousands of neurons. The BigNeuron project has contributed to improving existing automatic tracing standards, focusing on the interaction and discussion between neuroscientists and developers, and integrating data generation and development efforts from several laboratories. BigNeuron takes an innovative approach by including events for discussion and identification of major issues, and coordinated hackathons for the development of tools to overcome them.

The project has inspired technical approaches to measure and compare the similarity of neuronal trees[43,44]. The experience has highlighted major challenges in the field of neuron annotation and classification[45,46], and contributed to justify methods for imaging[47,48] and large-scale annotation of neurons[24,25,49]. BigNeuron enables developers to interactively benchmark their tools against the methods assessed in this work, and to introduce datasets and algorithms in the benchmarking set upon request. This collaborative standardized approach is notable because the problem of choosing a method to perform automatic reconstruction in project-specific datasets is an arduous task. The resource could also be adapted to use alternative tree edit distance definitions[50,51]. In our web app we also enable users to benchmark algorithms with an aggregated metric. However, we note that aggregated metrics should be interpreted with care, given that the ranking positions of teams participating in challenges change when the aggregation method is changed[52]. We also note that we focused on the challenge of tracing single neurons. However, this approach does not preclude the use of the presented tools in images with multiple neurons[53].

The diversity in the data that we analyzed enabled us to gain relevant insights into the future steps needed to improve automatic reconstruction tools. Based on our results, image quality should be taken into account in the process of choosing the most suitable tool for a specific dataset. Another source of variance could be the disagreement between annotations obtained by different persons. Exploring this problem was beyond the scope of this work. However, a study motivated by BigNeuron explored the variability of human annotators, showing that the differences are small[54]. Taking into account this complementary result, we consider that the solution to this problem is to improve the observability of complicated 3D neuron arborizations with the best possible multi-dimensional annotation tools to minimize the variation of human annotation.

Fig. 3 | Consensus provides best estimates of neuronal structure. **a**, Errors for a set of morphological metrics in a random subset of automatic tracing algorithms. The error was computed as the difference between the metrics obtained in the gold standard and the automatic reconstructions. Errors for the automatic reconstructions obtained from all image volumes with a given algorithm are shown as box plots. The center line is the median value, the box is bounded by Q1 and Q3, the whiskers add 1.5-fold the interquartile range, and outliers are indicated with points. *P* values were obtained using the Kruskal–Wallis (one-sided) test. Number of reconstructions for each method: Advantra, $n = 109$; app1, $n = 98$; app2, $n = 102$; nctuTW, $n = 54$; neuTube, $n = 112$; smartTracing, $n = 86$. **b**, Overview of the development of an algorithm to generate consensus trees. The algorithm first performs iterative match and center: for each node in each tree, it identifies the nearest corresponding location among input neurons, and shifts to the mean location. Nodes from all of the input neurons are merged to form consensus nodes, and reliability weights for the consensus nodes are established by collecting votes for the connections from individual input neuron trees. A maximum spanning tree (MST) algorithm is used to connect consensus nodes to form the consensus tree. **c**, An example of the results obtained with the consensus tree algorithm using a set of 12 automatic reconstructions. **d**,**e**, Bi-directional entire structure average distance between automatic reconstructions and the gold standard (GS) of the dataset cluster number 9 (Fig. 2i) (**d**) and cluster number 8 (Fig. 2i) (**e**). The mean ± s.e. is shown as a bar plot.

The information obtained through the pooling of datasets and the benchmarking of reconstruction tools can be used to generate knowledge-based improved algorithms. Previous work has used the BigNeuron datasets to benchmark tracing algorithms[55–58] and to define neuromorphological features for the analysis of dendritic trees[44]. However, to our knowledge there is only one study that explores the idea of



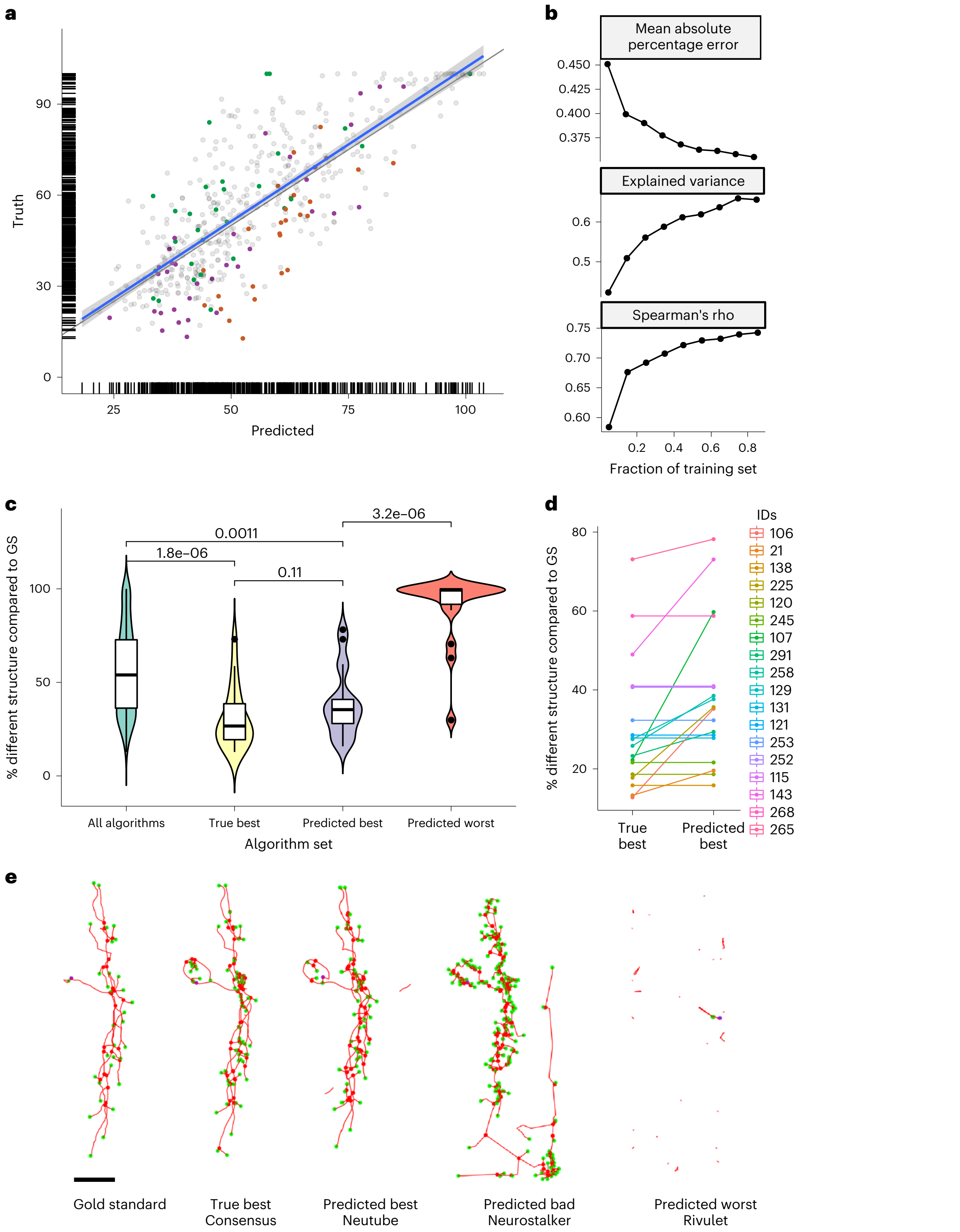


Fig. 4 | Support vector machine regression for predicting best algorithm performance. **a**, Support vector machine regression of the percentage of difference between automatic reconstructions and the gold standard. Each point represents the true or predicted percentage of difference for an automatic reconstruction. Colored points show subsets of results for individual image volumes. The rest of the data in the testing set are represented in gray. The gray line shows the function $y = x$, and the blue line shows a linear model fit of the data, with the 95% confidence level for predictions in light gray. **b**, Learning curves showing the quality of the regression as a function of the percentage of data used for training. Each point is the average of 5 repetitions using different random subsets of the data for training. **c**, Percentage of difference between automatic reconstructions and gold standard (GS) annotations represented as an overlay of box and violin plots. The center line is the median value, the box is bounded by Q1 and Q3, and the whiskers add 1.5-fold the interquartile range. Colored dot points are overlaid for all measures. Statistical comparisons between groups were performed using the Wilcoxon test (two-sided). The number of reconstructions in each group are: All algorithms, $n = 489$; True best, $n = 18$; Predicted best, $n = 18$; Predicted worst, $n = 18$. **d**, Pair plot showing the percentage of difference between automatic reconstructions and gold standard annotations for the true best algorithms and the predicted best algorithms. Colors indicate different neuron datasets in the testing set. **e**, Representative images of the gold standard annotation, the true best automatic reconstruction algorithm, and various predictions using the regression results. Red lines represent neuron tree branches, blue dots indicate the root of the trees, red dots indicate branch points, and green dots indicate terminal points. Scale bar, 100 µm.

