Petabyte-Scale Multi-Morphometry of Single Neurons for Whole Brains

Shengdian Jiang[1,2,+], Yimin Wang[1,3,+], Lijuan Liu[1,+], Sujun Zhao[1], Mengya Chen[3], Xuan Zhao[1,2], Peng Xie[1], Liya Ding[1], Zongcai Ruan[1], Hong-Wei Dong[4], Giorgio A. Ascoli[5], Michael Hawrylycz[6], Hongkui Zeng[6], Hanchuan Peng[1,6,*]

[1] SEU-ALLEN Joint Center, Institute for Brain and Intelligence, Southeast University, Nanjing, Jiangsu, China
[2] School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China
[3] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[4] Department of Neurobiology, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA
[5] Volgenau School of Engineering, George Mason University, Fairfax, VA 22030, USA
[6] Allen Institute for Brain Science, Seattle, WA 98109, USA
[+] Equal contribution
[*] Correspondence: Hanchuan Peng (h@braintell.org)

Abstract: Recent advances in brain imaging allow producing large amounts of 3-D volumetric data from which morphometry data is reconstructed and measured. Fine detailed structural morphometry of individual neurons, including somata, dendrites, axons, and synaptic connectivity based on digitally reconstructed neurons, is essential for cataloging neuron types and their connectivity. To produce quality morphometry at large scale, it is highly desirable but extremely challenging to efficiently handle petabyte-scale high-resolution whole brain imaging database. Here, we developed a multi-level method to produce high quality somatic, dendritic, axonal, and potential synaptic morphometry, which was made possible by utilizing necessary petabyte hardware and software platform to optimize both the data and workflow management. Our method also boosts data sharing and remote collaborative validation. We highlight a petabyte application dataset involving 62 whole mouse brains, from which we identified 50,233 somata of individual neurons, profiled the dendrites of 11,322 neurons, reconstructed the full 3-D morphology of 1,050 neurons including their dendrites and full axons, and detected 1.9 million putative synaptic sites derived from axonal boutons. Analysis and simulation of these data indicate the promise of this approach for modern large-scale morphology applications.

INTRODUCTION

Reconstructing the complete 3-D shape or morphology of a neuron, including its dendrites and axons in their entirety, as well as finer structures such as the somata, dendritic spines, and axonal terminal boutons, is recognized as a crucial step to profile the myriad types of neurons in brains[1–4]. This technique, which we refer to as *Multi-Morphometry*, has begun to generate intriguing information and hypotheses about brain circuits at the single-neuron / single-synapse level[5–7].

Mammalian brains, of at least hundreds of cubic millimeters in volume, are very large when sub-micrometer resolution imaging is used to acquire 3-D volumetric image datasets at the whole-brain

scale. A fundamental challenge in multi-morphometry is that sub-micrometer resolution is necessary to analyze synaptic patterns in a neuron's arborization, while whole-brain scale is essential to delineate long projecting axonal arbors[8]. As a result, even for the mouse brain, a widely used model system of mammalian brains, a typical 3-D brain-image dataset will have tens of teravoxels in volume[9,10]. On the other hand, neurons have a very complicated tree-like shape, and are often labelled and visualized sparsely using chemical[11,12], transgenic[13] or viral approaches[14,15]. The number of morphologically distinguishable neurons per brain is often limited. Therefore, to understand the vast complexity and variation of neurons, it is crucial to obtain a large collection of brain image datasets[16,17]. As each voxel is often stored as one or more bytes, the multi-morphometry problem arises as a petabyte-computing challenge, and as a paramount task for current bioimage informatics applications and technologies[18–21].

There is a long history of reconstructing individual neurons' morphology with image analysis[22,23]. Subneuronal structures including somata, spines and boutons have also been segmented and analyzed from images[5,24–28]. This is a challenge of high community interest. A number of algorithms have been examined and compared against each other in public initiatives, e.g. DIADEM[29] or in the global collaborative BigNeuron initiative[30]. However, most existing methods are applicable only to smaller image datasets and partial neuronal structures. For individual mammalian brain datasets, technologies that can handle teravoxels of image volume to trace millimeters long neurite fibers emerged only recently, including TeraFly[31], UltraTracer[32], BigDataViewer[33], and TeraVR[34]. Manual and semi-automatic methods were used to trace neurons' full skeletons in Janelia's MouseLight project[2]. Yet, it is largely an open problem how to scale up all these approaches to handle petabyte-scale multi-morphometry challenge that is becoming a compelling reality as whole-brain screening projects involving increasingly larger and more complicated animal models are being carried out internationally (BRAIN Initiative (https://braininitiative.nih.gov/), Blue Brain Project[35], etc.).

RESULTS

We attempted this petabyte (PB)-scale whole-brain computing challenge by introducing a technology that involves a hardware platform that is able to handle petabytes of data storage, sharing, computing and visualization, a software platform called *MorphoHub* that can utilize such hardware platform, and most importantly, a mechanism to scale up the synergized automated computation and multi-user collaboration for effective validation and correction. Our method is centered around reconstructing the multi-level information of single neuron's morphology (**Fig. 1a**) at the whole-brain scale for a number of brains. The *MorphoHub* software package is able to streamline the workflow of imaging data management, visualization, reconstruction and collaboration, and data sharing (**Fig. 1b**, **Methods**). Using this approach, we extended several state-of-the-art methods to the PB-scale (**Supplementary Table 1**) and produced multi-morphometry data from such massive image database (**Fig. 1c**). Our method allows a smooth transition from manual and interactive morphometry acquisition to increasingly routine work done by automatic algorithms as we show below.

A key component of our method is a three-level (L1, L2, and L3) reconstruction approach (**Fig. 1a**). We incrementally reconstruct morphological components of neurons, including somata, dendrites, axons, spines and boutons, *only* when such information can be produced faithfully and affordably. Specifically, an L1 reconstruction contains the full dendritic arbor and the skeleton of all axonal neurite tracts, excluding the fine structures of distal axonal arbors (**Fig. 1a** and **Fig. 2**). An L2 reconstruction contains the complete structures of all neuronal arbors (**Fig. 1a** and **Fig. 2**). An L3 reconstruction contains the identification of two key elements of synaptic connectivity, dendritic spines and axonal boutons, as well as other structures of potential interest (e.g., specific topology of axonal branching patterns, modeling of specific neuronal compartments' shape) (**Fig. 1a** and **Fig. 3**).

The proposed multi-level reconstruction method is generic and scalable to single neuron datasets of arbitrary size if proper data structure and data workflow are in place. To provide such capability for a real PB-scale computing environment, we developed *MorphoHub* to manage all data flow and processing procedures in an integrated way (**Fig. 1b** and **Supplementary Fig. 1**, **Methods**). *MorphoHub* handles four heterogeneous data types, including image volumes, neuron morphology, meta-data of user interactions, and data management (conversion, storage, transferring/sharing) schemas, for a PB-scale database. We also engineered a universal application interface (**Fig. 1b** and **Supplementary Fig. 2**) in *MorphoHub* so that it could invoke additional image analysis and validation tools when needed.

To demonstrate the capability of this approach, we built an image database called *D62* consisting of 62 whole mouse brain images. *D62* has in total 713.35 teravoxels, 1.43 petabytes in native image space, and 973 terabytes in compressed file space (**Supplementary Data 1**). *MorphoHub* ran smoothly on *D62* and allowed us to precisely pinpoint somata of 50,233 neurons using TeraFly (100% accuracy validated by independent annotators). For each neuron, we then reconstructed the dendrites automatically (**Fig. 3a**, **Methods**), followed by feature-based screening and visual validation (**Fig. 1c**, **Methods**). Using this workflow, we produced traceable dendritic results of 11,322 neurons. Due to the scale of the problem, similar results were hard to obtain using other software.

For each of the sparsely labeled neurons whose long axon projection could be separated, we first produced an L1-reconstruction corresponding to the key skeleton of a neuron along with its dendrites and axonal projection targets. We then requested human annotators to validate each L1-reconstruction. The resulting L1-reconstruction was then further refined to complete the L2-reconstruction that also added the distal axonal arbors projecting far away across various brain regions. In this study, we focused on reconstructing the full morphology of 1,050 neurons whose somata situate in the thalamus, striatum, and cortical regions of mouse brains (**Methods**). Each L1-L2 pair of the completed neuron reconstructions were validated by at least two annotators. This dataset, called *R1050*, was used to further examine whether or not the core multi-level reconstruction method would make sense.

We compared the L1 and L2 morphology in *R105*0. On average a pair of L1-L2 reconstructions have five to six fold difference in terms of their length and number of branches (**Fig. 2a** and **Fig. 2b**). However, a Sholl analysis[36] indicates that the L1-L2 pairs share branching patterns in dendrites and differ only in the additional axonal regions of L2 (**Supplementary Fig. 3**).

Additionally, 95% of the corresponding L1 and L2 reconstruction-pairs have at least 75% overlap in their projecting target brain-regions (**Fig. 2c**). Finally, we used random-sampling to simulate the potential reconstruction error that would be seen if the L1-L2 leveled protocol were not used (**Methods**). The L1-L2 reconstruction protocol can avoid reconstruction error of a brute-force approach that would reconstruct distal arbors directly without first validating the L1-reconstruction (**Fig. 2d**). In summary, **Fig. 2** shows that reconstructing L1 first without incurring the great complexity in producing the L2 data already offers an efficient way to analyze the key branching structures and the ballpark projection pattern of single neurons. The reconstructions in each level state were also transmitted and stored efficiently and safely.

On the other hand, the rich detail in the axonal arbors of the L2 reconstruction of a neuron as demonstrated in **Fig. 2** also serves as the basis of finer-resolution morphometry. For axonal arbors in *R1050*, we further generated the L3 morphometry by detecting putative synaptic sites (boutons) of neurons at very large scale (**Methods**). Due to the presence of axonal fibers, we are able reduce the conventionally required 3-D blob-segmentation problem for bouton detection to a much simpler 1-D Gaussian model fitting problem, that can be solved in both high accuracy and high speed. We fit a Gaussian kernel to the axonal varicosities and detected 1.9 million boutons (**Fig. 3b-d**). Random inspection of the results indicated that the detection precision was above 95% (**Supplementary Fig. 4**). We also found that nearly 80% of the average spacing of adjacent boutons along an axon ranged from 5 to 20 um (**Fig. 3c**). The L1-L2-L3 trio morphometry produced using this method allows studying the complete distribution of single neurons, their projection, and potential connectivity patterns at whole brain scale. A detailed analysis of the patterns in and statistics of these datasets will be reported elsewhere.

All multi-morphometry data produced were also registered to the Allen Common Coordinate Framework (CCF)[37] to see how the distributions of each data level correlate with others (**Fig. 4**). In this way, various brain regions (the white colored regions), dendrites, axons, and boutons can all be identified (**Fig. 4a-e**). As a result, we output a summary matrix with rows representing neurons, columns representing a unique CCF parcel, and numerical entries expressing, for each neuron and corresponding parcel, the axonal and dendritic length, the number of boutons, and the (binary) presence of soma. Such representation lends itself to highly informative quantitative analyses, such as pairwise probability of directional connection between neurons (dot product of presynaptic axonal values and postsynaptic dendritic values) and projection similarity (arccosine distance between axonal values of two neurons). Fig. 4 shows that the distributions of these neuronal entities do not correlate globally. Instead, they exhibit regional enrichment of which the pattern is hard to observe when only local brain areas are analyzed.

For PB-scale computing, the speed of data I/O for storage and data sharing across networks (Internet or intranets) becomes more critical than in applications at small scale. It is essential to reduce data volume without compromising visualization and analysis of such large data. We observed that an L1-L2-L3 morphology trio of a neuron will always be sparse and that the spine and bouton information in the L3 data could be described using a neighborhood around the neuron skeleton in preceding levels. To utilize this observation, we developed a compact L0 representation of a neuron for effective imaging data management, sharing, and computation (**Fig. 5a**, **Methods**). The key idea is that the L0 data of a neuron represents a tightly bounded image region that covers the L1-L2-L3 trio area. Because dendritic spines typically attach dendritic fiber orthogonally, and axonal boutons scatter along axons, for any given L2 (or L1) skeleton we conveniently extracted

a series of piecewise 512x512x512-voxel 3-D image-tiles around the skeleton to cover all spines and boutons and at the same time to allow fast image file I/O. In this way, an L0 representation identifies an image sub-region that contains all parts of a specific neuron.