Q23

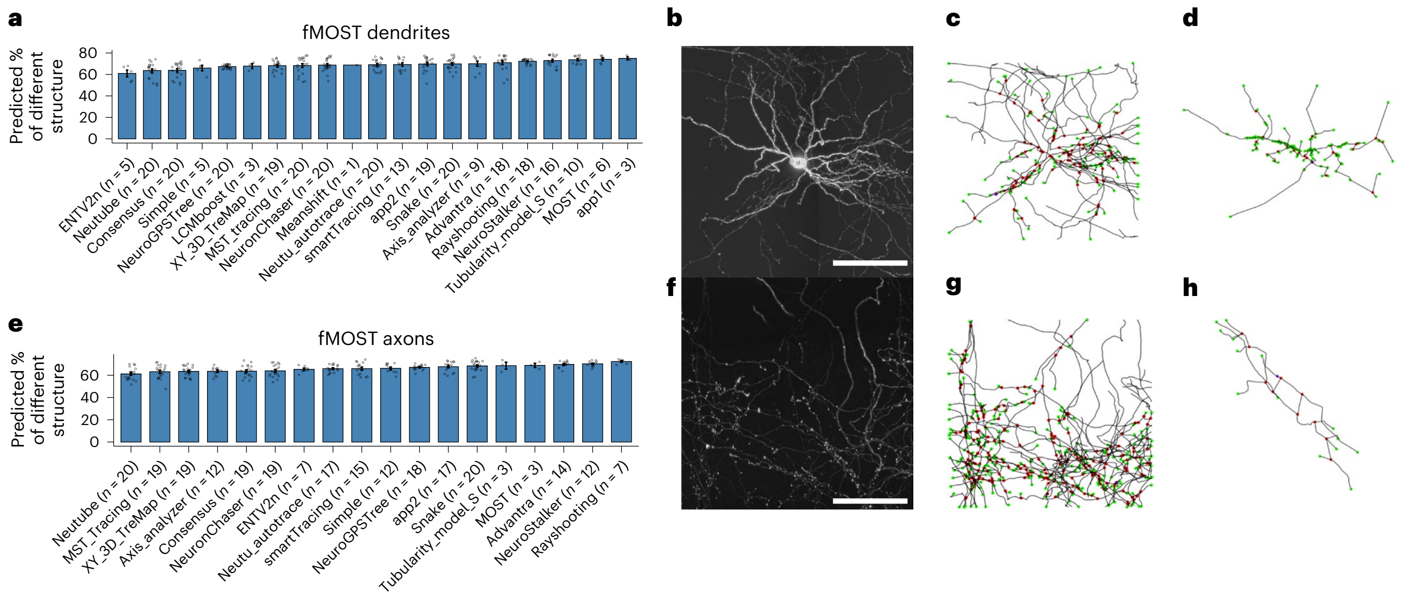


Fig. 5 | Showcase of best algorithm prediction in fMOST data. Comparison of results obtained for fMOST dendrite and axon image blocks, including the predicted percentage of different structure and representative visualizations. **a**, Predicted percentage of different structure for automatic reconstructions of fMOST dendrite image blocks. **b**, Maximum intensity projection of a representative fMOST dendrite image block. **c**,**d**, 2D projections of the neuTube (**c**) and NeuroStalker (**d**) algorithm automatic reconstructions in the image volume in **b**, predicted to provide the best and the lowest accuracy results, respectively. **e**, Predicted percentage of different structure for automatic reconstructions of fMOST axon image blocks. **f**, Maximum intensity projection of a representative fMOST axon image block. **g**,**h**, 2D projections of the neuTube (**g**) and Advantra (**h**) algorithm automatic reconstructions in the image volume represented in **f**, predicted to provide the best and the lowest accuracy results, respectively. Scale bars, 100 µm.

Q26

Q27

using an ensemble learning algorithm to generate improved automatic reconstruction results[37], and this learning algorithm has been included in our analysis. Our results with the consensus tree algorithm highlight the potential of such an approach, and provide best estimates of the ground truth neuron structure in most of the datasets we analyzed.

Q29

While the ensemble neuron tracer and our consensus tree algorithms implicitly rely on the identification and discarding of unrealistic features in automatically generated trees, our datasets and web app enable interactive exploration of the relationship between reconstruction quality and the features of the analyzed trees. Our results indicate that image features (for example, focus score) are the main correlates of reconstruction quality, and we have also identified correlations with neuromorphological features such as the parent–daughter ratio. Those observations can inform future algorithm development by adding morphological constraints to the generated trees. Even though we have generated data for a few variations in parameter sets for some of the methods, in this work we have focused on comparison between the base methods. We think that future efforts could focus on the iterative search of optimal parameter sets by applying optimization algorithms with the maximization of reconstruction quality as the objective function.

In most cases our regression method to estimate best algorithm performance provides accurate estimates of the automatic reconstruction quality. However, there is a small number of datasets for which there is a substantial difference between the true and the predicted best algorithms. Thus, this tool should be used with caution. Interestingly, in a few cases, none of the automatic reconstruction methods was able to produce any result close to the ground truth (>50% of different structure). This happened in datasets with extremely low signal-to-noise ratios and overall bad imaging quality. Quantitative analysis of those features can help users to identify objective thresholds for image quality features below which datasets should be discarded. Recent projects are generating unprecedented volumes of gold standard reconstructions (for example, ref. 25). We expect that application of the methods presented in standardized, larger gold standard datasets will provide a valuable tool toward full automation of neuron tracing in specific combinations of microscopic imaging and labeling techniques that are becoming standards in the field.

We believe BigNeuron is a valuable resource, given that the gold standard manual annotations can be used (and have already been used in a number of works[55–58]) for standardized benchmarking. We also provide here an example of how the generated bench-testing data can be used to learn how diverse algorithms can be the best performer in specific case scenarios. Researchers can also explore these data interactively with our web app, and identify which Gold166 images are most similar to theirs, enabling them to select the corresponding most effective algorithms. The set of automatic tracing algorithms unified in Vaa3D also enables neuroscientists to test them easily, while both the consensus tree algorithm and the support vector machine predictor can be introduced into their pipelines to leverage the results of a bench-testing phase when facing the challenge of tracing their datasets.

Q30

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-023-01848-5.

References

1. Meijering, E. Neuron tracing in perspective. *Cytometry A* **77**, 693–704 (2010).
2. Parekh, R. & Ascoli, G. A. Neuronal morphology goes digital: a research hub for cellular and system neuroscience. *Neuron* **77**, 1017–1038 (2013).
3. Capowski, J. J. An automatic neuron reconstruction system. *J. Neurosci. Methods* **8**, 353–364 (1983).

4. Senft, S. L. A brief history of neuronal reconstruction. *Neuroinformatics* **9**, 119–128 (2011).
5. Cai, D., Cohen, K. B., Luo, T., Lichtman, J. W. & Sanes, J. R. Improved tools for the Brainbow toolbox. *Nat. Methods* **10**, 540–547 (2013).
6. Nern, A., Pfeiffer, B. D. & Rubin, G. M. Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proc. Natl Acad. Sci. USA* **112**, E2967–E2976 (2015).
7. Daigle, T. L. et al. A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. *Cell* **174**, 465–480 (2018).
8. Chung, K. & Deisseroth, K. CLARITY for mapping the nervous system. *Nat. Methods* **10**, 508–513 (2013).
9. Hama, H. et al. Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain. *Nat. Neurosci.* **14**, 1481–1488 (2011).
10. Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J. & Stelzer, E. H. K. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science* **305**, 1007–1009 (2004).
11. Verveer, P. J. et al. High-resolution three-dimensional imaging of large specimens with light sheet-based microscopy. *Nat. Methods* **4**, 311–313 (2007).
12. Li, A. et al. Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science* **330**, 1404–1408 (2010).
13. Chiang, A.-S. et al. Three-dimensional reconstruction of brain-wide wiring networks in *Drosophila* at single-cell resolution. *Curr. Biol.* **21**, 1–11 (2011).
14. Meissner, G. W. et al. A searchable image resource of *Drosophila* GAL4-driver expression patterns with single neuron resolution. *eLife* **12**, e80660 (2023).
15. Markram, H. The blue brain project. *Nat. Rev. Neurosci.* **7**, 153–160 (2006).
16. Ecker, J. R. et al. The BRAIN Initiative Cell Census Consortium: lessons learned toward generating a comprehensive brain cell atlas. *Neuron* **96**, 542–557 (2017).
17. Stockton, D. B. & Santamaria, F. Integrating the Allen Brain Institute Cell Types Database into automated neuroscience workflow. *Neuroinformatics* **15**, 333–342 (2017).
18. Yamasaki, T., Isokawa, T., Matsui, N., Ikeno, H. & Kanzaki, R. Reconstruction and simulation for three-dimensional morphological structure of insect neurons. *Neurocomputing* **69**, 1043–1047 (2006).
19. Wang, Y., Narayanaswamy, A., Tsai, C.-L. & Roysam, B. A broadly applicable 3-D neuron tracing method based on open-curve snake. *Neuroinformatics* **9**, 193–217 (2011).
20. Zhao, T. et al. Automated reconstruction of neuronal morphology based on local geometrical and global structural models. *Neuroinformatics* **9**, 247–261 (2011).
21. Xiao, H. & Peng, H. APP2: automatic tracing of 3D neuron morphology based on hierarchical pruning of a gray-weighted image distance-tree. *Bioinformatics* **29**, 1448–1454 (2013).
22. Santamaría-Pang, A., Hernandez-Herrera, P., Papadakis, M., Saggau, P. & Kakadiaris, I. A. Automatic morphological reconstruction of neurons from multiphoton and confocal microscopy images using 3D tubular models. *Neuroinformatics* **13**, 297–320 (2015).
23. Peng, H. et al. Automatic tracing of ultra-volumes of neuronal images. *Nat. Methods* **14**, 332–333 (2017).
24. Winnubst, J. et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell* **179**, 268–281 (2019).
25. Peng, H. et al. Morphological diversity of single neurons in molecularly defined cell types. *Nature* **598**, 174–181 (2021).
26. Zhong, Q. et al. High-definition imaging using line-illumination modulation microscopy. *Nat. Methods* **18**, 309–315 (2021).
27. Shillcock, J. C., Hawrylycz, M., Hill, S. & Peng, H. Reconstructing the brain: from image stacks to neuron synthesis. *Brain Inform.* **3**, 205–209 (2016).
28. Peng, H., Meijering, E. & Ascoli, G. A. From DIADEM to BigNeuron. *Neuroinformatics* **13**, 259–260 (2015).
29. Gillette, T. A., Brown, K. M. & Ascoli, G. A. The DIADEM metric: comparing multiple reconstructions of the same neuron. *Neuroinformatics* **9**, 233–245 (2011).
30. Peng, H. et al. BigNeuron: large-scale 3D neuron reconstruction from optical microscopy images. *Neuron* **87**, 252–256 (2015).
31. Stockley, E. W., Cole, H. M., Brown, A. D. & Wheal, H. V. A system for quantitative morphological measurement and electronic modelling of neurons: three-dimensional reconstruction. *J. Neurosci. Methods* **47**, 39–51 (1993).
32. Polavaram, S., Gillette, T. A., Parekh, R. & Ascoli, G. A. Statistical analysis and data mining of digital reconstructions of dendritic morphologies. *Front. Neuroanat.* **8**, 138 (2014).
33. Akram, M. A., Nanda, S., Maraver, P., Armañanzas, R. & Ascoli, G. A. An open repository for single-cell reconstructions of the brain forest. *Sci. Data* **5**, 180006 (2018).
34. Bird, A. D. & Cuntz, H. Dissecting Sholl analysis into its functional components. *Cell Rep.* **27**, 3081–3096 (2019).
35. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical Distributions* (Wiley Hoboken, 2011).
36. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
37. Wang, C.-W., Lee, Y.-C., Pradana, H., Zhou, Z. & Peng, H. Ensemble neuron tracer for 3D neuron reconstruction. *Neuroinformatics* **15**, 185–198 (2017).
38. Muñoz-Castañeda, R. et al. Cellular anatomy of the mouse primary motor cortex. *Nature* **598**, 159–166 (2021).
39. Zheng, T. et al. Visualization of brain circuits using two-photon fluorescence micro-optical sectioning tomography. *Opt. Express* **21**, 9839–9850 (2013).
40. Jiang, S. et al. Petabyte-scale multi-morphometry of single neurons for whole brains. *Neuroinformatics* **20**, 525–536 (2022).
41. Alivisatos, A. P. et al. The Brain Activity Map Project and the challenge of functional connectomics. *Neuron* **74**, 970–974 (2012).
42. Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R. & Koch, C. Neuroscience thinks big (and collaboratively). *Nat. Rev. Neurosci.* **14**, 659–664 (2013).
43. Costa, M., Manton, J. D., Ostrovsky, A. D., Prohaska, S. & Jefferis, G. NBLAST: rapid, sensitive comparison of neuronal structure and construction of neuron family databases. *Neuron* **91**, 293–311 (2016).
44. Kanari, L. et al. A topological representation of branching neuronal morphologies. *Neuroinformatics* **16**, 3–13 (2018).
45. Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. & Olivo-Marin, J.-C. Imagining the future of bioimage analysis. *Nat. Biotechnol.* **34**, 1250–1255 (2016).
46. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
47. Treweek, J. B. et al. Whole-body tissue stabilization and selective extractions via tissue-hydrogel hybrids for high-resolution intact circuit mapping and phenotyping. *Nat. Protoc.* **10**, 1860–1896 (2015).
48. Ke, M.-T. et al. Super-resolution mapping of neuronal circuitry with an index-optimized clearing agent. *Cell Rep.* **14**, 2718–2732 (2016).
49. Gong, H. et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat. Commun.* **7**, 12142 (2016).

50. Li, Y., Wang, D., Ascoli, G. A., Mitra, P. & Wang, Y. Metrics for comparing neuronal tree shapes based on persistent homology. *PLoS One* **12**, e0182184 (2017).
51. Ljungquist, B., Akram, M. A. & Ascoli, G. A. Large scale similarity search across digital reconstructions of neural morphology. *Neurosci. Res.* **181**, 39–45 (2022).
52. Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 5217 (2018).
53. Li, R. et al. Precise segmentation of densely interweaving neuron clusters using G-Cut. *Nat. Commun.* **10**, 1549 (2019).
54. Wang, Y. et al. TeraVR empowers precise reconstruction of complete 3-D neuronal morphology in the whole brain. *Nat. Commun.* **10**, 3474 (2019).
55. Li, R., Zeng, T., Peng, H. & Ji, S. Deep learning segmentation of optical microscopy images improves 3-D neuron reconstruction. *IEEE Trans. Med. Imaging* **36**, 1533–1541 (2017).
56. Liu, S., Zhang, D., Song, Y., Peng, H. & Cai, W. Automated 3-D neuron tracing with precise branch erasing and confidence controlled back tracking. *IEEE Trans. Med. Imaging* **37**, 2441–2452 (2018).
57. Gu, L. et al. Semi-supervised learning in medical images through graph-embedded random forest. *Front. Neuroinform.* **14**, 601829 (2020).
58. Radojević, M. & Meijering, E. Automated neuron reconstruction from 3D fluorescence microscopy images using sequential Monte Carlo estimation. *Neuroinformatics* **17**, 423–442 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

Linus Manubens-Gil[1,64], Zhi Zhou[2], Hanbo Chen[3], Arvind Ramanathan[4], Xiaoxiao Liu[5], Yufeng Liu[1], Alessandro Bria[6], Todd Gillette[7], Zongcai Ruan[1], Jian Yang[8,9], Miroslav Radojević[10], Ting Zhao[11], Li Cheng[12], Lei Qu[1,13], Siqi Liu[14], Kristofer E. Bouchard[15,16], Lin Gu[17,18], Weidong Cai[19], Shuiwang Ji[20], Badrinath Roysam[21], Ching-Wei Wang[22], Hongchuan Yu[23], Amos Sironi[24], Daniel Maxim Iascone[25,26], Jie Zhou[27], Erhan Bas[28], Eduardo Conde-Sousa[29,30], Paulo Aguiar[29], Xiang Li[31], Yujie Li[32,33], Sumit Nanda[7], Yuan Wang[34], Leila Muresan[35], Pascal Fua[36], Bing Ye[37], Hai-yan He[38], Jochen F. Staiger[39], Manuel Peter[40], Daniel N. Cox[41], Michel Simonneau[42], Marcel Oberlaender[43], Gregory Jefferis[11,44,45], Kei Ito[11,46,47], Paloma Gonzalez-Bellido[48], Jinhyun Kim[49], Edwin Rubel[50], Hollis T. Cline[51], Hongkui Zeng[32], Aljoscha Nern[11], Ann-Shyn Chiang[52], Jianhua Yao[53], Jane Roskams[32,54], Rick Livesey[55], Janine Stevens[11], Tianming Liu[33], Chinh Dang[50], Yike Guo[56], Ning Zhong[8,9,57], Georgia Tourassi[58], Sean Hill[59,60,61,62], Michael Hawrylycz[32], Christof Koch[32], Erik Meijering[63] ✉, Giorgio A. Ascoli[7] ✉ & Hanchuan Peng[1,64] ✉

[1]Institute for Brain and Intelligence, Southeast University, Nanjing, China. [2]Microsoft Corporation, Redmond, WA, USA. [3]Tencent AI Lab, Bellevue, WA, USA. [4]Computing, Environment and Life Sciences Directorate, Argonne National Laboratory, Lemont, IL, USA. [5]Kaya Medical, Seattle, WA, USA. [6]University of Cassino and Southern Lazio, Cassino, Italy. [7]Center for Neural Informatics, Structures and Plasticity, Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA, USA. [8]Faculty of Information Technology, Beijing University of Technology, Beijing, China. [9]Beijing International Collaboration Base on Brain Informatics and Wisdom Services, Beijing, China. [10]Nuctech Netherlands, Rotterdam, the Netherlands. [11]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA. [12]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada. [13]Ministry of Education Key Laboratory of Intelligent Computation and Signal Processing, Anhui University, Hefei, China. [14]Paige AI, New York, NY, USA. [15]Scientific Data Division and Biological Systems and Engineering Division, Lawrence Berkeley National Lab, Berkeley, CA, USA. [16]Helen Wills Neuroscience Institute and Redwood Center for Theoretical Neuroscience, UC Berkeley, Berkeley, CA, USA. [17]RIKEN AIP, Tokyo, Japan. [18]Research Center for Advanced Science and Technology (RCAST), The University of Tokyo, Tokyo, Japan. [19]School of Computer Science, University of Sydney, Sydney, New South Wales, Australia. [20]Texas A&M University, College Station, TX, USA. [21]Cullen College of Engineering, University of Houston, Houston, TX, USA. [22]Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan. [23]National Centre for Computer Animation, Bournemouth University, Poole, UK. [24]PROPHESEE, Paris, France. [25]Department of Neuroscience, Columbia University, New York, NY, USA. [26]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [27]Department of Computer Science, Northern Illinois University, DeKalb, IL, USA. [28]AWS AI, Seattle, WA, USA. [29]i3S, Instituto de Investigação E Inovação Em Saúde, Universidade Do Porto, Porto, Portugal. [30]INEB, Instituto de Engenharia Biomédica, Universidade Do Porto, Porto, Portugal. [31]Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [32]Allen Institute for Brain Science, Seattle, WA, USA. [33]Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA, USA. [34]Program in Neuroscience, Department of Biomedical Sciences, Florida State University College of Medicine, Tallahassee, FL, USA. [35]Cambridge Advanced Imaging Centre, University of Cambridge, Cambridge, UK. [36]Computer Vision Laboratory, EPFL, Lausanne, Switzerland. [37]Life Sciences Institute and Department of Cell and Developmental Biology, University of Michigan, Ann Arbor, MI, USA. [38]Department of Biology, Georgetown University, Washington, DC, USA. [39]Institute for Neuroanatomy, University Medical Center Göttingen, Georg-August- University Göttingen, Goettingen, Germany. [40]Department of Stem Cell and Regenerative Biology and Center for Brain Science, Harvard University, Cambridge, MA, USA. [41]Neuroscience Institute, Georgia State University, Atlanta, GA, USA. [42]Université Paris-Saclay, CNRS, ENS Paris-Saclay, CentraleSupélec, LuMIn, Gif-sur-Yvette, France. [43]Max Planck Group: In Silico Brain Sciences, Max Planck Institute for Neurobiology of Behavior – caesar, Bonn, Germany. [44]Division of Neurobiology, MRC Laboratory of Molecular Biology, Cambridge, UK. [45]Department of Zoology, University of Cambridge, Cambridge, UK. [46]Institute for Quantitative Biosciences, University of Tokyo, Tokyo, Japan. [47]Institute of Zoology, Biocenter Cologne, University of Cologne, Cologne, Germany. [48]Department of Ecology, Evolution, and Behavior, University of Minnesota, St Paul, MN,

USA. [49]Brain Science Institute, Korea Institute of Science and Technology (KIST), Seoul, South Korea. [50]Virginia Merrill Bloedel Hearing Research Center, University of Washington, Seattle, WA, USA. [51]The Scripps Research Institute, San Diego, CA, USA. [52]Brain Research Center, National Tsing Hua University, Hsinchu, Taiwan. [53]Tencent AI Lab, Shenzhen, China. [54]Department of Zoology, Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada. [55]Zayed Centre for Rare Disease Research, UCL Great Ormond Street Institute of Child Health, London, UK. [56]Data Science Institute, Imperial College London, London, UK. [57]Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi, Japan. [58]BSEC, ORNL, Oak Ridge, TN, USA. [59]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [60]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [61]Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [62]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [63]School of Computer Science and Engineering, University of New South Wales, Sydney, New South Wales, Australia. [64]These authors contributed equally: Linus Manubens-Gil, Hanchuan Peng. ✉e-mail: meijering@imagescience.org; ascoli@gmu.edu; h@braintell.org

Methods

Dataset gathering

Fourteen neuroscience research laboratories and institutions worldwide acquired the imaging datasets used in this work, combining different imaging methods, model organisms and neuron types[6,13,25,59–73]. Supplementary Table 1 lists the metadata describing the relevant aspects of each dataset.

Quantification and statistical analysis

Image quality measurement. To quantify image quality, we implemented a plugin in Vaa3D (RRID:SCR_002609, v3.497, http://vaa3d.org) that computes the features described in detail in ref. 74. A brief description of the features measured with the plugin is given in the Supplementary Information. The Vaa3D plugin used is available at https://github.com/Vaa3D/vaa3d_tools/tree/master/hackathon/linus/image_quality. To obtain image quality measurements associated with each individual reconstruction, we also computed these features in the 3D data only in the image voxels belonging to SWC nodes of the trees. Specifically, we extended the image_quality plugin, enabling us to specify an SWC file associated with an image volume and compute the image quality metrics using only the intensity information of those voxels. Thus, the plugin computes image quality metrics in two modes: the first mode accounts for all voxels in the image (including background, blobs and non-traced neurites), and the second mode accounts only for the intensity information along the traces. We used both sets of image quality metrics to perform the analysis. Supplementary Table 3 lists the definitions of the image quality features used in the analysis.

Image preprocessing. Due to varying properties of image acquisition pipelines between institutions, it is challenging to define a universal data preprocessing protocol for reconstruction. We performed Brainbow color separation, 8-bit data type conversion and color inversion for brightfield images as the preprocessing tasks prior to automatic reconstruction algorithm bench-testing, using Vaa3D and its plugins. A brief description of the preprocessing steps is given in the Supplementary Information. Supplementary Fig. 1 shows examples of the effect of each preprocessing step on the imaging datasets.