For *R1050* reconstructed from *D62*, on average the L0 data generated based on an L1 skeleton contains more than 77% of that generated based on the corresponding L2 reconstruction (**Fig. 5b**). In addition, the L0 data of a neuron typically occupies three orders of magnitude less image volume compared to the whole brain image (**Fig. 5c**). The L0 data of the largest neuron in this work has ~80 gigavoxels, while the mean value and standard deviation of the volume of L0 data of 1,050 fully reconstructed neurons are 6.75 and 5.94 gigavoxels, respectively (**Fig. 5d, Supplementary Table 2**). Practically, even the union of all L0 data of neurons, denoted as Super L0, in a sparsely labeled brain still has 1 to 3 orders of magnitude fewer voxels compared to the total volume of the brain (**Fig. 5c**). In this way, the multi-morphometry framework allows thousands of fold better efficiency in both storing and transferring the essential image data and quantitative shape information of neurons. This utility greatly simplifies the previously challenging data sharing task. Indeed, without accelerated content delivery, currently it is possible to transfer the L0 data in *R1050* between the data production center (SEU-ALLEN) in Asia and one data releasing facility (BICCN Image Library, Pittsburgh supercomputing center) in North America (**Fig. 5d**). This direct data sharing replaced a previous way to bulk ship hard drives containing the massive amount data back-and-forth across continents.

With the L0-L1-L2-L3 quadruple data, we further enhanced the scale and faithfulness of our multi-morphometry produced in two ways. First, since the L0 data has a much smaller volume and thus is much easier to share across network, we developed a collaborative mode that interconnects a number of formerly autonomous TeraFly/TeraVR users to synergistically work on the L0 data directly (**Supplementary Fig. 5**). This method parallelizes the workflow and thus improves the data production rate. The cooperative work of multiple annotators also elevates the faithfulness of the resultant morphometry. Second, we used a deep learning network to learn from the L0-Lx (x = 1,2,3) pair. The trained model was used to detect specific neuronal features, such as neurite skeleton or axonal terminals (**Supplementary Fig. 6**). This process can be repeatedly optimized based on progressively more and more accurate L0-L1-L2-L3 quadruple data. Such automation also increases the data production rate for PB-scale computing.

DISCUSSION

In this study, we demonstrated a robust PB-scale informatics platform to generate large-scale single neuron reconstructions suitable for multi-scale biological analysis. Our approach has several advantages: (1) efficient multi-level production and management of whole-brain neuron reconstructions; (2) conducting morphological analysis and cell typing globally and at multi-resolution; (3) enabling the investigation of the convergence or divergence of neuronal projections by analyzing distribution of neuronal arbors across brain regions; (4) comparison of various neuronal elements and sub-structures with respect to the types of cells. Taken altogether, our whole-brain multi-morphometry approach provides a framework to produce hierarchical datasets that synchronize brain anatomy, single neuron morphology, sub-neuronal structures, and potential pre-synaptic sites, all mapped onto a standardized atlas. Our method will be useful for further

studies of neuronal circuits based on whole-brain imaging, not only for mouse brains but also for other model systems such as monkey brains.

Our work furthers previous effort to use light microscopy (LM) to visualize and detect synapses around neurite tracts labeled by genetic markers[5,6] or antibodies[17], which were limited to partial neuronal structures in local brain regions. In this study we used fMOST data as a showcase and for putative synaptic sites we have focused on axonal boutons. As a generic computational framework, our approach is applicable to various datasets produced with different methods and collected with different imaging modalities.

Of note, our method stores both the original neuronal tracing in the native coordinates of the individual brain specimen, allowing efficient extraction of precise geometric measurements (e.g., synaptic distance from the soma along the axonal path) as well as a registered version of the same morphology mapped to CCF. The registration of the axonal reconstructions is efficiently expressed by augmenting each tracing coordinate, from the center of the soma all the way to the boutons, with the unique identifier of the anatomical parcel in which it is embedded[38]. On the one hand, such compact representation immediately enables real-time computation of potential circuit connectivity[39]. On the other, it provides critical information regarding neuronal identity by encoding its somatic region and quantitative projection targets. This information, together with the specification of essential details regarding the brain specimen and imaging modality, fulfills the recommended requirements for standard metadata description of neuromorphological reconstructions[40]. Thus, our IT infrastructure is seamlessly compliant for effective pipelining with the NeuroMorpho.Org public data sharing platform, ensuring maximum impact through expanded community outreach[41].

Another line of major current effort is to reconstruct dense neurite tracts and synapses, or "connectomes", using electron microscopy (EM) followed by computationally intensive 3D image segmentation and modeling, such as the MICRON project[42] and the Janelia FlyEM project[43]. For mammalian brains, EM-based approaches are still restricted to local brain regions due to their very high resolution and various challenges in sample preparation and imaging. Our method complements the EM-based approach in whole brain scale profiling. Importantly, the lower cost of LM enables one to integrate morphometry information from many brains that is crucial to understand the variability of neurons and their circuits across brains and conditions. In comparison with EM-based reconstruction of individual neurons, our work reconstructs the whole morphology, including (1) soma locations; (2) dendritic arbors, which entails the capacity for and locations of synaptic inputs; (3) axonal trajectories with collateral projections and terminal boutons, which indicate innervation of projection targets. Together, such information is directly relevant to neuronal function.

The multi-level reconstruction approach is being enhanced in various ways. In addition to various workstation/PC clients, virtual reality consoles, super-computing, and big-display walls that are already integrated in our software *MorphoHub*, mobile applications (APPs) for more intelligent and automated neuron tracing are being developed and added onto our software. We are also deploying *MorphoHub* for data servers in the cloud and scaling up the capability for concurrent data serving of distributed users. We hope these engineering efforts would lead to a new globally accessible platform that has potential to bring the current productivity to the next level, especially

addressing challenges in completing neuron morphometry more efficiently, producing more fine-scale morphometry such as synapses with their shapes and statistics, integrating more automation through the use of AI, sharing of imaging data remotely at affordable cost, and international collaboration of neuroanatomists and other interested users.

CONCLUSION

Neuronal morphology is an essential component of cell type identity in the brain and an essential determinant of connectivity and circuit function. Large scale accurate neuronal profiling necessitates advanced methods in computational processing to effectively manage storage and bandwidth for collaborative segmentation and annotation. The data in this study shows our petabyte-scale computing framework is able to provide a solution to modern anatomic workflow requirements that are now demanded for very large-scale morphometry.

DATA AND CODE AVAILABILITY

The whole brain image datasets are released under BICCN's Brain Image Library (BIL) at Pittsburgh Supercomputing Center. The multi-morphometry datasets can be downloaded at https://github.com/SD-Jiang/MorphoHub/releases/tag/v1.0. The software is available at https://github.com/SD-Jiang/MorphoHub.

ACKNOWLEDGMENTS

This work is funded by Southeast University (SEU) to support Open Science collaboration. SEU supported the development of the methods and informatics data management and analysis pipeline at the SEU-Allen Joint Center. We are grateful to the Allen Institute for imaging datasets collected through multiple grant awards from institutes under the National Institutes of Health (NIH), including award number R01EY023173 from The National Eye Institute to H.Z., U01MH105982 from the National Institute of Mental Health and Eunice Kennedy Shriver National Institute of Child Health & Human Development to H.Z., and U19MH114830 from the National Institute Of Mental Health to H.Z. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH and its subsidiary institutes. The Allen Institute affiliated authors wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support. G.A.A. acknowledges NIH grants R01NS39600, U01MH14829, and R01NS86082. We thank Zhi Zhou, Yuanyuan Song, Lulu Yin, Shichen Zhang, Jintao Pan, Yanting Liu, Guodong Hong, Jia Yuan, Yanjun Duan, Yaping Wang, Qiang Ouyang, Zijun Zhao, Wan Wan, Peng Wang, Ping He, Lingsheng Kong, Feng Xiong, and other team mmebers for the support of data production.

AUTHOR CONTRIBUTIONS

H.P. conceptualized the study, envisioned and led the development of this platform and various analyses. S.J., Y.W., and L.L. co-developed the *MorphoHub* system and the multi-level neuron reconstruction protocols. L.L. led the reconstruction of *R1050*. Y.W. provided the support of *TeraVR*. M.C. led the identification of somata in D62. S.Z., X.Z. assisted in *MorphoHub* development. P.X. and L.D. assisted in data analysis. Z.R. and H.P. constructed the hardware platform for this study. H.Z. provided the raw imaging dataset for producing D62 and advised on neurobiology. M.H. advised on morphometry and data management. H.D. and G.A. advised on whole brain neuronal structure analysis. H.P. led the writing of the manuscript in consultation with all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Adkins, R. S. et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.19.343129 (2020).
2. 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.e13 (2019).
3. Gong, H. et al. Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. *Neuroimage* **74**, 87–98 (2013).
4. Peng, H. et al. Brain-wide single neuron reconstruction reveals morphological diversity in molecularly defined striatal, thalamic, cortical and claustral neuron types. Preprint at *bioRxiv* https://doi.org/10.1101/675280 (2020).
5. Iascone, D. M. et al. Whole-Neuron Synaptic Mapping Reveals Spatially Precise Excitatory/Inhibitory Balance Limiting Dendritic and Somatic Spiking. *Neuron* **106**, 566-578.e8 (2020).
6. Kim, J. et al. mGRASP enables mapping mammalian synaptic connectivity with light microscopy. *Nature Methods* **9**, 96–102 (2012).
7. Parekh, R. & Ascoli, G. A. Neuronal Morphology goes Digital: A Research Hub for Cellular and System Neuroscience. *Neuron* **77**, 1017–1038 (2013).
8. Ropireddy, D., Scorcioni, R., Lasher, B., Buzsáki, G. & Ascoli, G. A. Axonal morphometry of hippocampal pyramidal neurons semi-automatically reconstructed after in vivo labeling in different CA3 locations. *Brain Struct Funct* **216**, 1–15 (2011).
9. Gong, H. et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nature Communications* **7**, 12142 (2016).
10. Economo, M. N. et al. A platform for brain-wide imaging and reconstruction of individual neurons. *eLife* **5**, e10566 https://doi.org/10.7554/eLife.10566 (2016).
11. Zeng, H. Mesoscale Connectomics. *Curr Opin Neurobiol* **50**, 154–162 (2018).
12. Zingg, B. et al. Neural Networks of the Mouse Neocortex. *Cell* **156**, 1096–1111 (2014).
13. Rotolo, T., Smallwood, P. M., Williams, J. & Nathans, J. Genetically-Directed, Cell Type-Specific Sparse Labeling for the Analysis of Neuronal Morphology. *PLoS ONE* **3**, e4099 (2008).

14. Aransay, A., Rodríguez-López, C., García-Amado, M., Clascá, F. & Prensa, L. Long-range projection neurons of the mouse ventral tegmental area: a single-cell axon tracing analysis. *Front. Neuroanat.* **9**, (2015).
15. Karube, F., Kubota, Y. & Kawaguchi, Y. Axon branching and synaptic bouton phenotypes in GABAergic nonpyramidal cell subtypes. *J Neurosci* **24**, 2853–2865 (2004).
16. Sivagnanam, S. et al. Early experiences in developing and managing the neuroscience gateway. *Concurrency and Computation: Practice and Experience* **27**, 473–488 (2015).
17. Micheva, K. D. & Smith, S. J. Array tomography: a new tool for imaging the molecular architecture and ultrastructure of neural circuits. *Neuron* **55**, 25–36 (2007).
18. Swedlow, J. R., Goldberg, I., Brauner, E. & Sorger, P. K. Informatics and quantitative analysis in biological imaging. *Science* **300**, 100–102 (2003).
19. Myers, G. Why bioimage informatics matters. *Nature Methods* **9**, 659–660 (2012).
20. Eliceiri, K. W. et al. Biological imaging software tools. *Nature Methods* **9**, 697–710 (2012).
21. Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. & Olivo-Marin, J.-C. Imagining the future of bioimage analysis. *Nature Biotechnology* **34**, 1250–1255 (2016).
22. Cohen, A. R., Roysam, B. & Turner, J. N. Automated tracing and volume measurements of neurons from 3-D confocal fluorescence microscopy data. *Journal of Microscopy* **173**, 103–114 (1994).
23. Acciai, L., Soda, P. & Iannello, G. Automated Neuron Tracing Methods: An Updated Account. *Neuroinformatics* **14**, 353–367 (2016).
24. Cheng, S. et al. DeepBouton: Automated Identification of Single-Neuron Axonal Boutons at the Brain-Wide Scale. *Front. Neuroinform.* **13**, (2019).
25. Bass, C., Helkkula, P., Paola, V. D., Clopath, C. & Bharath, A. A. Detection of axonal synapses in 3D two-photon images. *PLoS ONE* **12**, e0183309 (2017).
26. Gala, R. et al. Computer assisted detection of axonal bouton structural plasticity in in vivo time-lapse images. *eLife* **6**, e29315 https://doi.org/10.7554/eLife.29315 (2017).
27. Yan, C. et al. Automated and Accurate Detection of Soma Location and Surface Morphology in Large-Scale 3D Neuron Images. *PLoS ONE* **8**, e62579 (2013).
28. 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).
29. Gillette, T., Brown, K., Svoboda, K., Liu, Y. & Ascoli, G. DIADEMchallenge.Org: A Compendium of Resources Fostering the Continuous Development of Automated Neuronal Reconstruction. *Neuroinformatics* **9**, (2011).
30. Peng, H. et al. BigNeuron: Large-scale 3D Neuron Reconstruction from Optical Microscopy Images. *Neuron* **87**, 252–256 (2015).
31. Bria, A., Iannello, G., Onofri, L. & Peng, H. TeraFly: real-time three-dimensional visualization and annotation of terabytes of multidimensional volumetric images. *Nature Methods* **13**, 192–194 (2016).
32. Peng, H. et al. Automatic tracing of ultra-volumes of neuronal images. *Nature Methods* **14**, 332–333 (2017).
33. Pietzsch, T., Saalfeld, S., Preibisch, S. & Tomancak, P. BigDataViewer: visualization and processing for large image data sets. *Nature Methods* **12**, 481–483 (2015).
34. Wang, Y. et al. TeraVR empowers precise reconstruction of complete 3-D neuronal morphology in the whole brain. *Nature Communications* **10**, 3474 (2019).
35. Markram, H. The Blue Brain Project. *Nature Reviews Neuroscience* **7**, 153–160 (2006).

36. Langhammer, C. G. et al. Automated Sholl analysis of digitized neuronal morphology at multiple scales: Whole cell Sholl analysis versus Sholl analysis of arbor subregions. *Cytometry A* **77**, 1160–1168 (2010).
37. Wang, Q. et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181**, 936-953.e20 (2020).
38. Nanda, S. et al. Design and implementation of multi-signal and time-varying neural reconstructions. *Scientific Data* **5**, 170207 (2018).
39. Rees, C. L., Moradi, K. & Ascoli, G. A. Weighing the Evidence in Peters' Rule: Does Neuronal Morphology Predict Connectivity? *Trends in Neurosciences* **40**, 63–71 (2017).
40. Bijari, K., Akram, M. A. & Ascoli, G. A. An open-source framework for neuroscience metadata management applied to digital reconstructions of neuronal morphology. *Brain Inf.* **7**, 2 (2020).
41. Ascoli, G. A., Maraver, P., Nanda, S., Polavaram, S. & Armañanzas, R. Win–win data sharing in neuroscience. *Nature Methods* **14**, 112–116 (2017).
42. Yin, W. et al. A petascale automated imaging pipeline for mapping neuronal circuits with high-throughput transmission electron microscopy. *Nature Communications* **11**, 4949 (2020).
43. Xu, C. S. et al. A Connectome of the Adult Drosophila Central Brain. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.21.911859 (2020).
44. Peng, H., Bria, A., Zhou, Z., Iannello, G. & Long, F. Extensible visualization and analysis for multidimensional images using Vaa3D. *Nature Protocols* **9**, 193–208 (2014).
45. Peng, H., Ruan, Z., Long, F., Simpson, J. H. & Myers, E. W. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nature Biotechnology* **28**, 348–353 (2010).
46. 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).

METHODS

Hardware and software platform

Our system is designed for the task of multi-morphometry data generation from petabyte-scale whole-brain imaging database. Built upon customized hardware infrastructure and software tools, the system is capable of organizing whole-brain imaging datasets, generating multi-morphometry data, managing data and workflow, and visualization and analysis of the generated data (**Supplementary Figure 1**).

The hardware infrastructure includes the following parts. VR-equipped annotation workstations are used for data visualization, interactive neuron reconstruction, proofreading, etc. A petabyte-scale storage is configured to store the whole-brain imaging datasets, while the multi-morphometry data is managed using cloud-based storage. A computing cluster is deployed for parallel execution of batch work assignments. Moreover, a wall-mounted display array is also available for monitoring the data generation status. The storage server and the computing cluster are connected with a 100 Gbps wired local area network for peta-scale data storage and parallel computing. The annotation workstations, cloud storage, and monitor system reside in a 10 Gbps local area network.

There are four major software components in the *MorphoHub* software package. The *MorphoHub*-DBMS is the core of the entire system which is responsible for the coordination of the overall data generation process. The DendriteGenerator is in charge for parallel generation of dendritic arbors. The L0Generator is useful for creating compact image representation of the reconstructions. The BoutonGenerator is capable of automatically detection of the synaptic boutons located on axons. These components are developed as plugins of Vaa3D[44,45], thus making *MorphoHub* cross-platform and deployable on various operating systems.

MorphoHub-DBMS

The *MorphoHub*-DBMS manages the generation of multi-morphometry data by either collaborating teams or automatic routines. All the morphometry data files are organized in the *MorphoHub*-Database, where they follow consistent naming rule and are assigned with unique IDs. The validity of the database is monitored by an error checking routine running in the background. Should there be any issue regarding to the morphometry data files, and error message would be generated and displayed on the screen wall system.

During the multi-level generation of the neuron morphometry data, multiple brains, neurons, annotators are involved. Also, for each neuron, the L1, L2, and L3 data, together with a number of intermediate data levels are produced. The *MorphoHub*-DBMS maintains the correct working state for all the neurons and supports a number of operations, including commit (submitting a reconstruction to a higher level), rollback (sending a reconstruction to a previous level), proofreading (requesting for quality check of the reconstruction), etc. Besides, the *MorphoHub*-DBMS also provides supports for user management and task assignment. All the data are periodically synchronized with private or public cloud-based storages for version control, data sharing and collaboration.

A screen wall display system is connected with *MorphoHub* to present useful information for the ongoing morphometry generation. On the display array, some of the latest reconstructed neurons are displayed; the status of the data repository is tracked; preliminary analysis results are also generated dynamically to provide potential insights of the data.

Generation of somata and dendritic arbors (DendriteGenerator)

The annotators browsed each whole brain image in *D62* using TeraFly and pinpointed soma locations. 2-D maximum intensity projection (MIP) images were generated for each soma-centered region for further validation. Then, centering at each identified soma, a local image volume of 1024x1024x512 was cropped from the from highest resolution of whole brain image. Next, the APP2[46] algorithm was invoked for the tracing of dendritic arbors. In particular, a number of background thresholds (15, 20, …, 40) were adopted for each tracing routine. The tasks were submitted to the cluster server for parallel computation. The results were retrieved and stored only if the execution time was under 30 seconds. Then, we leveraged the gold-standard datasets, e.g., a set of manually annotated and validated dendritic arbors, to form rules for further screening the automatic reconstruction results. The [min, max] interval of the following five features of the dendritic arbors were considered, including 'Tips', 'Length', 'Max Path Distance', 'Average Bifurcation Angle Remote', and 'Max Branch Order'. An automatic tracing result qualified if more than four features conformed with the gold-standard. In case more than one tracing qualified for a certain soma location, the result with larger overall tracing length was selected. In visual screening, tracing results were removed when multiple cells were connected.

Generation of L1 and L2 data

In *MorphoHub*, we use a multi-level approach to reconstruct neuron morphology: the L1 reconstruction contains the dendritic trees and the long axonal projections, while the structures of distal axonal arbors are annotated in the L2 reconstruction. The generation of either L1 or L2 data requires several rounds of checking and correction and is essentially an iterative procedure. In each iteration, there is a generation step (GS) and a validation step (VS) (**Supplementary Fig. 7**). In the generation step, an annotator tries to reconstruct the neuron's morphology until the person considers that the reconstruction meets the standard of the current level, i.e., L1 or L2. Then, in the validation step, a second annotator examines the reconstruction while labeling the over-traced structures (false positive, FP) and missing structures (false negative, FN) and confirming the correct structures (true positive, TP). After that, the precision rate (P= TP/ (TP+ FP)) and recall rate (R= TP/ (TP + FN)) can be calculated. If the F1 score = (2*RP/ (R+P)) is greater than the preset threshold, the current level is considered finalized. Otherwise, another iteration is needed. Normally, the reconstructions of both L1 and L2 converge in two iterations.

Calculation of the overall reconstruction accuracy in random-sampling simulation

The reconstruction of a neuron's morphology, normally starting from the soma and gradually extending all the way to the remote terminals, is essentially a consecutive decision-making process. Thus, it is likely to suffer from the problem of error propagation, i.e., the reconstruction errors at upstream structures will affect the downstream structures. We assume that for each primitive structure in neuron reconstruction, e.g., a neuronal segment, there is a constant probability p for error occurrence. Besides, if error occurs at some structure, it will also propagate to all its child structures. During the validation step, reconstruction errors could be identified and corrected. However, new errors would still be likely to be introduced at the given chance. Based on such

assumptions, we could calculate the overall reconstruction accuracy after *n* rounds of iterations. With the L1-L2 leveled protocol, we assign *n/2* iterations for L1 reconstructions, and other *n/2* iterations for L2. With the one-level strategy, we simply repeat the generation-validation steps *n* times for the reconstruction of whole neuron. The simulation can be carried out several times to achieve stable results.

L3 (bouton) generation (BoutonGenerator)

As a typical L3 data, the bouton distributions are generated based on the guidance of L2 axonal arbors. Boutons are mainly located at distal axonal arbors rather than the long axonal projections. We used a four-step approach to extract L3 (bouton) data automatically. Firstly, the neuronal tree representation of L2 data is resampled using a fixed-length interval. In this work, the interval is set to be 4 microns. Secondly, for each node in the neuronal tree, the corresponding image intensity value is retrieved from the whole brain datasets. Since the nodes cannot be guaranteed to locate at the centers of the putative boutons, the nodes are allowed to be locally shifted to the maximum intensity position within a small area (e.g., 2 voxels). We assume that the intensity of imaging signal along axons obeys the 1-D Gaussian distribution and a bouton site tends to have intensity jump compared with its neighboring nodes. Thus, the third step is to calculate the intensity jump threshold for each small image block (e.g., 128x128x123) as the standard deviation of the block, and extract bouton candidate in a block-wise manner. Finally, in the last step, we remove any possibly duplicated bouton site if it is too close to its neighboring site (e.g., within 5 voxels).

Generation of the L0 imaging data (L0Generator)

The L0 imaging data is generated based on the corresponding morphometry data, e.g., L1 data, L2 data, or even a dendritic arbor. The L0 data contains the image regions that cover all the anatomical structures of the morphometry and is organized in a TeraFly-compatible hierarchical form, just as the whole-brain imaging data. The approach to generate L0 data is described below. For a given neuron, each fundamental morphological element, i.e., nodes and edges, is examined. The local image block in the whole-brain dataset that contains the element is found and marked as "relevant". Then, all the "relevant" images are combined into the union set from which the compact L0 data are finally generated by building a multi-resolution image hierarchy.

Quality control workflow and public release

The reconstructions are concurred to be released after following the single-tree criterion, which includes correct types, no loops, and no trifurcation or multi-furcation. Combination of manual modifications and automatic algorithms were used to control the quality of reconstructions. Automatic routines were invoked to detect any presence of gaps, loop and multi-furcation, followed by manual correction of such issues. Other procedures include examining the reconstruction quality at branch terminals and checking whether all the neurites are centered at image signals. After 2 to 3 rounds of checking, the reconstructions are then ready for ingestion and mirroring in open-source repositories such as NeuroMorpho.Org (http://neuromorpho.org/).