Generation of gold standard reconstructions. A total of 14 laboratories contributed 17 datasets spanning a broad diversity of species, brain regions, neuron types, labeling methods, microscopy techniques and imaging resolution (Supplementary Table 1). Gold standard annotations were produced in the BigNeuron Neuron Annotation Workshop at Allen Institute, Seattle, 15–17 June 2015. Each reconstruction entry was manually validated by at least six annotators working collaboratively. Notably, due to limited image quality, some reconstructions can be ambiguous. Annotators were asked to make their best judgment in such cases and vote to maximize the agreement between them. Thus, we are confident that the final reconstructions reflect the best possible reconstruction generated by human experts from the respective image given the limited image quality, time and resources; we thus adopted this dataset as the gold standard for the automated reconstruction algorithms. After manual curation and postprocessing, we obtained a gold standard set of 166 reconstructed neurons.

Development of automatic reconstruction algorithms. We ported 44 implementations of 32 methods for automatic tracing as BigNeuron plugins to Vaa3D (Supplementary Table 2[18–21,37,56,58,75–91]), including 16 already existing tracing algorithms and 16 unpublished algorithms specifically developed within the scope of this project (for these latter ones, we provide a brief description in the Supplementary Information). Of the 44 implementations, 7 used a two-step process of inclusion of filamentary processes[78] to improve the results of several algorithms, and were not included to ensure a fair comparison between all base algorithms. Two variants were found to be too slow or did not generate results (PSF and LCMboost_2). We therefore considered only the remaining 35 base implementations (called 'algorithms' hereafter) for bench-testing on the BigNeuron image data (Supplementary Table 2). The consensus tree algorithm introduced in this article was used to combine the results of all of the algorithms. It is worth noting that development and porting of the tested algorithms was performed without access to the gold standard data.

Algorithm bench-testing. To bench-test the algorithms, we ran all algorithms on the image stacks of gold standard reconstructions after preprocessing. Given that the testing of 35 algorithms in the Gold166 release image volumes (163 unique 3D stacks) implied running 5,705 automatic reconstruction processes, those were parallelized in high-performance computing facilities. With a maximum time per process of 1 h, this implies 5,705 hours of computing time. In terms of memory the biggest image volume in the Gold166 set was 8 GB, and the algorithms often need more than 16 GB of RAM (random-access memory) to run. Additionally, we performed comprehensive bench-testing on the 30,000 single-neuron 3D image volumes that were gathered throughout the project, implying more than 10 million CPU (central processing unit) hours to generate more than 1.4 million tracing results. The processing was distributed using the TITAN supercomputer at Oak Ridge National Laboratory (United States), as well as supercomputers at Lawrence Berkeley National Laboratory (United States) and Human Brain Project (Europe) to ensure that the technical platform of BigNeuron could reproduce the results over different machines. Processes that took longer than 1 h of computing time were terminated and did not provide results for the algorithm–dataset pair being tested. As a result, 5,571 automatic reconstructions were generated (https://github.com/BigNeuron/Data/releases/tag/gold166_bt_v1.0). An example of a bench-testing script for calling the tested algorithms can be found at https://github.com/Vaa3D/vaa3d_tools/blob/d8e434c93708ab2a5bd349a79d9093d11aecf9d1/bigneuron_ported/bench_testing/ornl/script_MPI/gen_bench_job_text_scripts_short.sh.

Reconstruction quality benchmarking. To measure differences between automatic and gold standard neuron reconstructions, we used the neuron_distance plugin in Vaa3D. This plugin quantifies the distance between neurons, defined as the average distance between two neurons in all nearest point pairs. Given that the number of nodes can differ between pairs of reconstructions, distances are obtained twice using each reconstruction as a starting set for the search of nearest points in the other. Finally, the average bi-directional distance is calculated. Together with the average distance, the plugin also provides the percentage of nodes with pairwise distances greater than 2 voxels for each of the compared reconstructions[86]. To assess reconstruction quality, we plotted the bi-directional average distance between pairs of neurons for each reconstruction method. Other approaches include the DIADEM metric, which quantifies the similarity between two reconstructions of the same neuron by matching the locations of bifurcations and terminations as well as their topology between the two reconstructed arbors[29], or tree distances based on persistence homology[44,50]. Our approach to calculating tree distances is a simple and fast approximation, given that solving this problem is beyond the scope of this study. Still, the reported tree edit distances are robust with node density variations (Supplementary Fig. 2). Supplementary Table in summarizes the definitions of the tree edit distances used in the analysis. We additionally combined the distance metrics we measured in a single normalized aggregated metric by taking the logarithm base 10 of Euclidean distances, normalizing all metrics between 0 and 1, and obtaining the mean. Note that aggregated metric benchmark results may change depending on the definition of the aggregated metric[52].

Morphological analysis of neuron reconstructions. To consistently compare morphological features that are dependent on the size of the trees, we first scaled both gold standard and automatic reconstructions using the pixel size information of each dataset (available at https://github.com/lmanubens/BigNeuron/blob/main/scaling_gold.csv). Subsequently, to ensure consistent distance measurements in the 3D space and generation of consensus trees, we resampled the reconstruction nodes so that all of the reconstruction segments had a length of 2 µm. To avoid disconnected subtrees and inconsistent hierarchies, we sorted the reconstructions using the sort_neuron_swc plugin (using the gold standard soma location as the root) in Vaa3D. Some of the tested algorithms did not perform radius estimation by default. To provide a comparable set of reconstructions, we estimated the radius of the trees using the neuron_radius plugin. We analyzed the morphological features of post-processed trees with the batch_compute function of the blast_neuron plugin. Supplementary Table 3 lists the definitions of the morphological features used in the analysis. Furthermore, we quantified the Sholl intersection profile scale, centripetal bias and root angle distributions of the trees using the TREES toolbox functions dissectSholl_tree and rootangle_tree (ref. 34).

Interactive data analysis app. The datasets collected in this project are large, and the exploration is time-consuming. To enable fast interaction with the data, we developed an interactive analysis web app in Shiny (https://www.shinyapps.io/, v1.6.0). Shiny enables the development of web applications that build on the R programming language utilities for statistical analysis and data plotting. On the server, users can analyze dataset images, gold standard annotations, automatic reconstructions and metadata associated with each dataset. The app enables users to interactively choose the image quality and tree morphology metrics used for dimensionality reduction and clustering analyses and perform reconstruction quality benchmarking. A detailed description of the web app organization and the analyses it performs can be found in the Supplementary Information.

Generation of consensus trees. Next, we developed an algorithm to iteratively merge the reconstructions obtained by different automatic algorithms. The aim of the algorithm is to conserve tree regions reliably retrieved by different algorithms and discard possible algorithm-dependent artifacts to obtain a consensus reconstruction that is potentially closer to the ground truth. The consensus tree algorithm performs the following steps: (1) K-centroid clustering of all of the nodes in input neurons (the number of clusters is defined as the average number of nodes of the input neurons); (2) for each cluster resulting from the K-centroid clustering, the center of the cluster is taken as a consensus node; (3) by iterating through all node connections in input trees, the weights of the consensus nodes are established by collecting votes from the connections from individual input neuron trees (every time a pair of nodes of two different clusters are connected in input trees, a vote is added to the connection between the consensus nodes of the respective clusters); and (4) use of a maximum spanning tree algorithm[92] to connect consensus nodes to form the consensus tree. An implementation of the algorithm can be found as a Vaa3D plugin called 'consensus_skeleton_2'. We did statistical tests for errors in morphological metrics of automatic reconstructions using the stat_compare_means function of the ggpubr package (v0.4.0). It is worth noting that development of the consensus tree algorithms was performed without access to the gold standard data.

Prediction of the best automatic reconstruction algorithm. To predict the best algorithm in a set of automatic reconstruction methods, we used the neuromorphological features of the automatic reconstruction and the image quality features of a dataset. The statistics were transformed using the Box–Cox method to ensure normality. We generated a support vector machine regression learner using the mlr3 (v0.12.0) package in R (v3.4.1). To generate learning curves, we used the generateLearningCurveData of the mlr (v2.18.0). To generate the regression results of Fig. 5, the data were split by the IDs of the datasets into 15% for testing and 85% for training sets. We predicted the percentage of different structure between the automatic reconstruction and the gold standard with the regression model. We obtained the regression coefficient of determination using the msr function of the mlr3 package (v0.12.0). We did Wilcoxon tests (two-sided) using the stat_compare_means function of the ggpubr package (v0.4.0). The code for this analysis can be found at https://github.com/lmanubens/BigNeuron/blob/main/mlr_regression/mlr3_regression_3DIQ.R.

Showcase of best algorithm prediction on fMOST data. We predicted best-performing algorithms in fMOST datasets. A total of 40 fMOST image volumes were kindly provided by L. Ding from the Institute for Brain and Intelligence, Southeast University (20 dendritic trees and 20 axonal trees). We processed the images with the same steps used in the Gold166 dataset and obtained automatic and consensus reconstructions. We correspondingly obtained image quality and neuromorphological features. After applying a Box–Cox transformation, all of the features obtained were used as inputs to the support vector machine regression model obtained as described in the previous section.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

3D image volumes of the Gold166 dataset and gold standard reconstructions are available at http://web.bii.a-star.edu.sg/bigneuron/gold166.zip. Bench-testing automated reconstructions can be downloaded from https://github.com/BigNeuron/Data/releases/tag/gold166_bt_v1.0. The fMOST showcase datasets can be found at https://zenodo.org/record/7556104 ref. [93]. The complete set of image volumes gathered throughout the project, amounting to ~4 TB of data, is available upon request. Databases of the Allen Mouse and Human Cell Types projects (http://celltypes.brain-map.org/), Taiwan FlyCircuits (http://www.flycircuit.tw/), and Janelia FlyLight (https://www.janelia.org/project-team/flylight) can be found in the given links. Source data are provided with this paper.

Code availability

The source code developed is released as open source and is available at https://github.com/lmanubens/BigNeuron. The Shiny web app can be used at https://linusmg.shinyapps.io/BigNeuron_Gold166/ and https://neuroxiv.net/bigneuron/. The Shiny app source code can be found at https://github.com/lmanubens/BigNeuron/tree/main/shiny_app ref. [94]. With a slightly revised MIT license, see BigNeuron Shiny app license in the Supplementary Note . Source code of automated reconstruction algorithms developed throughout the project can be found at https://github.com/Vaa3D/vaa3d_tools/tree/master/released_plugins/v3d_plugins. The Vaa3D plugins license is also a slightly revised MIT license that can be found at: https://github.com/Vaa3D/vaa3d_tools/blob/master/LICENSE. The source code for the consensus tree algorithm, licensed as a Vaa3D plugin, is publicly available at https://github.com/Vaa3D/vaa3d_tools/tree/master/hackathon/xiaoxiaol/consensus_skeleton_2.