FIGURES

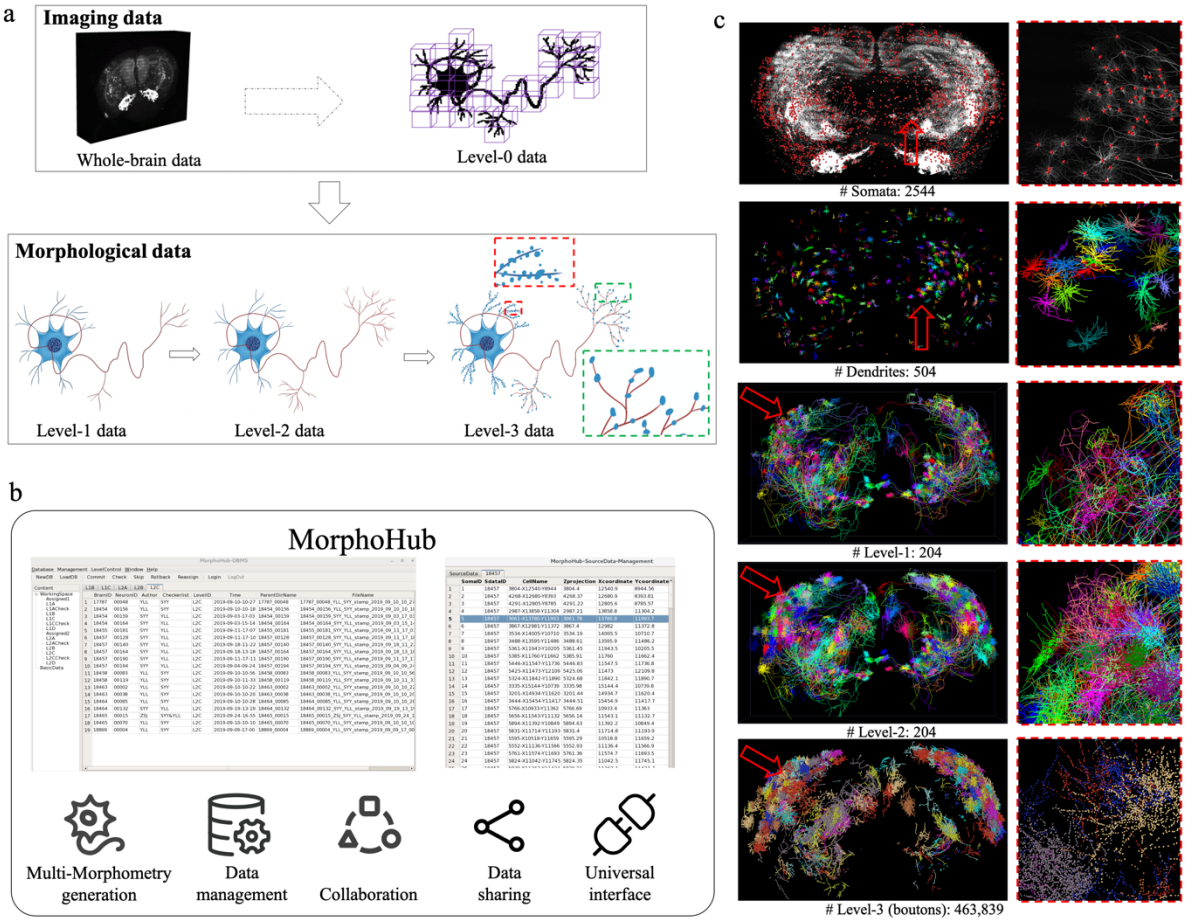


Fig 1. Multi-morphometry data generation from whole-brain imaging datasets. ***a*** *An illustration of the multi-level reconstruction approach. From a whole-brain image containing trillions of voxels (top left), the Level-1 (L1), Level-2 (L2), and Level-3 (L3) data are reconstructed in sequence (bottom). Moreover, a concise Level-0 (L0) imaging data is also generated based on the reconstructed morphometry (top right).* ***b*** *The MorphoHub system for the generation of multi-morphometry data, management and visualization of all related data and workflow, data sharing and extended functions.* ***c*** *Examples of the multi-morphometry data reconstructed from one Brain (Brain id: 18454). From top to bottom are the somata, dendrites, L1, L2, and L3 data, respectively, with zoom-in panels for red arrows shown on the right.*

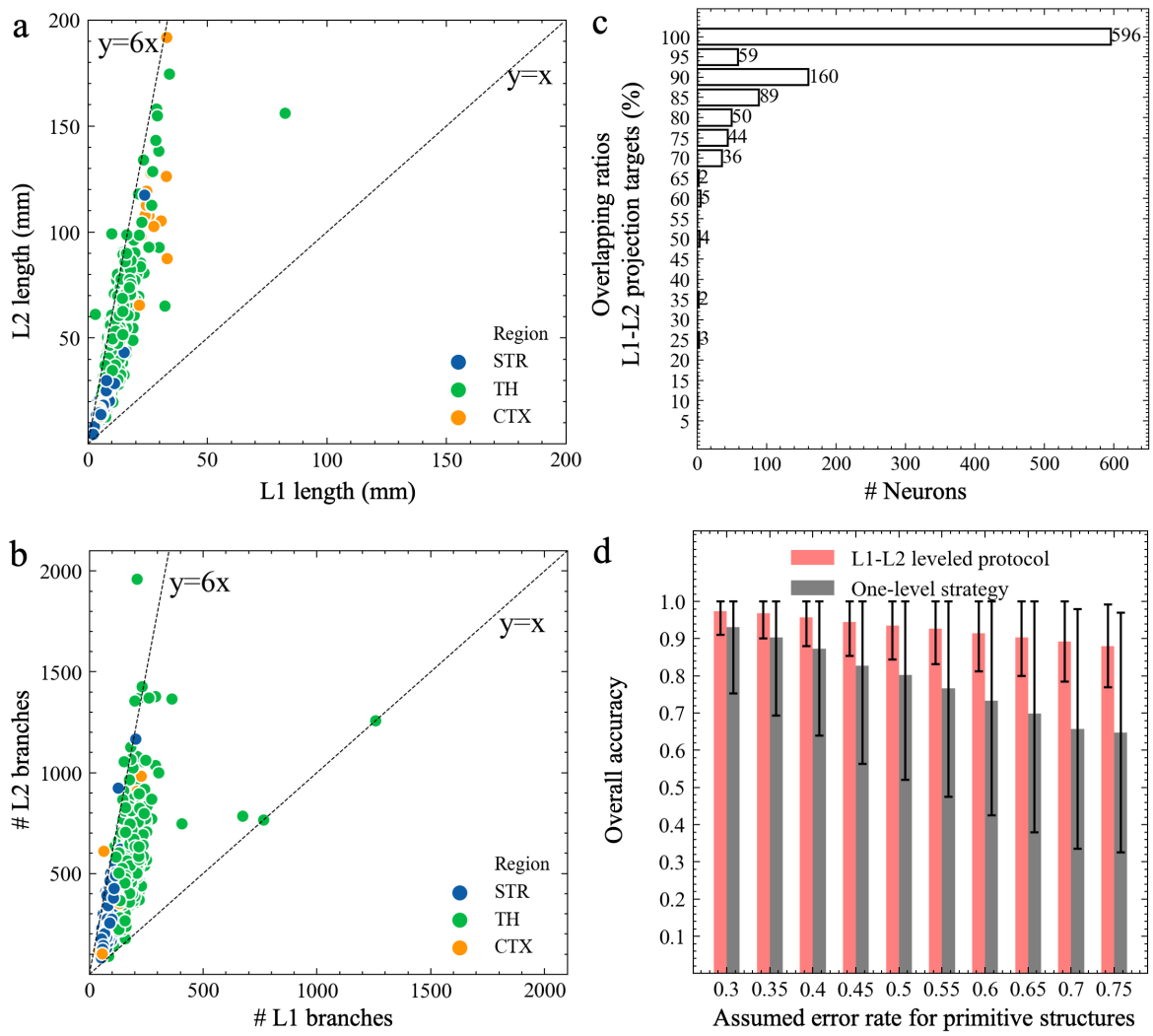


Fig 2. Comparisons of L1 and L2 data. ***a*** *and* ***b*** *Two features (****a****: total length;* ***b****: branch number) of L1 and L2 data. Neurons are from three brain regions (TH (Thalamus): 705 neurons; CTX (Cortex): 23 neurons; STR (Striatum): 322 neurons).* ***c*** *Comparison of the projection targets of the L1-L2 pairs. Horizontal axis: number of neurons. Vertical axis: overlapping ratios of projection targets of L1-L2 pairs.* ***d*** *Random-sampling simulation of potential reconstruction errors of the L1-L2 leveled protocol and the brute-force one-level strategy. Horizontal axis: the assumed error rate for each primitive structure (e.g., a tract). Vertical axis: overall accuracy of the reconstructions after validation. Error bar: SD.*

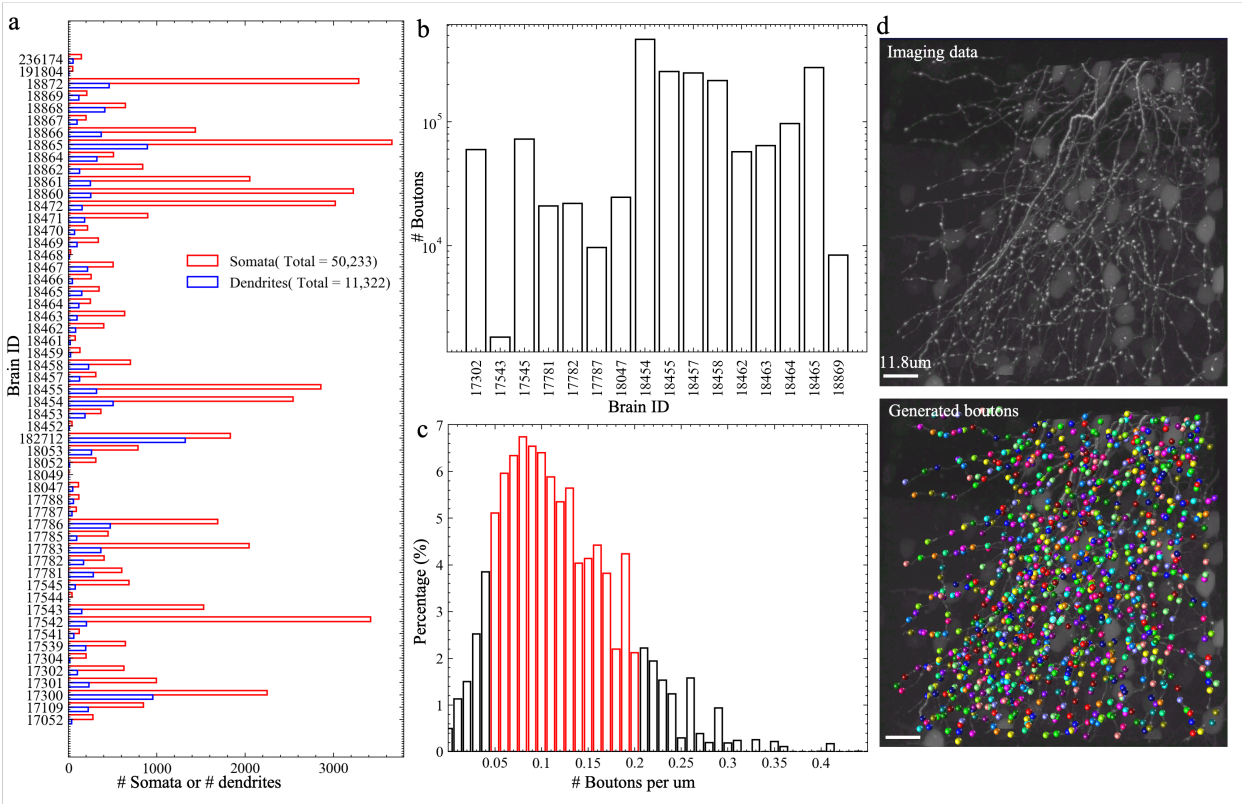


Fig 3. Mass data production of somata, dendrites, and boutons. ***a*** *Number of identified somata and automatically reconstructed dendritic arbors from D62.* ***b*** *Number of detected boutons based on the L2 morphometry in R1050.* ***c*** *Distribution of the linear density of boutons along the axonal skeletons. Red bar highlights the range [0.05, 0.2].* ***d*** *Examples of detected boutons. Upper: a local image region containing an axonal cluster. Bottom: putative boutons shown in distinct colors.*