References

59. Pfeiffer, B. D. et al. Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 9715–9720 (2008).
60. Nanda, S., Das, R., Bhattacharjee, S., Cox, D. N. & Ascoli, G. A. Morphological determinants of dendritic arborization neurons in *Drosophila* larva. *Brain Struct. Funct.* **223**, 1107–1120 (2018).
61. Ikeno, H. et al. Development of a scheme and tools to construct a standard moth brain for neural network simulations. *Comput. Intell. Neurosci.* **2012**, e795291 (2012).

62. Mumm, J. S. et al. In vivo imaging reveals dendritic targeting of laminated afferents by zebrafish retinal ganglion cells. *Neuron* **52**, 609–621 (2006).
63. Yoshimatsu, T. et al. Transmission from the dominant input shapes the stereotypic ratio of photoreceptor inputs onto horizontal cells. *Nat. Commun.* **5**, 3699 (2014).
64. Bleckert, A., Schwartz, G. W., Turner, M. H., Rieke, F. & Wong, R. O. L. Visual space is represented by nonmatching topographies of distinct mouse retinal ganglion cell types. *Curr. Biol.* **24**, 310–315 (2014).
65. Druckmann, S. et al. Structured synaptic connectivity between hippocampal regions. *Neuron* **81**, 629–640 (2014).
66. Prönneke, A. et al. Characterizing VIP neurons in the barrel cortex of VIPcre/tdTomato mice reveals layer-specific differences. *Cereb. Cortex* **25**, 4854–4868 (2015).
67. Peter, M. et al. Transgenic mouse models enabling photolabeling of individual neurons in vivo. *PLoS One* **8**, e62132 (2013).
68. Gao, Y., Liu, L., Li, Q. & Wang, Y. Differential alterations in the morphology and electrophysiology of layer II pyramidal cells in the primary visual cortex of a mouse model prenatally exposed to LPS. *Neurosci. Lett.* **591**, 138–143 (2015).
69. Chen, H. et al. Fast assembling of neuron fragments in serial 3D sections. *Brain Inform.* **4**, 183–186 (2017).
70. Brito, J. et al. Neuronize: a tool for building realistic neuronal cell morphologies. *Front. Neuroanat.* **7**, 15 (2013).
71. Shi, Y., Kirwan, P., Smith, J., Robinson, H. P. C. & Livesey, F. J. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat. Neurosci.* **15**, 477–486 (2012).
72. Wang, Y. et al. Intense and specialized dendritic localization of the fragile X mental retardation protein in binaural brainstem neurons: a comparative study in the alligator, chicken, gerbil, and human. *J. Comp. Neurol.* **522**, 2107–2128 (2014).
73. He, H.-Y., Shen, W., Hiramoto, M. & Cline, H. T. Experience-dependent bimodal plasticity of inhibitory neurons in early development. *Neuron* **90**, 1203–1214 (2016).
74. Bray, M.-A. & Carpenter, A. E. Quality control for high-throughput imaging experiments using machine learning in cellprofiler. In *High Content Screening: A Powerful Approach to Systems Cell Biology and Phenotypic Drug Discovery* (eds Johnston, P. A. & Trask, O. J.) 89–112 (Springer, 2018).
75. Homan, A. C., van Knippenberg, D., Van Kleef, G. A. & De Dreu, C. K. W. Bridging faultlines by valuing diversity: diversity beliefs, information elaboration, and performance in diverse work groups. *J. Appl. Psychol.* **92**, 1189–1199 (2007).
76. Peng, H., Long, F. & Myers, G. Automatic 3D neuron tracing using all-path pruning. *Bioinformatics* **27**, i239–i247 (2011).
77. Yang, J. et al. FMST: an automatic neuron tracing method based on fast marching and minimum spanning tree. *Neuroinformatics* **17**, 185–196 (2019).
78. Gu, L. & Cheng, L. Learning to boost filamentary structure segmentation. In *2015 IEEE International Conference on Computer Vision (ICCV)* 639–647 (2015). https://doi.org/10.1109/ICCV.2015.80
79. Wan, Z., He, Y., Hao, M., Yang, J. & Zhong, N. M-AMST: an automatic 3D neuron tracing method based on mean shift and adapted minimum spanning tree. *BMC Bioinformatics* **18**, 197 (2017).
80. Wu, J. et al. 3D BrainCV: simultaneous visualization and analysis of cells and capillaries in a whole mouse brain with one-micron voxel resolution. *Neuroimage* **87**, 199–208 (2014).
81. Lee, P.-C., Chuang, C.-C., Chiang, A.-S. & Ching, Y.-T. High-throughput computer method for 3D neuronal structure reconstruction from the image stack of the *Drosophila* brain and its applications. *PLoS Comput. Biol.* **8**, e1002658 (2012).
82. Quan, T. et al. NeuroGPS: automated localization of neurons for brain circuits using L1 minimization model. *Sci. Rep.* **3**, 1414 (2013).
83. Zhao, T., Olbris, D. J., Yu, Y. & Plaza, S. M. NeuTu: software for collaborative, large-scale, segmentation-based connectome reconstruction. *Front. Neural Circuits* **12**, 101 (2018).
84. Bas, E. & Erdogmus, D. Principal curves as skeletons of tubular objects: locally characterizing the structures of axons. *Neuroinformatics* **9**, 181–191 (2011).
85. Sironi, A., Turetken, E., Lepetit, V. & Fua, P. Multiscale centerline detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 1327–1341 (2016).
86. Peng, H., Ruan, Z., Atasoy, D. & Sternson, S. Automatic reconstruction of 3D neuron structures using a graph-augmented deformable model. *Bioinformatics* **26**, i38–i46 (2010).
87. Liu, S. et al. Rivulet: 3D neuron morphology tracing with iterative back-tracking. *Neuroinformatics* **14**, 387–401 (2016).
88. Minemoto, T. et al. SIGEN: system for reconstructing three-dimensional structure of insect neurons. In *Proceedings of the Asia Simulation Conference, JSST2009,* CDROM 1–6 (2009).
89. Yang, J., Gonzalez-Bellido, P. T. & Peng, H. A distance-field based automatic neuron tracing method. *BMC Bioinformatics* **14**, 93 (2013).
90. Chen, H., Xiao, H., Liu, T. & Peng, H. SmartTracing: self-learning-based neuron reconstruction. *Brain Inform.* **2**, 135–144 (2015).
91. Zhou, Z., Liu, X., Long, B. & Peng, H. TReMAP: automatic 3D neuron reconstruction based on tracing, reverse mapping and assembling of 2D projections. *Neuroinformatics* **14**, 41–50 (2016).
92. Prim, R. C. Shortest connection networks and some generalizations. *The Bell System Technical Journal* **36**, 1389–1401 (1957).
93. Manubens-Gil, L. BigNeuron fMOST showcase image data. https://doi.org/10.5281/zenodo.7556104 (2023).
94. Manubens-Gil, L. lmanubens/BigNeuron: BigNeuron Shiny app v1.0.0 code base. https://doi.org/10.5281/ZENODO.7556112 (2023).

Acknowledgements

This project was supported by Allen Institute for initialization and a series of events, and a DOE Oak Ridge National Laboratory Leadership Computing award to H.P.; the cross-platform bench-test was also supported intensively at Lawrence Berkeley National Lab, Allen Institute for Brain Science, Blue Brain Project at EPFL, Southeast University at Nanjing, and various other facilities. This project also used large-scale display wall and immersive virtual reality facilities at Imperial College of London, Oak Ridge National Laboratory, Janelia Research Campus of HHMI, and Southeast University. This project was also supported by Tencent Inc. and Southeast University for interactive online analysis. Much of the work of this project reported here is a joint community effort with many sponsors over 7 years. The authors acknowledge Prabhat, Z. Wan, J. Yang, H. Zhou, A. Narayanaswamy, S. Zeng, P. Glowacki, D. Jin, Z. Zheng, P. Hong, T. Zeng, R. Li, H. Ikeno, Y.-T. Ching, T. Quan, J.-F. Evers, C. Murtin, S. Gao, Y. Zhu, Y. Yang, H. Ai, S. Ishii, E. Hottendorf, T. Kawase, H. Dong and a number of other colleagues for assistance in developing and porting the automated reconstruction algorithms, and discussion. The authors thank S. Sorensen for assistance in discussing and tracing some gold standard neuron reconstructions, and R. Kanzaki (University of Tokyo), D. Miyamoto (University of Tokyo), R. Wong (University of Washington), Y. Wang (Allen Institute for Brain Science), E. Lein (Allen Institute for Brain Science), C. Bass (King's College London), S. Danzer (Cincinnati Children's Hospital) and many other colleagues for providing neuron image datasets. The authors also acknowledge A. Jones for suggesting the project name and support throughout the project, and thank R.

Yuste and D. Van Essen for discussion; J. Isaac and K. Moses for support from the Wellcome Trust to organize the University of Cambridge hackathon; G. Rubin and N. Spruston for support, data sharing and event organization at the Janelia Research Campus of HHMI; INCF for financial support for organization of a series of meetings; and Beijing University of Technology, Imperial College of London and Southeast University for support of hackathons and workshops. The authors also thank X. Zhao for assistance in hosting the Shiny app in the neuroXiv website; P. Qian, Z. Zhao and X. Chen for assistance in formatting the manuscript; and A. Carpenter for discussion in organizing a tracing hackathon. The authors acknowledge the support of the National Science and Technology Innovation 2030 – 'Brain Science and Brain-Inspired Research' Program of China (Grant 2021ZD0204002 Q33 and 2022ZD0205200 to H.P, L.M.-G., Y.L., Z.R.). B.Y. was supported by the NIH grant R01 EB028159. G.A.A. was supported by NIH grants R01 NS39600, RF1 MH128693 and R01 NS86082. A.-S.C. was supported by the Higher Education Sprout Project co-funded by the Ministry of Education and the Ministry of Science and Technology in Taiwan. M.S. was supported by the European Programs JPND TransPathND (ANR-17-JPCD-0002), EuroNanoMed III MoDiaNo (ANR-18-ENM3-0002) and CNES. The authors acknowledge the National Center for High-Performance Computing in Taiwan for managing the FlyCircuit data. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Q34 H.P. conceived and conceptualized this project. H.P., G.A.A. and E.M. formed the managing team and communicated with other key investigators including A.R., J.Y., J.R., K.I., J.K., G.J., P.G.-B., Y.G., N.Z., G.T., S.H., M.H. and C.K. to secure key resources to execute the project. H.P. envisioned and led the development of the computational and data analysis platform. L.M.-G. developed the data analysis for this study. H.C. assisted in hosting the Shiny app in the neuroXiv website. Z.Z., with the help of H.P., bench-tested neuron reconstruction algorithms and generated the automatic morphology reconstructions. Y.L. bench-tested neuron reconstruction algorithms in the fMOST datasets. H.C. and H.P. color-separated single neurons from fruitfly brain images generated by A.N. and colleagues in Gerald Rubin lab and the Janelia FlyLight Project Team. H.C., S.N., A.N. and C.D. participated in tracing hackathons and contributed to the generation of the gold standard test dataset. Y.W., E.R., L.M., B.Y., H.Z., H.T.C., A.-S.C., J.F.S., M.P., R.L., D.N.C., H.-Y.H, M.S., K.I., J.K., G.J., P.G.-B. and M.O. contributed testing neuron images. A.B., T.G., Z.R., X.L., Y.L., K.E.B., L.G., L.C., J.Y., L.Q., S.L., H.Y., W.C., S.J., B.R., C.-W.W., A.S., P.F., M.R., T.Z., D.M.I., J.Z., T.L., E.B., E.C.-S., P.A. and J.S. contributed reconstruction algorithms. L.M.-G., G.A.A., E.M. and H.P. wrote the manuscript with assistance from co-authors.