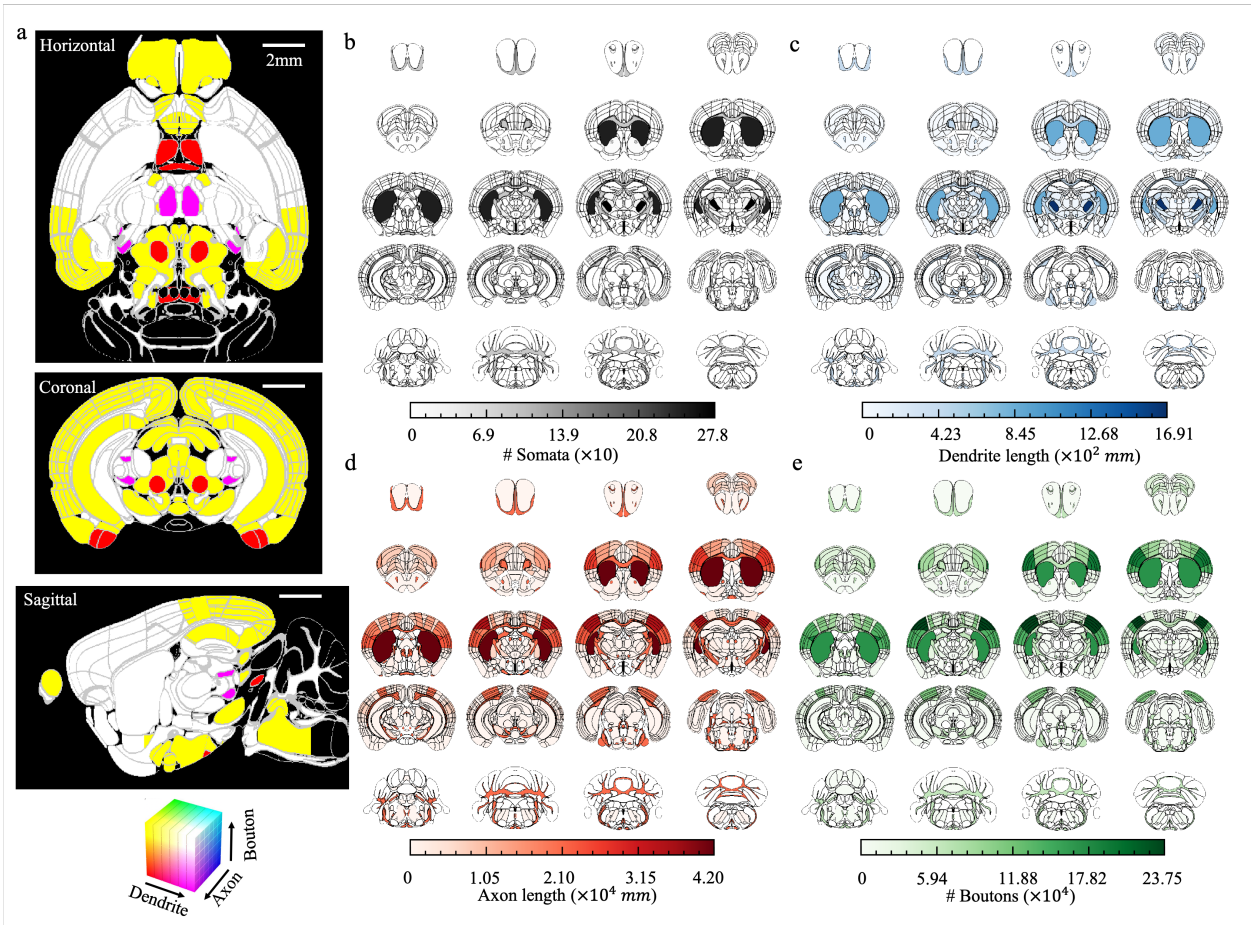


Fig 4. Visualization of the multi-morphometry data in R1050. ***a*** *The color-coded joint distribution of dendrites, axons, and boutons. From top to bottom: horizontal view (slice no. 165), coronal view (slice no. 335), and sagittal view (slice no. 148). Colors indicate the densities of dendrites, axons, and boutons normalized to the standard RGB color space. Scale bar: 2mm.* ***b*** *Individual distributions of somata in R1050. Each inset corresponds to the combination of 25 consecutive coronal slices, in which the brain regions were colored according to the densities of somata. The darker the color, the higher the density.* ***c-e*** *Similar visualizations for dendrites, axons, and boutons in R1050, respectively.*

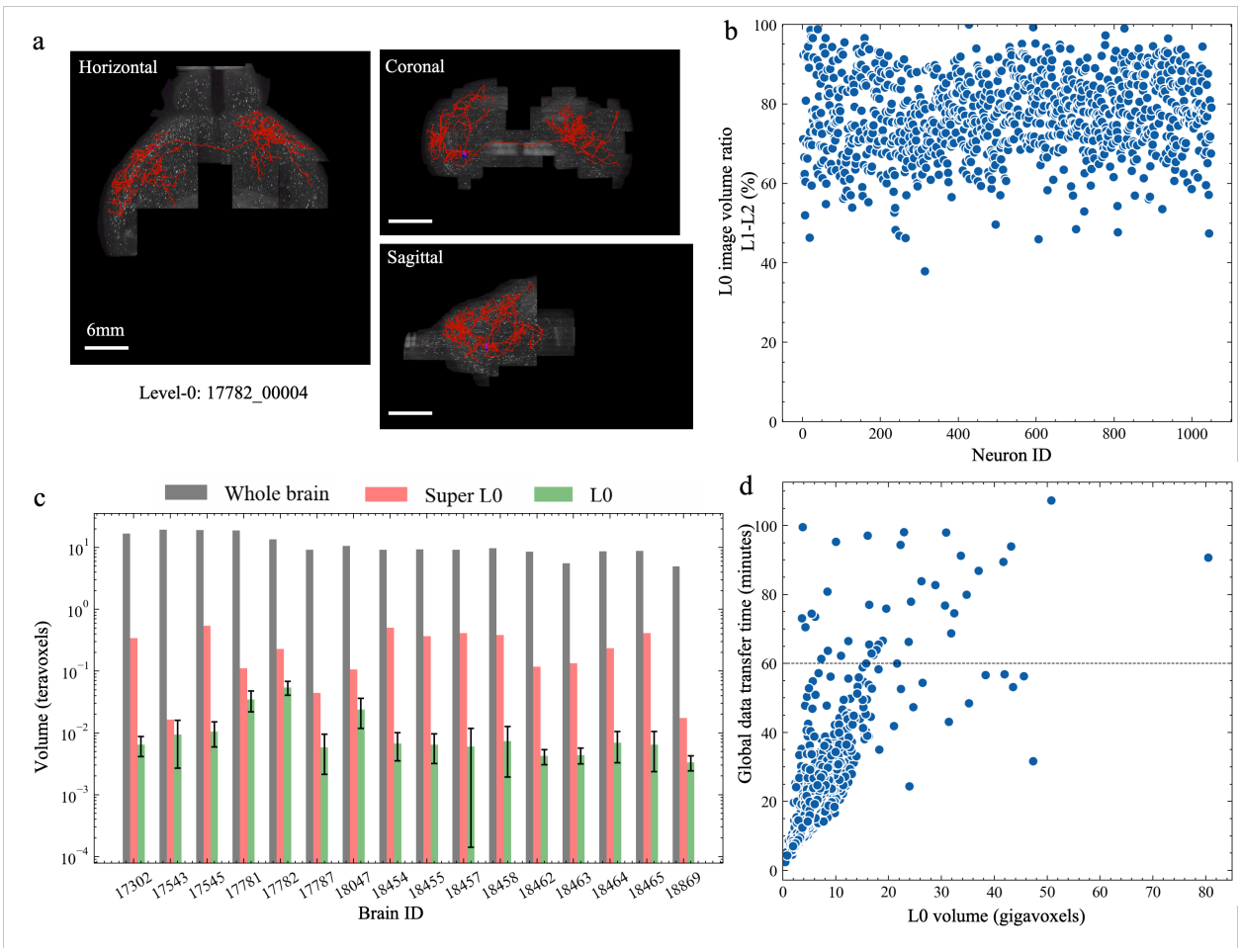


Fig 5. The L0 representation of imaging data. ***a*** *The L0 image (shown in horizontal, coronal, and sagittal views) for neuron 17782_00004, overlaid with its L2 reconstruction.* ***b*** *For neurons in R1050, the ratio of L0 image volume generated from L1 data over that generated from L2 data.* ***c*** *Comparisons of the size of whole-brain images, the average size of L0 data, and the size of the "super L0-data" (union of all L0 data of neurons). Error bar: SD.* ***d*** *Time for transferring 1050 L0 images between two research centers in Asia (SEU-ALLEN) and America (BIL).*

SUPPLEMENTARY TABLES

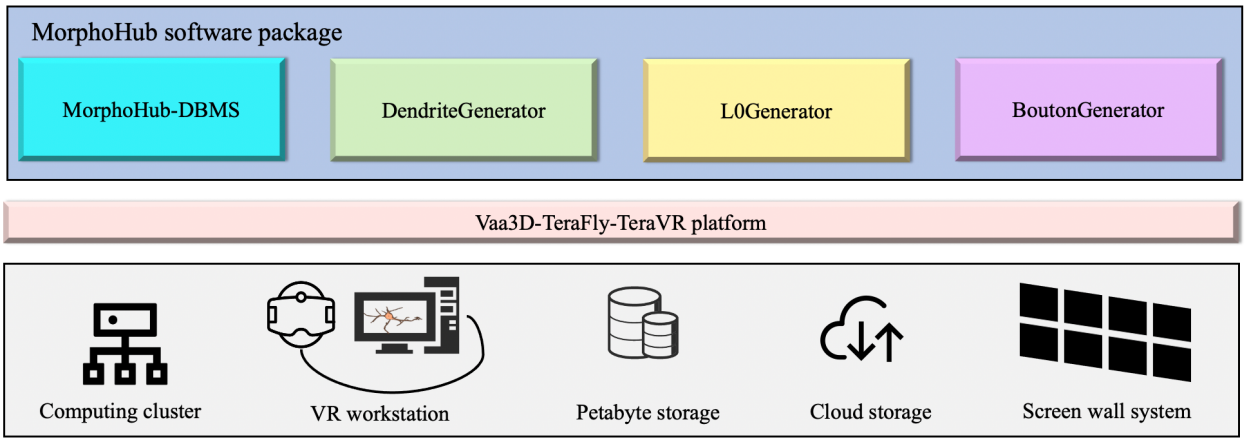
Supplementary Table 1. Comparisons between MorphoHub and relevant software for whole mouse-brain imaging data

	TeraFly / TeraVR	UltraTracer	BigDataViewer	MorphoHub
Whole-brain data handling	Yes, handle whole-brain data	No	No	Yes, handle PB-scale whole-brain datasets
Multi-level neuron reconstruction	No	No	No	Yes
Automatic dendrite tracing	No	Yes	No	Yes, in massive scale
Automatic bouton generation	No	No	No	Yes
Compact imaging data sharing	No	No	No	Yes
Realtime visualization and analysis	No	No	No	Yes

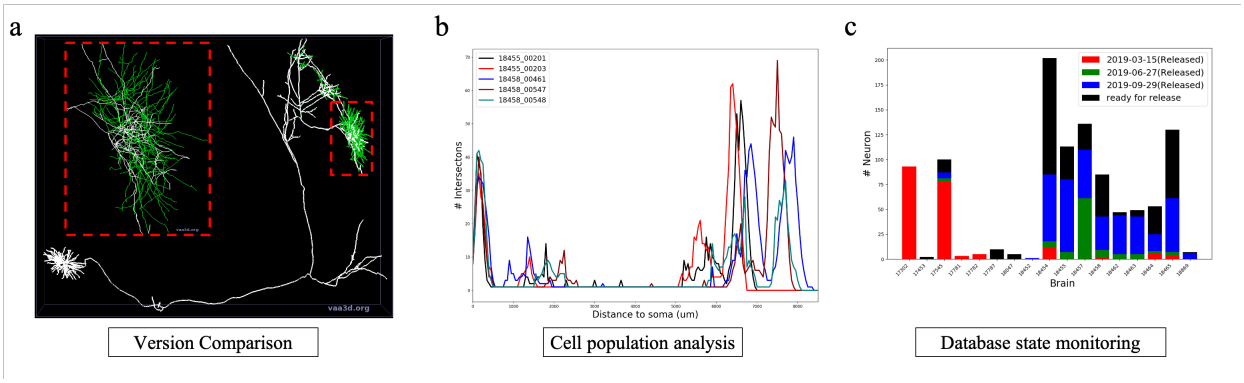
Supplementary Table 2. The statistics of L1 and L2 reconstructions.

Level \ Features	Length (mm)						#Voxels of L0 (giga)	
	Total	Average	Standard Deviation	Minimal	Maximum	Median	Average	Standard Deviation
Level-1	9964.96	9.49	5.12	1.88	82.55	8.14	5.29	3.75
Level-2	39233.88	37.36	22.60	3.97	191.76	31.91	6.75	5.94

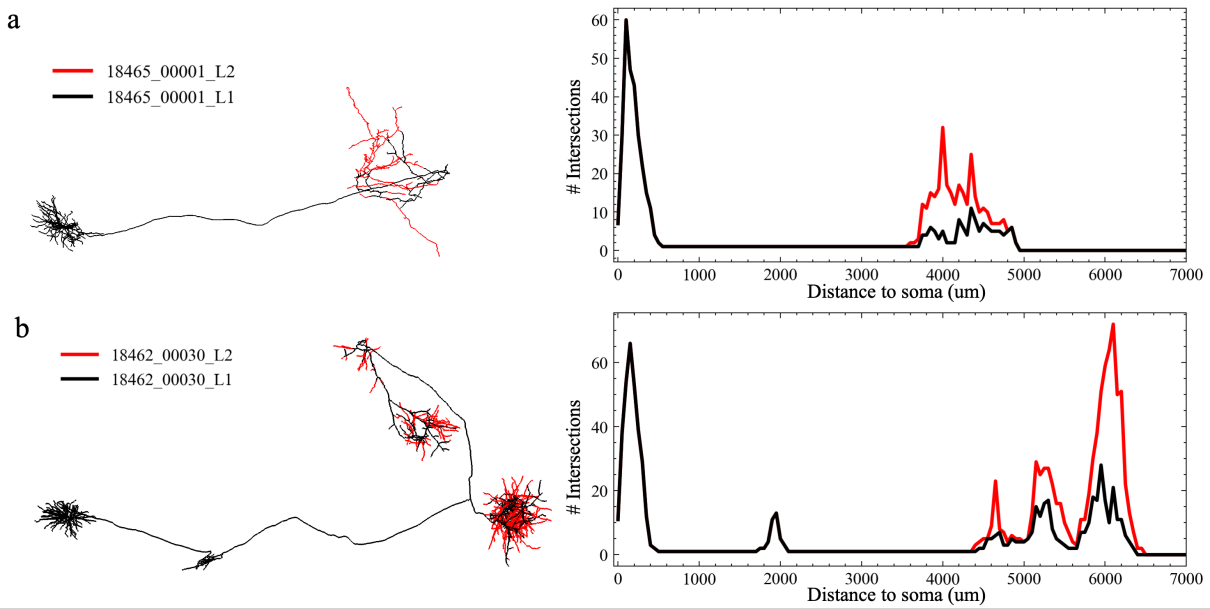
SUPPLEMENTARY FIGURES



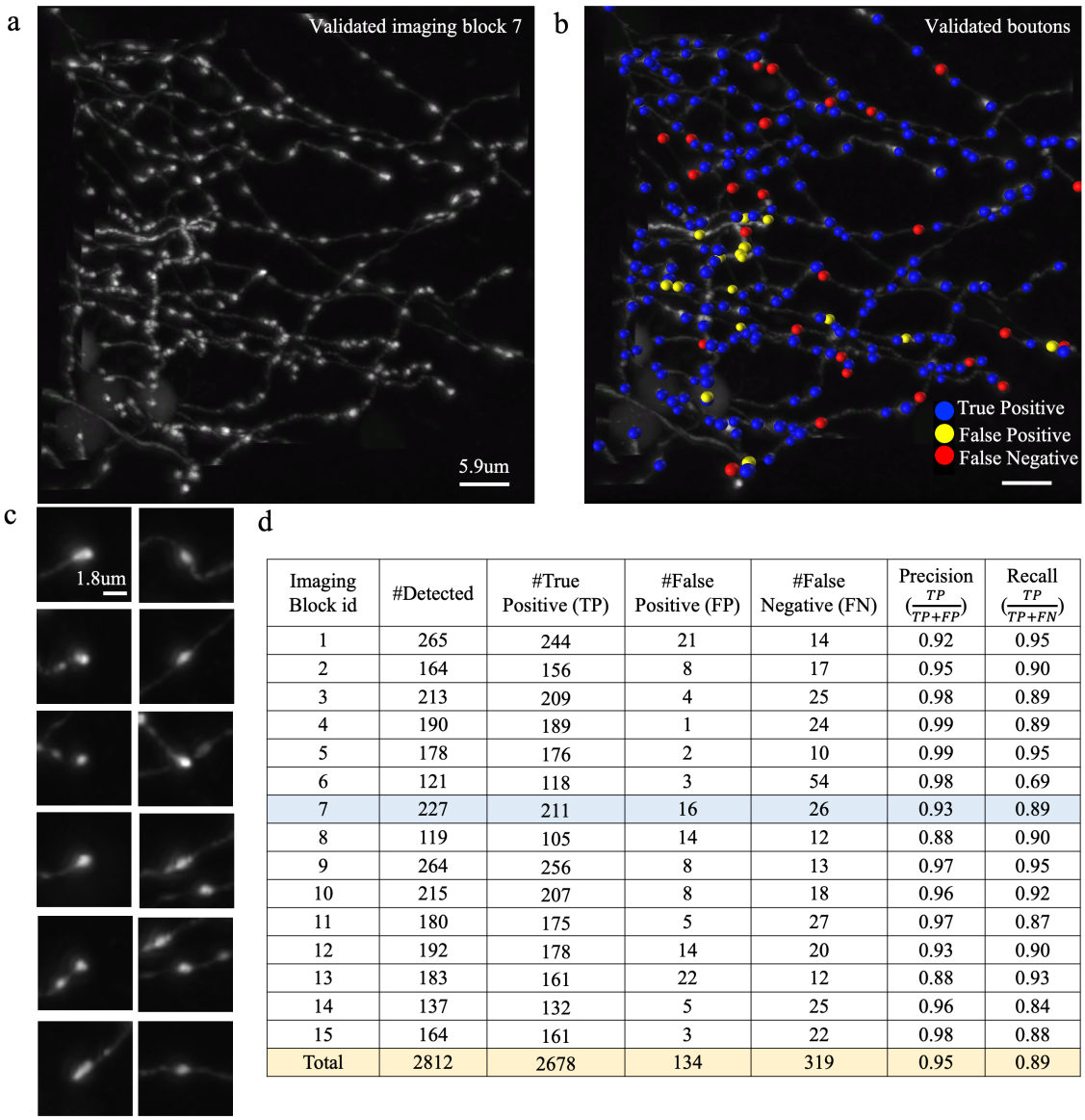
Supplementary Figure 1. An illustration of our system. A petabyte hardware platform is constructed (bottom row), upon which the MorphoHub software package is developed (upper row). There are four major modules in MorphoHub, which are the MorphoHub-DBMS, DendriteGenerator, L0Generator, and BoutonGenerator, respectively.



Supplementary Figure 2. Extended functions via the universal application interface of MorphoHub. ***a*** *Version control tool for comparing the structural difference of two reconstruction versions. (white: identical structures, green: different structures).* ***b*** *Sholl analysis for neurons of the same cell type.* ***c*** *Monitoring of the morphometry data production. The numbers of generated neurons are shown for each brain and each release date.*

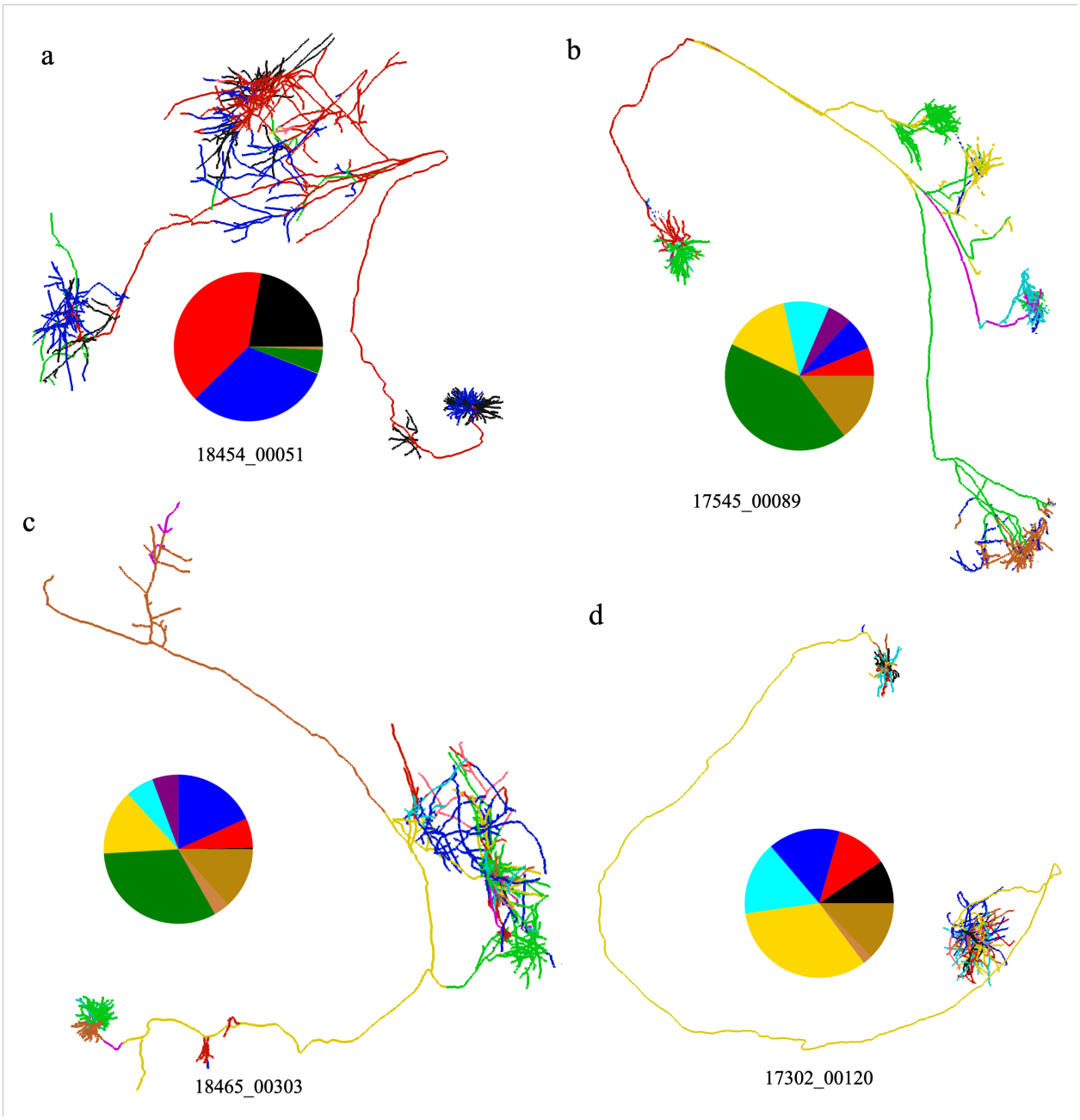


Supplementary Figure 3. Comparison between the L1-L2 data pairs. ***a*** *Left: a neuron (18465_00001) with one axonal cluster. Neurites in black are the common parts for both L1 and L2; neurites in red belong to L2 only. Right: Sholl analysis of the L1 and L2 data.* ***b*** *A neuron (18462_00030) with two axonal clusters.*

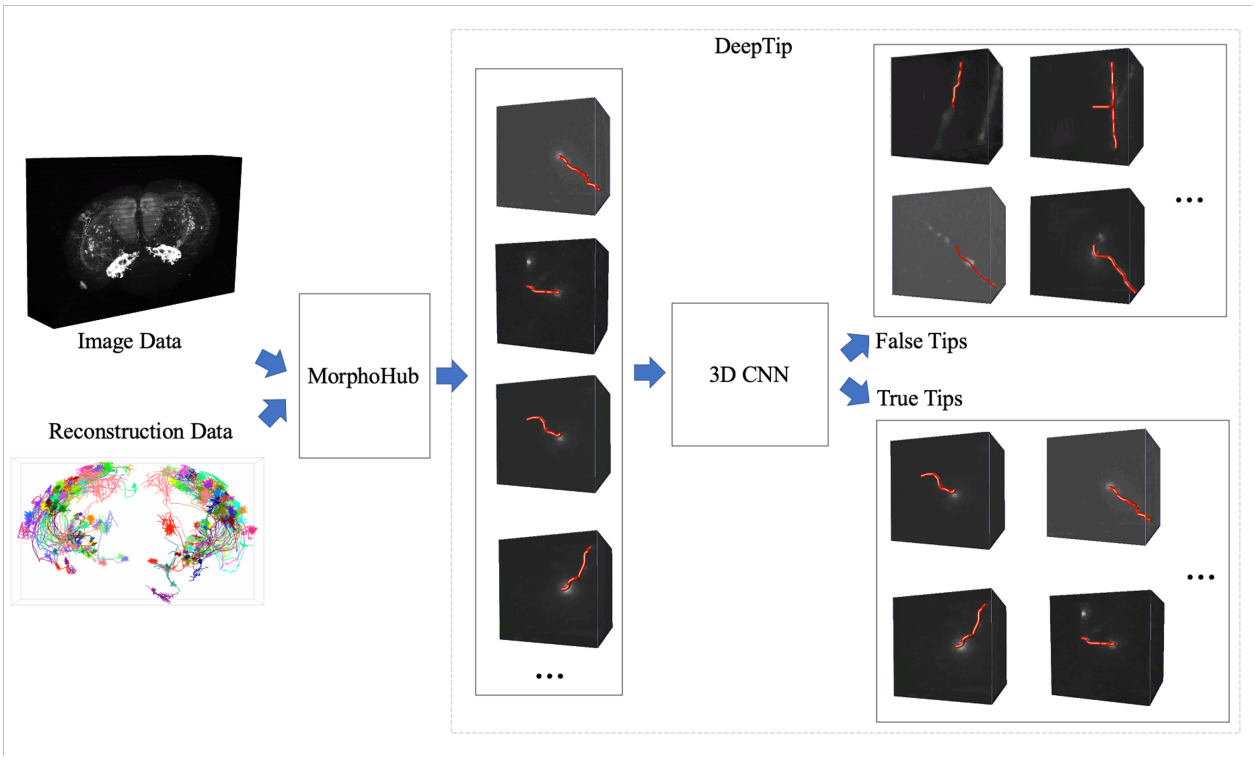


Imaging Block id	#Detected	#True Positive (TP)	#False Positive (FP)	#False Negative (FN)	Precision $(\frac{TP}{TP+FP})$	Recall $(\frac{TP}{TP+FN})$
1	265	244	21	14	0.92	0.95
2	164	156	8	17	0.95	0.90
3	213	209	4	25	0.98	0.89
4	190	189	1	24	0.99	0.89
5	178	176	2	10	0.99	0.95
6	121	118	3	54	0.98	0.69
7	227	211	16	26	0.93	0.89
8	119	105	14	12	0.88	0.90
9	264	256	8	13	0.97	0.95
10	215	207	8	18	0.96	0.92
11	180	175	5	27	0.97	0.87
12	192	178	14	20	0.93	0.90
13	183	161	22	12	0.88	0.93
14	137	132	5	25	0.96	0.84
15	164	161	3	22	0.98	0.88
Total	2812	2678	134	319	0.95	0.89