Competing interests

Z.Z. is employed by Microsoft. The company did not influence the Q35 research. H.C. and J. Yao are employed by Tencent AI Lab. Tencent provides the initial hosting of the Shiny website. The company does not expect to benefit financially. X. Liu is employed by Kaya Medical. The company did not influence the research. M.R. is employed by Nuctech. The company did not influence the research. S.L. is employed by Paige AI. The company did not influence the research. A.S. is employed by PROPHESEE. The company did not influence the research. E.B. is employed by AWS AI. The company did not influence the research. All other authors have no competing interests. Q36

Additional information

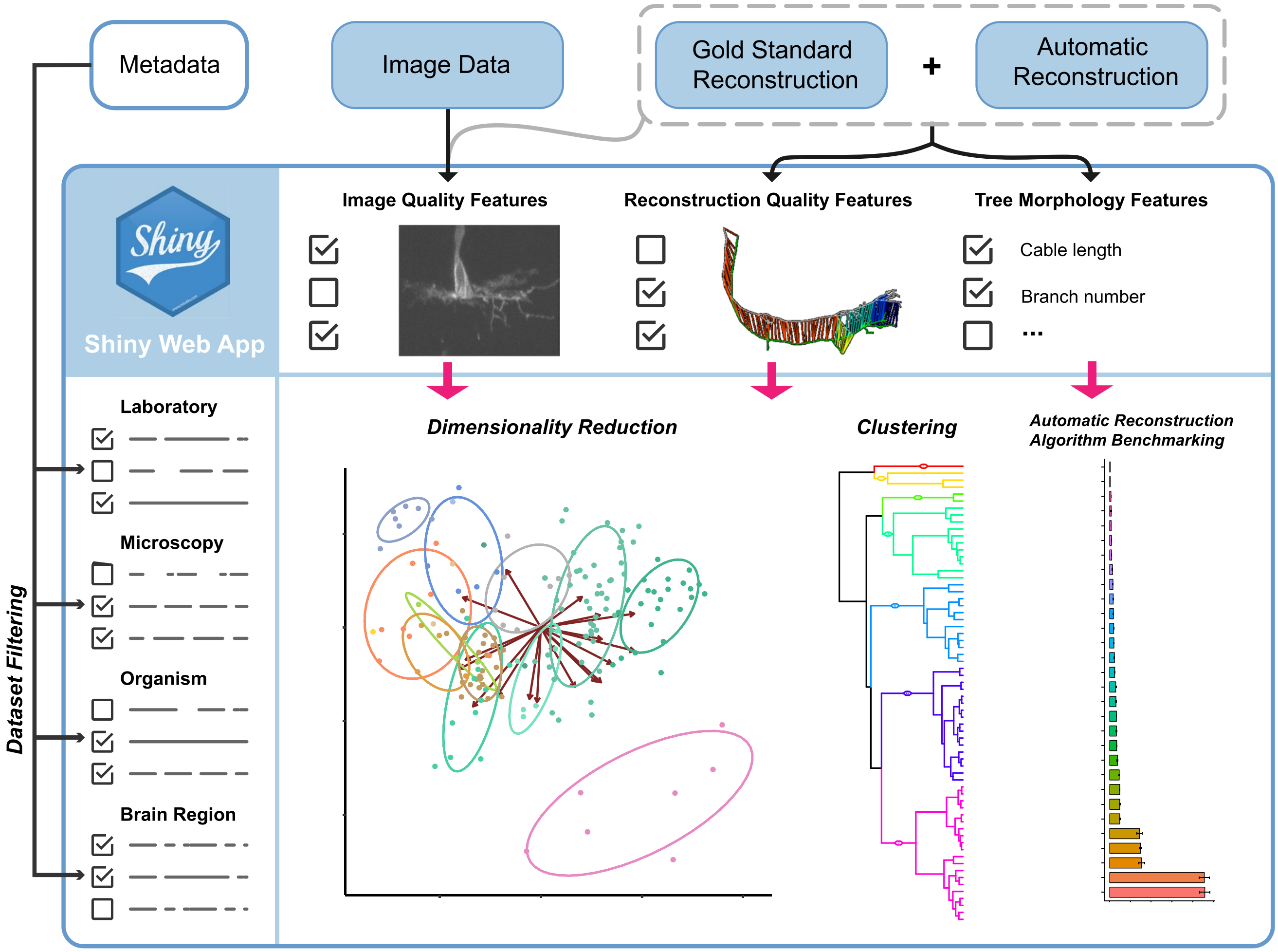
Extended data are available for this paper at https://doi.org/10.1038/s41592-023-01848-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01848-5.

Correspondence and requests for materials should be addressed to Erik Meijering, Giorgio A. Ascoli or Hanchuan Peng.

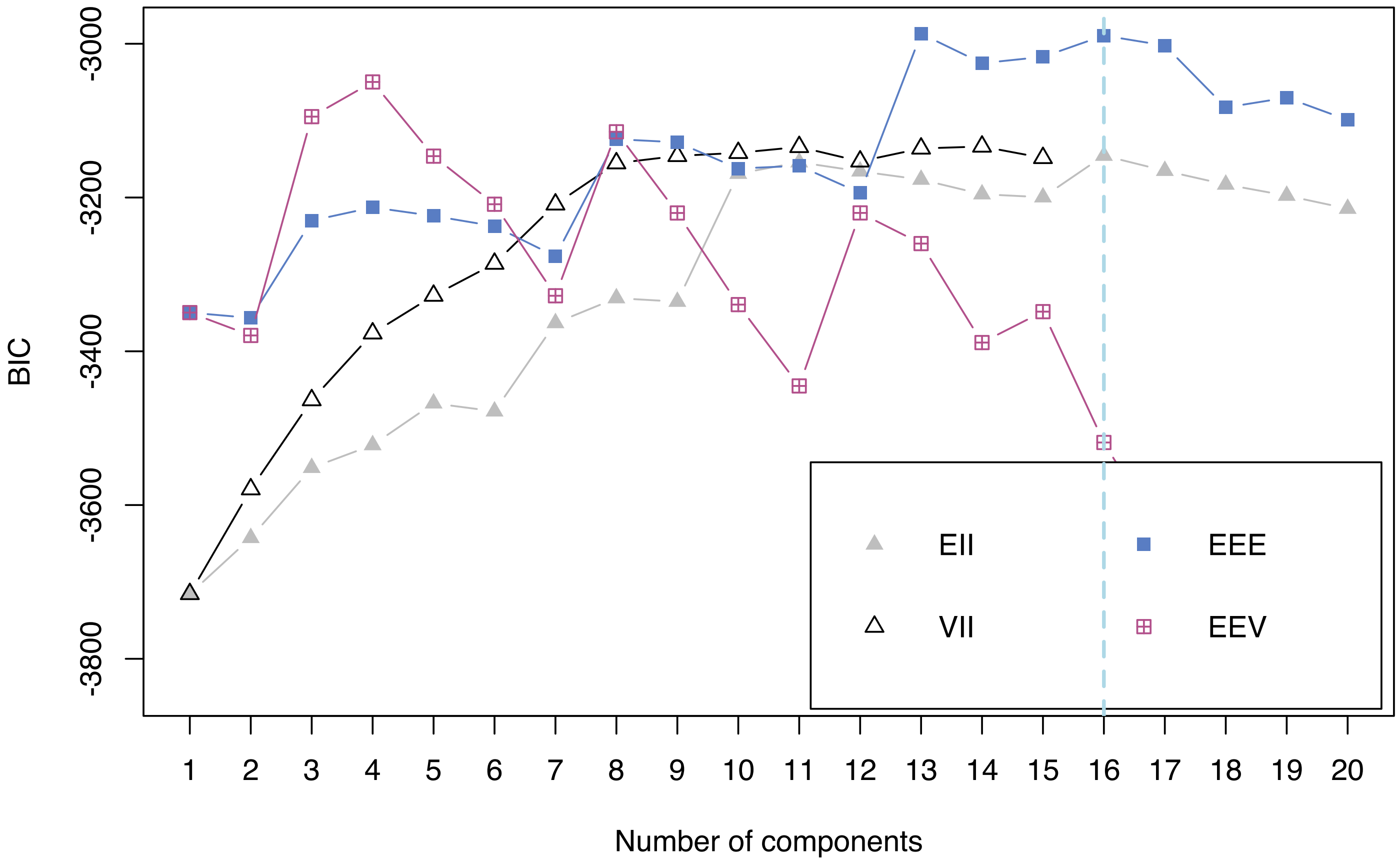
Peer review information *Nature Methods* thanks Chao Chen and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

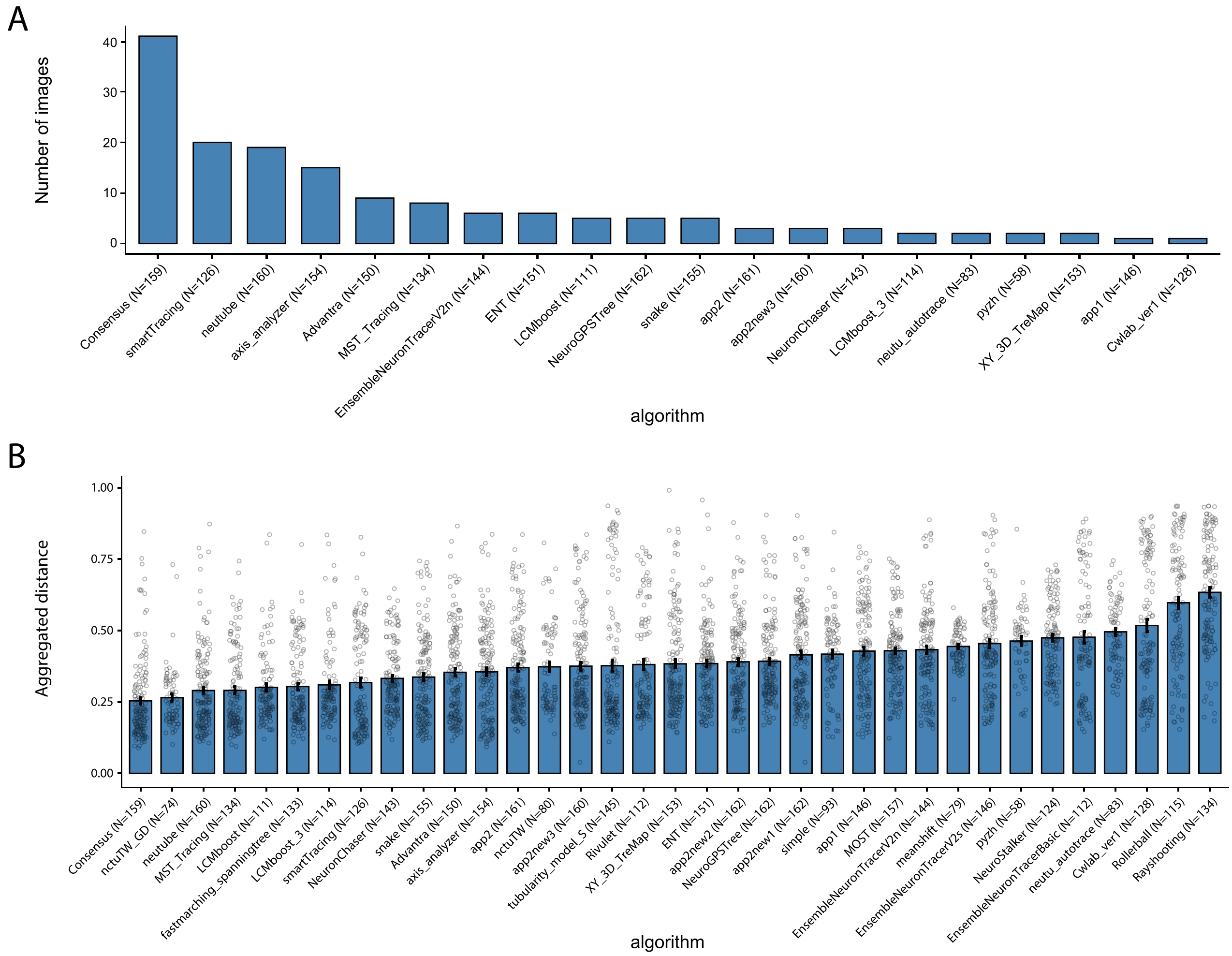


Extended Data Fig. 1 | A web app to allow interactive navigation of heterogeneous bench-testing results. Visualization of the Shiny interactive web app (https://linusmg.shinyapps.io/BigNeuron_Gold166/ and https://neuroxiv.net/bigneuron/). The data loaded into the app includes the dataset images, gold standard annotations, automatic reconstructions, and metadata associated with each dataset. Users can interactively choose the image quality and tree morphology metrics used for dimensionality reduction and cluster analysis, and perform reconstruction quality benchmarking. Documentation for the usage of the app can be found at https://github.com/lmanubens/BigNeuron.

https://doi.org/10.1038/s41592-023-01848-5

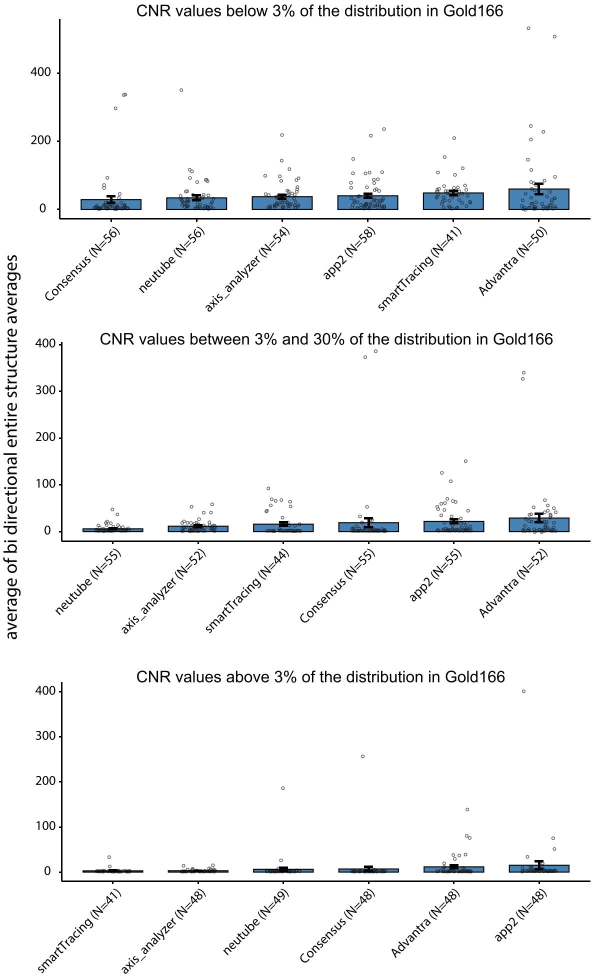


Extended Data Fig. 2 | Bayesian Information Criterion (BIC) for parametrized Gaussian Mixture models fitted by the expectation-maximization algorithm. Each colored symbol indicates the BIC for a given mixture model with a number of components specified in the x axis. 'EII': spherical, equal volume; 'VII': spherical, unequal volume; 'EEE': ellipsoidal, equal volume, shape, and orientation; 'EEV': ellipsoidal, equal volume and equal shape. The dashed light blue line indicates the maximum BIC. The Bayesian Information Criterion is a measure for the comparative evaluation among a finite set of statistical models, the measure is based on maximizing the likelihood function while penalizing for the number of parameters in the models[36].

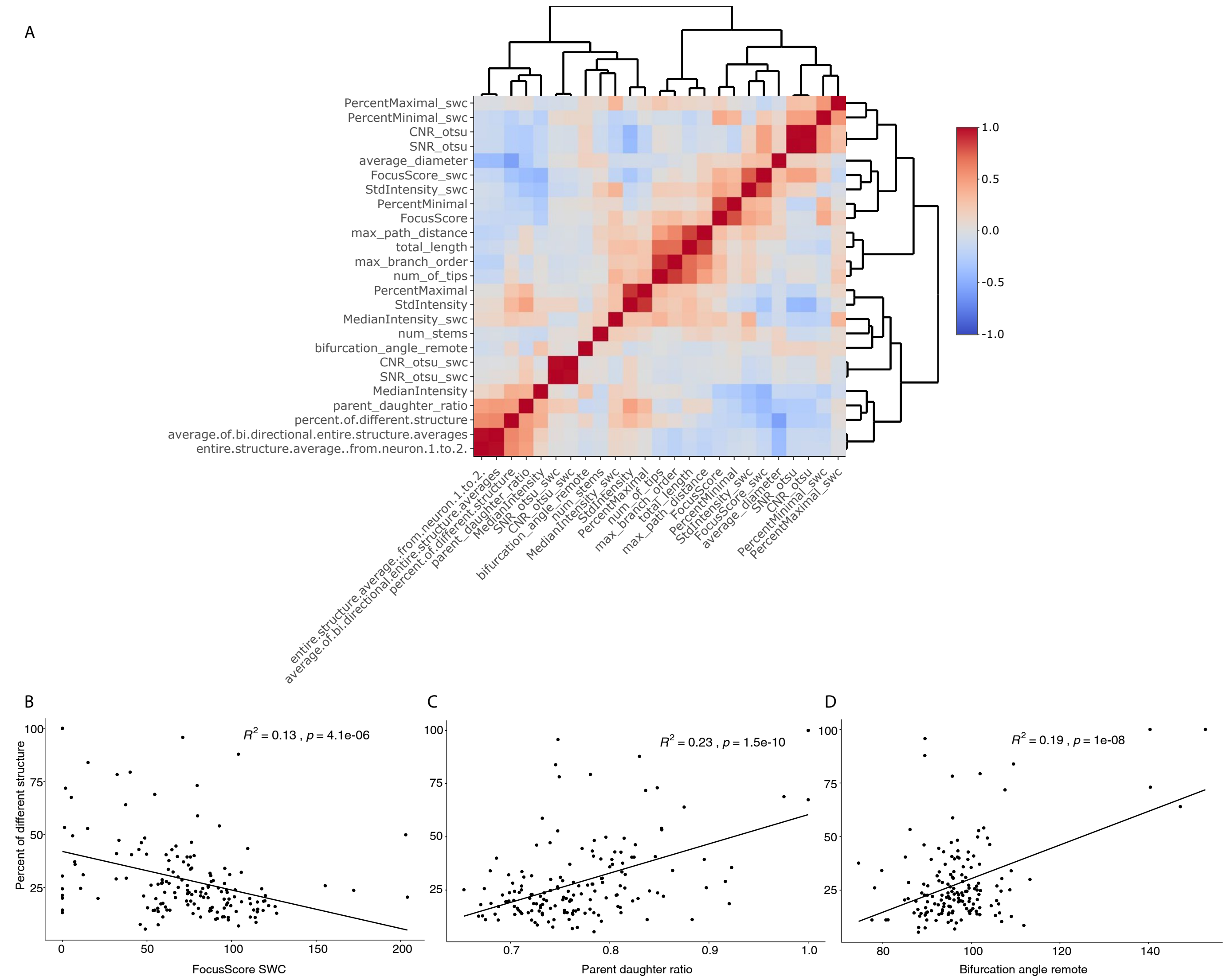


Extended Data Fig. 3 | Overall benchmark of best-performing algorithms. **a** Number of images in which specific automatic tracing algorithms outperform the others. For each image, the algorithm having the smallest average bi-directional entire structure average distance against the gold standard was considered the best. The number of times each algorithm was found to be best is shown as a bar plot. The number of times each algorithm produced a result in the full Gold166 dataset is indicated in parentheses in the labels. **b** Overall benchmark of all algorithms accounting for all measured distances to gold standards with an aggregated metric. Mean + /− Standard Errors are shown as bar plots. Each dot represents the distance quantification for each neuron. The number of times each algorithm produced a result in the full Gold166 dataset is indicated in parentheses in the labels.

https://doi.org/10.1038/s41592-023-01848-5

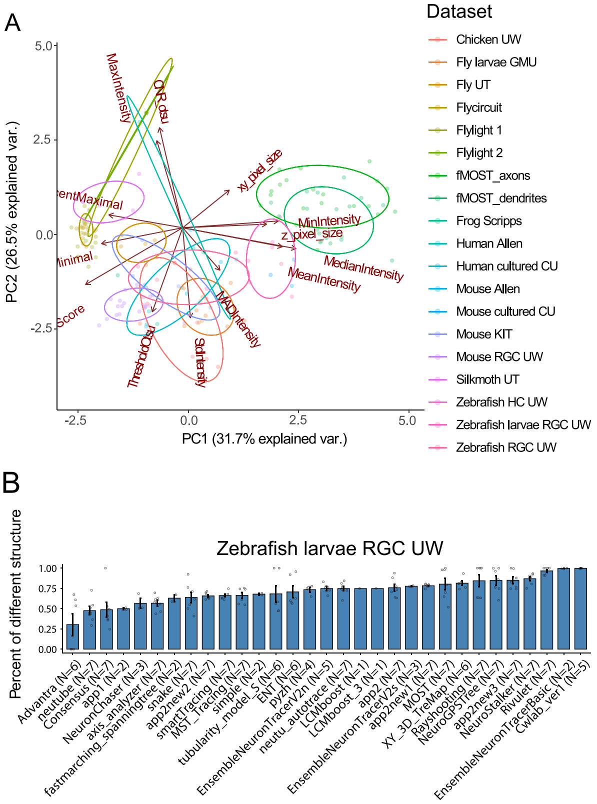


Extended Data Fig. 4 | Supplementary benchmarks of best-performing algorithms. Benchmarks of the 6 overall best-performing algorithms based on Extended Data Fig. 3a for subsets of Gold166 with different CNR based on the Otsu threshold. Means + /− Standard Errors are presented as bar plots. Each dot represents the distance quantification for each neuron. The number of times each algorithm produced a result in the full Gold166 dataset is indicated in parentheses in the labels.