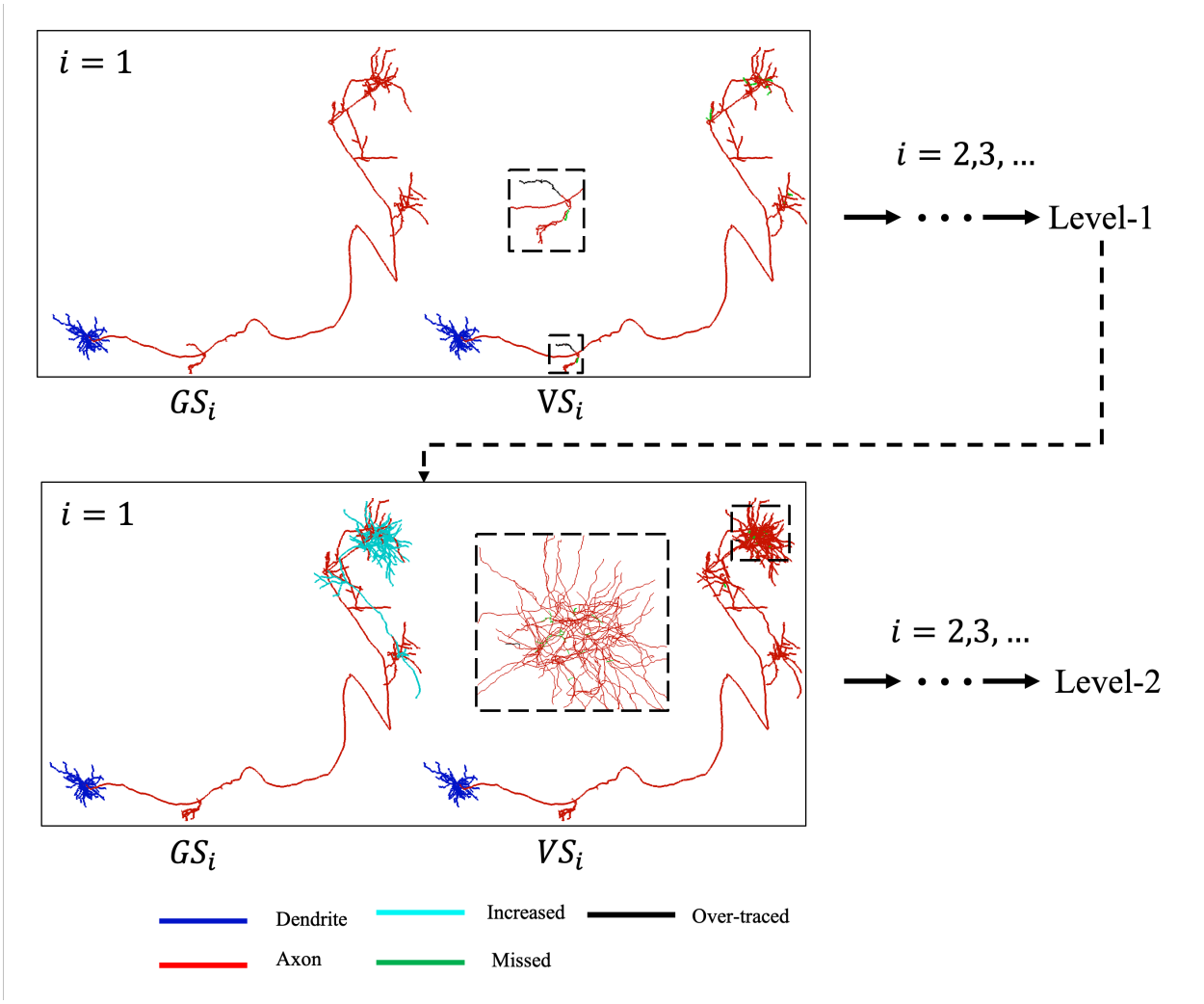
Supplementary Figure 4. Validation of the bouton detection method. ***a*** *A local image volume (id 7) containing axonal boutons.* ***b*** *The automatically detected boutons. The detected boutons were compared with manual annotations and were marked by blue (true positive), yellow (false positive), or red (false negative) spheres. The size of the spheres indicates the level of confidence for detection.* ***c*** *Close-up views of the detected terminal-boutons (first column) and intermediate-boutons (second column). The cropped images were centered at the boutons.* ***d*** *A list of validation results involving 15 image blocks and 2,812 detected boutons.*



Supplementary Figure 5. Collaborative neuron reconstruction among several TeraFly/TeraVR clients. ***a-d****. L2 neurons that are reconstructed and proofread by a group of annotators. The portion reconstructed by each annotator is represented by a different color. The pie chart corresponds to the reconstruction length of distinct annotators.*



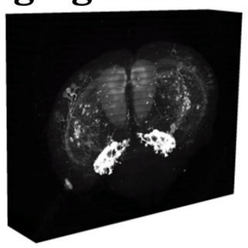
Supplementary Figure 6. 3D-CNN based tip detection. MorphoHub is compatible for integrating various AI components to assist with data production, proofreading and analysis. For example, DeepTip is a deep learning-based component that is integrated in MorphoHub for differentiating true terminal tips of neurites and even automatically correcting the false ones.



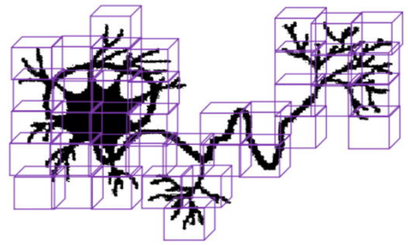
Supplementary Figure 7. The generation of L1/L2 data. Each box stands for an iteration and contains a generation step and a validation step. Various colors are used for distinguishing the changed structures in each iteration (blue: dendrites; red: axons; cyan: increased structures from previous step; green: under-traced structures; black: over-traced structures).

a

Imaging data

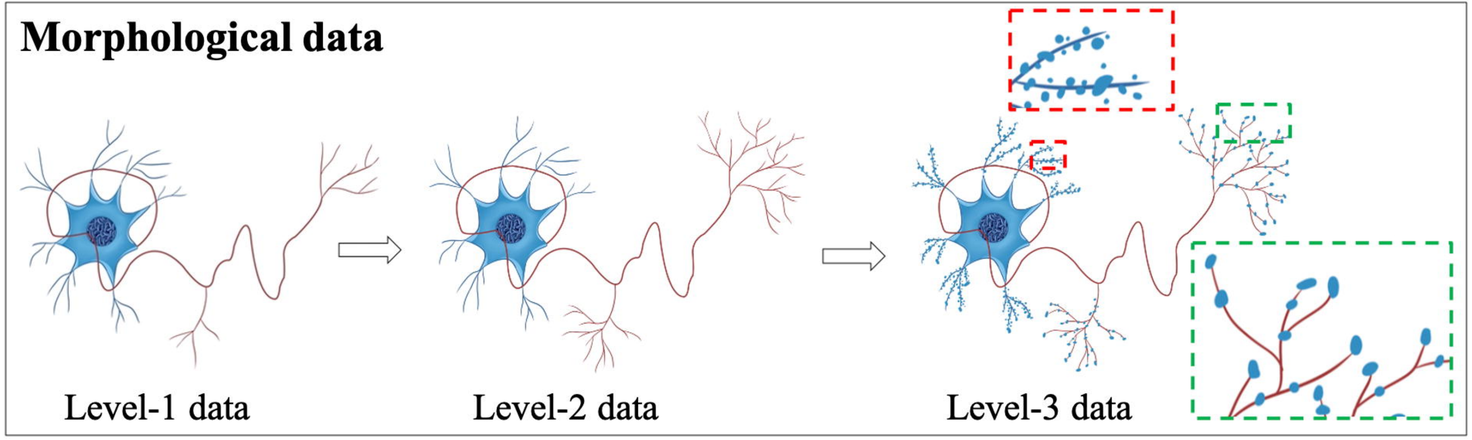


Whole-brain data



Level-0 data

Morphological data



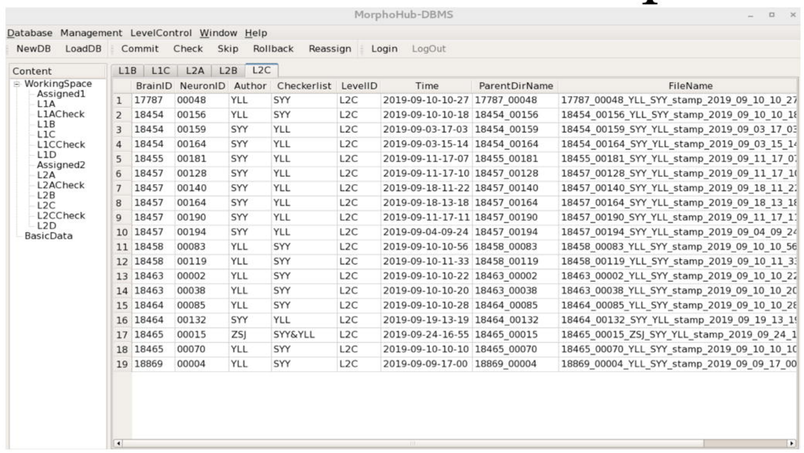
Level-1 data

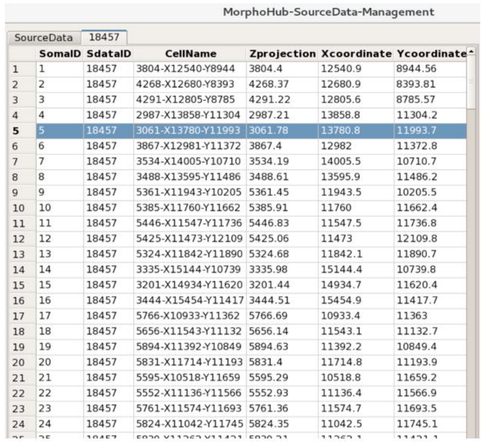
Level-2 data

Level-3 data

b

MorphoHub

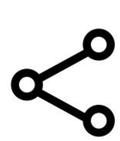




Multi-Morphometry generation

Data management

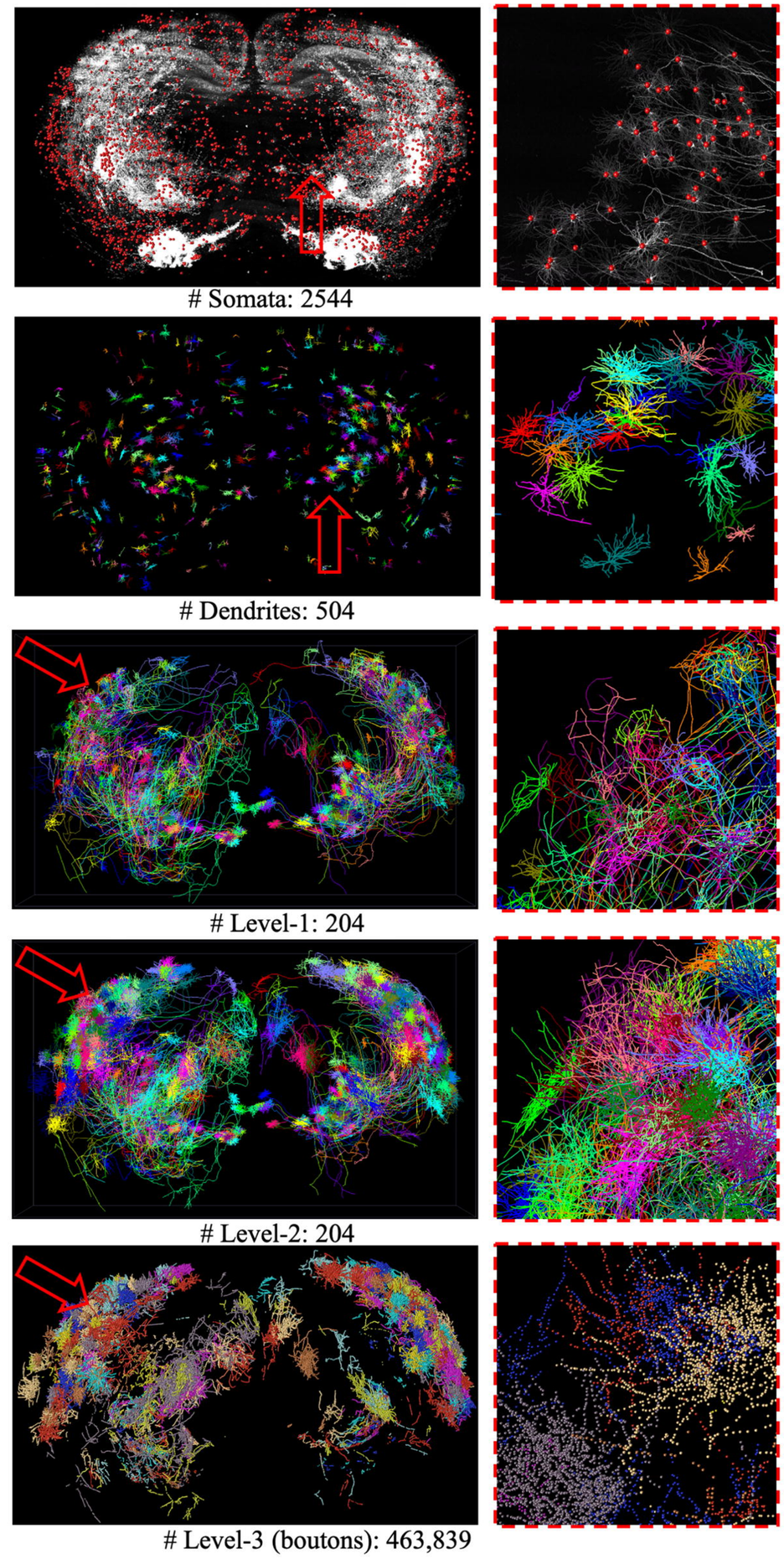
Collaboration

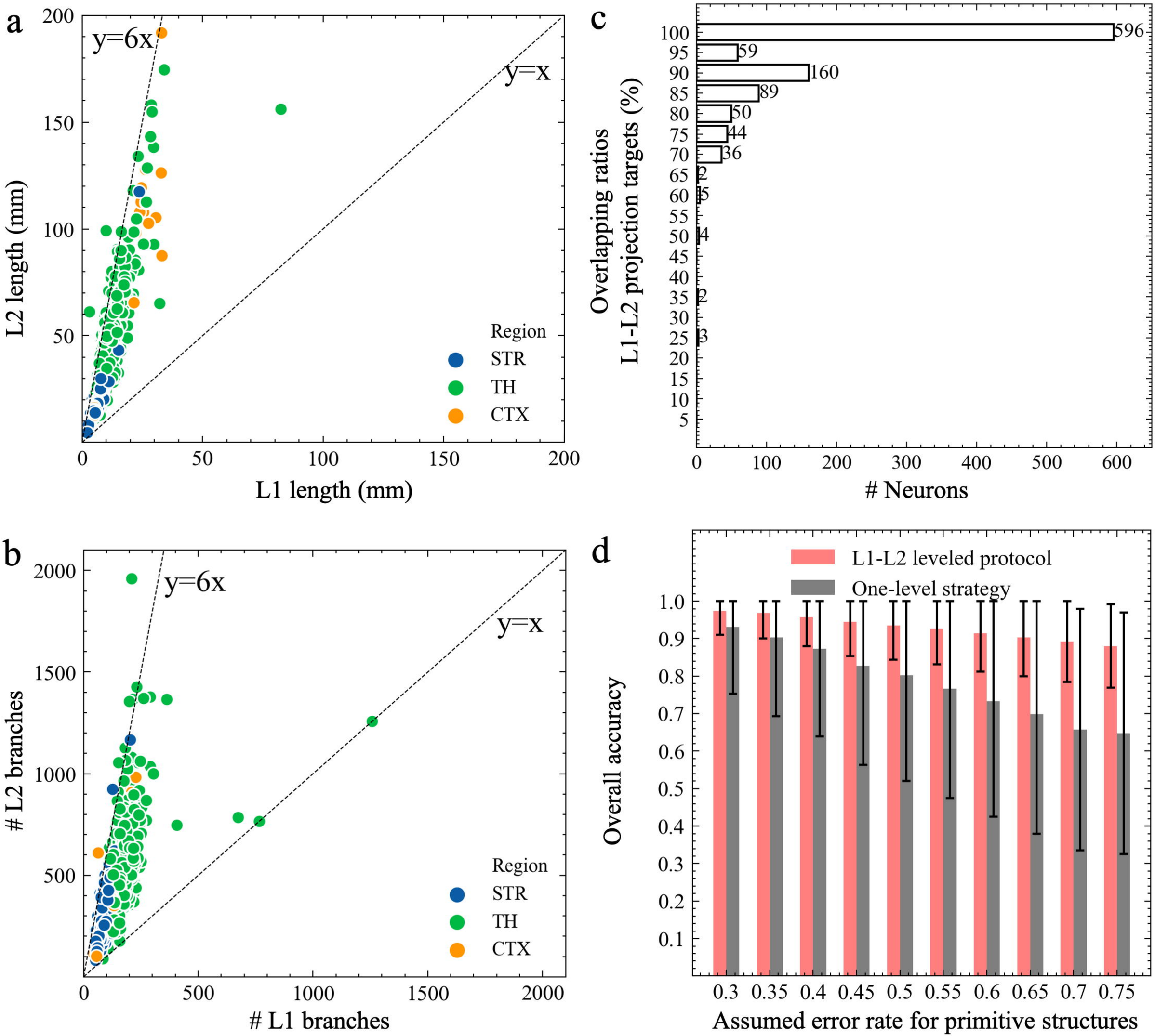


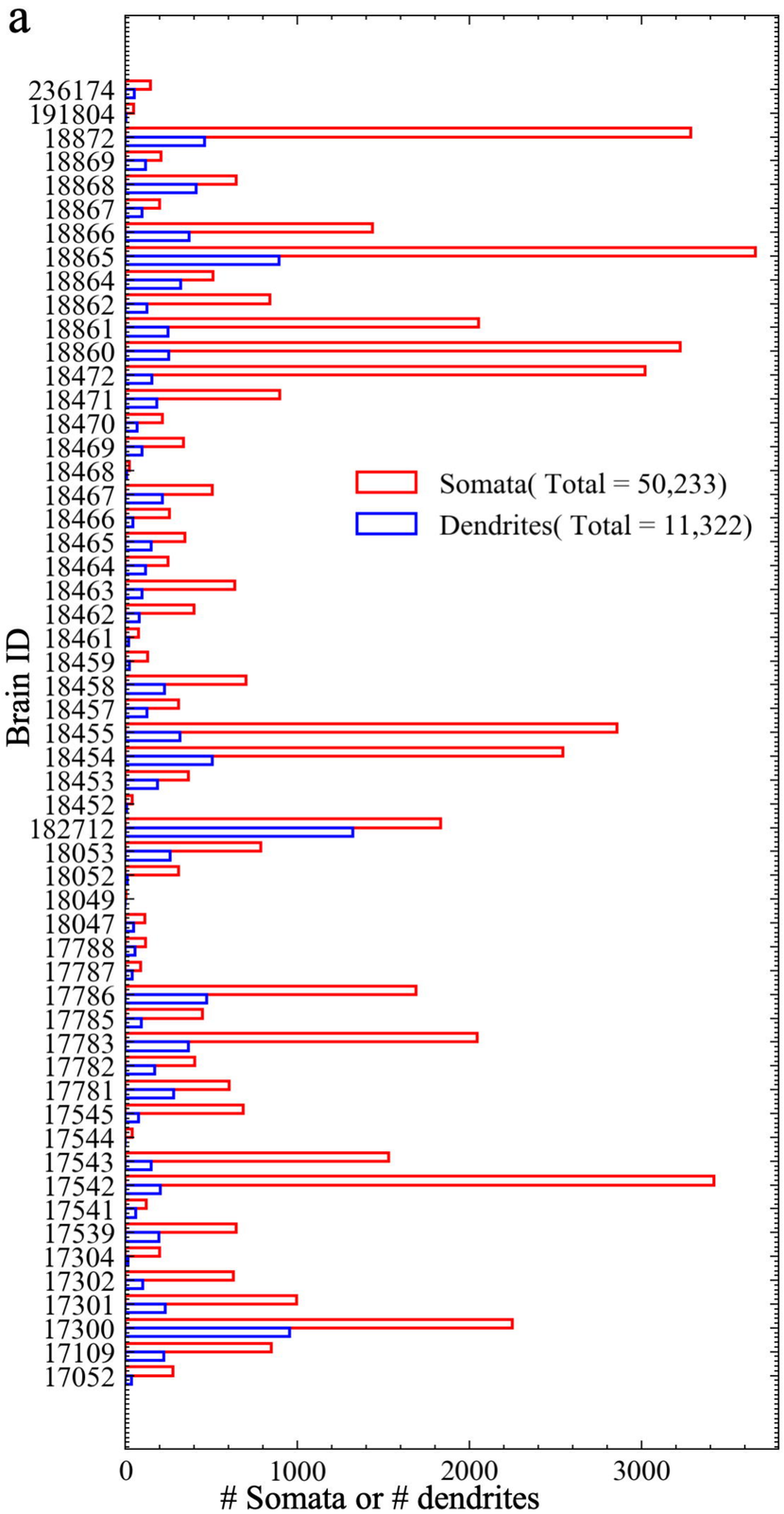
Data sharing

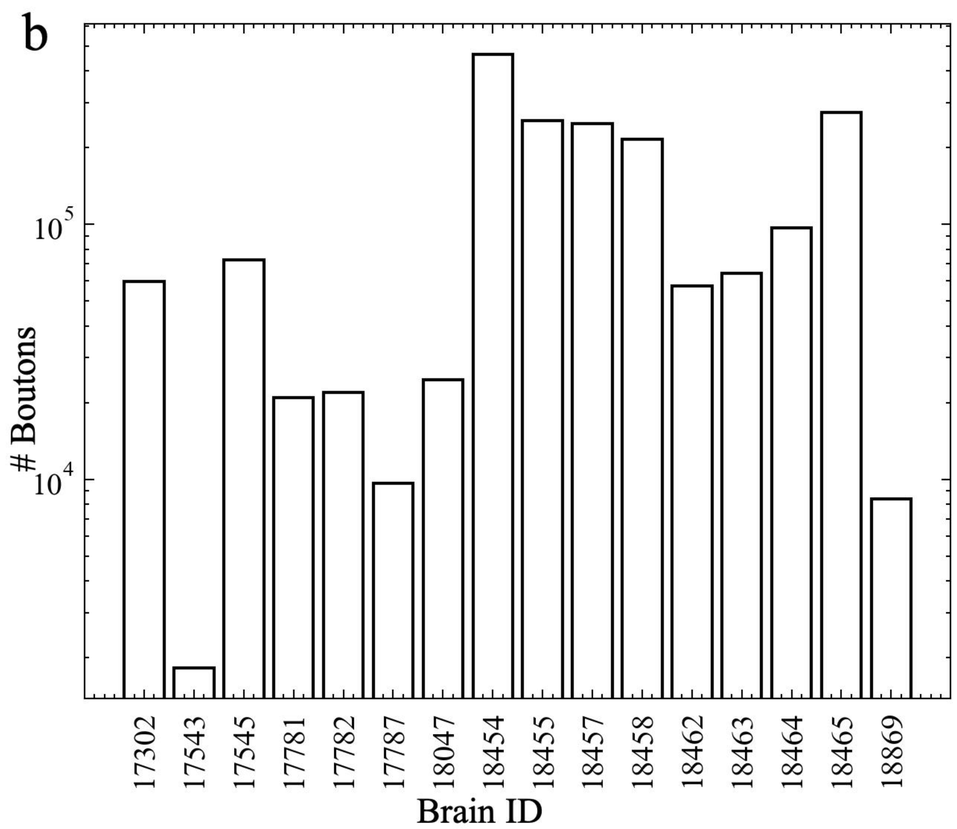
Universal interface