Extended Data Fig. 5 | Image quality metrics correlate with the accuracy of automated tracing. Hierarchical clustering among image quality metrics, tree morphological features, and reconstruction quality. Reconstruction quality correlates with a set of features, indicating that more focused images of big neurons tend to provide better automatic reconstruction results. **a** The heatmap indicates color-coded pairwise Pearson correlations between metrics obtained for consensus tree reconstructions. **b**–**d** Correlation plots for image quality and dendritic tree morphology features (B: Focus Score in SWC nodes, C: parent–daughter ratio, and D: bifurcation angle remote) and consensus reconstruction quality (% of different structure). P values indicate the result of two-sided tests for correlation.

https://doi.org/10.1038/s41592-023-01848-5



Extended Data Fig. 6 | Gold166 subset that most closely resembles fMOST data according to image quality features. **a** Principal Component Analysis of gold standard datasets accounting for their image quality metrics. Each point is one gold standard image volume, and the color indicates the dataset it comes from. Arrows represent the direction of each variable in the PCA space. Longer arrows belong to variables that are well represented by the two principal components. Given that 68% of the density of multi variate normal distributions are within 1 Mahalanobis distance of the mean, 68% confidence normal data ellipses for each group are drawn with solid lines. **b** Percentage of different structure between automatic reconstructions and gold standard trees for the Zebrafish larvae RGC neurons. Mean + /− Standard Errors of percentage of different structure are shown as bar plots.

QUERY FORM

Manuscript ID	[Art. Id: 1848]
Author	Linus Manubens-Gil

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	*Nature of Query*
Q1:	Reference [59] is a duplicate of [6] and hence the repeated version has been deleted. Please check.
Q2:	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.
Q3:	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.
Q4:	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).
Q5:	You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes. If these are not considered scientific changes, any altered Supplementary files will not be used, only the originally accepted version will be published.
Q6:	Please check Figures for accuracy as they have been relabelled. Please markup minor changes in the eProof. For major changes, please provide revised figures. (Please note that in the eProof the figure resolution will appear at lower resolution than in the pdf and html versions of your paper.)
Q7:	If applicable, please ensure that any accession codes and datasets whose DOIs or other identifiers are mentioned in the paper are scheduled for public release as soon as possible, we recommend within a few days of submitting your proof, and update the database record with publication details from this article once available.
Q8:	In the Abstract please confirm that the edits to the sentence 'We observed that diverse algorithms….' preserve the originally intended meaning.
Q9:	Your paper has been copyedited. Please review every sentence to ensure that it conveys your intended meaning. If changes are required, please provide further clarification rather than revert-

Uncorrected proofs

QUERY FORM

Manuscript ID	[Art. Id: 1848]
Author	Linus Manubens-Gil

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	*Nature of Query*
	ing to the original text. Please note that formatting (including hyphenation, Latin words, and any reference citations that may be mistaken for exponents) has been made consistent with our house style.
Q10:	Please check that all funders have been appropriately acknowledged and that all grant numbers are correct.
Q11:	Please check that the Competing Interests declaration is correct as stated. If you declare competing interests, please check the full text of the declaration for accuracy and completeness.
Q12:	Please ensure that genes are correctly distinguished from gene products: for genes, official gene symbols (e.g. NCBI Gene) for the relevant species should be used and italicized; gene products such as proteins and non-coding RNAs should not be italicized.
Q13:	Please confirm that the edits to the Fig 1 legend preserve the originally intended meaning.
Q14:	In the Fig 1 legend, in the sentence 'We developed…' is the insertion of 'in any single-neuron imaging dataset' correct?
Q15:	Please confirm that the edits to the two sentences ‘To quantify the heterogeneity ….’ and 'For seven out of…' preserve the originally intended meaning.
Q16:	In the Fig 2b legend please provide the definitions for 'CU, XXX; GMU, XXX; HC, XXX; KIT, XXX; UT, XXX; UW, XXX' and please confirm/correct the inserted defintion of RGC 'retinal ganglion cell'.
Q17:	In the sentence 'As shown in the PCA….' is the insertion of 'while the silkmoth neurons have' correct?
Q18:	In the sentence 'Our analysis shows…' is the insertion of ‘cornu ammonis (CA1) and dentate gyrus’ as the spelling out of 'CA1 and DG' correct?
Q19:	Please confirm that the edits to the sentence ‘We assigned a confidence value ….’ preserve the originally intended meaning.

Uncorrected proofs

QUERY FORM

Manuscript ID	**[Art. Id: 1848]**
Author	**Linus Manubens-Gil**

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	*Nature of Query*
Q20:	In Fig 3 parts d and e please check the edits to the y axis label
Q21:	The Fig 3d,e legends refer to clusters number 8 and 9 in Fig 2i, but the clusters are not numbered in Fig 2i therefore it is not clear which clusters are being referred to. Please check.
Q22:	Please confirm that the edits to the sentence 'We assessed the quality....' preserve the originally intended meaning.
Q23:	In the FIg 4a legend please confirm that the edits to the sentence 'Each point represents....' preserve the originally intended meaning.
Q24:	Please confirm that the edits to the sentence ' Visual inspection of the reconstructions ....' preserve the originally intended meaning.
Q25:	In the sentence 'A comparison of image...' is the insertion of 'zebrafish larvae retinal ganglion cell dataset' correct?
Q26:	In Fig 5 parts a and e please clarify the data format of the bar plots. Do they represent the mean +/- s.e.? Please provide the correct meaning.
Q27:	In the Fig 5g,h legend is the insertion of 'in the image volume represented in f' correct?
Q28:	Please confirm that the edits to the sentence 'The diversity in the data....' preserve the originally intended meaning.
Q29:	In the sentence 'However, to our knowledge...' is the insertion of 'and this learning algorithm has been included in our analysis' correct?
Q30:	In the sentence 'We also provide here...' is the insertion of 'can be the best performer' correct?
Q31:	In the sentence 'This plugin quantifies...' is the insertion of 'defined as the average distance between two neurons in all nearest point pairs' correct?

Uncorrected proofs

QUERY FORM

Manuscript ID	[Art. Id: 1848]
Author	Linus Manubens-Gil

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	*Nature of Query*
Q32:	Please reword 'Supplementary Table in summarizes' because the meaning is unclear. Please also insert the Supplementary Table number.
Q33:	In the Acknowledgements section, in 'Grant 2021ZD0204002 and 2022ZD0205200 to H.P, L.M.-G., Y.L, Z.R.)' please clarify 'Y.L.': does this refer to 'Y. Liu' or 'Y. Li'?
Q34:	In the Author Contributions section please spell out the two uses of J.Y. to indicate which is J. Yang and which is J. Yao. Please spell out the two uses of Y.L. to indicate which is Y. Liu and which is Y Li. And please spell out the one use of X.L. as either X. Liu or X. Li, and please indicate where the other author should be added because X.L. appears only once.
Q35:	In the Competing Interests section 'J.Y.' has been changed to 'J. Yao' and and 'X.L.' has been changed to 'X. Liu'. Are these the correct authors?
Q36:	Please check and correct the source data information for 'Extended Data Fig. 8' because only six extended data figs have been supplied.

Uncorrected proofs

nature portfolio

Corresponding author(s): Erik Meijering, Giorgio A. Ascoli and Hanchuan Peng

Last updated by author(s): Jan 19, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Gold standard neuron reconstructions were obtained using Vaa3D (RRID:SCR_002609, version 3.497, http://vaa3d.org) .
Data analysis	The source code developed is released open source and is available at https://github.com/lmanubens/BigNeuron. The Shiny web app can be used at https://linusmg.shinyapps.io/BigNeuron_Gold166/ and https://neuroxiv.net/bigneuron/. The Shiny app source code can be found at: https://github.com/lmanubens/BigNeuron/tree/main/shiny_app. Source code of automated reconstruction algorithms developed throughout the project can be found at https://github.com/Vaa3D/vaa3d_tools/tree/master/released_plugins/v3d_plugins. The source code for the consensus tree algorithm is publicly available at: https://github.com/Vaa3D/vaa3d_tools/tree/master/hackathon/xiaoxiaol/consensus_skeleton_2. We developed the source code in R (version 3.4.1.) and relied in the following R packages: ggpubr (version 0.4.0), mlr3 (version 0.12.0), Shiny (https://www.shinyapps.io/, version 1.6.0), prcomp (version 4.0.3), ggbiplot (version 0.55), factoextra (version 1.0.7), Rtsne (version 0.15), nat (version 1.8.16), heatmaply (version 1.1.0), mclust (version 5.4.7), ggpubr (version 0.4.0). Additionally, we used the TREES Toolbox (version 2.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

3D image volumes of the Gold166 dataset and gold standard reconstructions are available at http://web.bii.a-star.edu.sg/bigneuron/gold166.zip. Bench-testing automated reconstructions can be downloaded from https://github.com/BigNeuron/Data/releases/tag/gold166_bt_v1.0. The fMOST showcase datasets can be found at https://zenodo.org/record/7556104. The complete set of image volumes gathered throughout the project, amounting to ~4TB of data, is available upon request.
Databases of the Allen Mouse and Human Cell Types projects (http://celltypes.brain-map.org/), Taiwan FlyCircuits (http://www.flycircuit.tw/), and Janelia Fly Light (https://www.janelia.org/project-team/flylight) can be found in the links mentioned in parentheses.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of gold standard datasets used for the study was determined by the throughput of manual annotation hackathons. We did a sample size estimation based on the maximum branch order measured in two types of neurons with N=6 (m1=12.5, m2=16.7, SD=4.45). A sample size of N=18 would have a Type I error rate of 0.05 and a power of 0.80. Thus, we expected ~160 datasets would be sufficient to assess the accuracy of automated tracing algorithms.
Data exclusions	No data were excluded for the analysis.
Replication	All attempts at replicating the results were sucessful using the Shiny app and published code.
Randomization	Randomization was not applicable to this study because tracing algorithms were applied to all datasets. Random training and testing samples were used to train the SVM regression.
Blinding	Blinding was not applicable in this study because analysis was performed a few years after development of the algorithms by different researchers.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging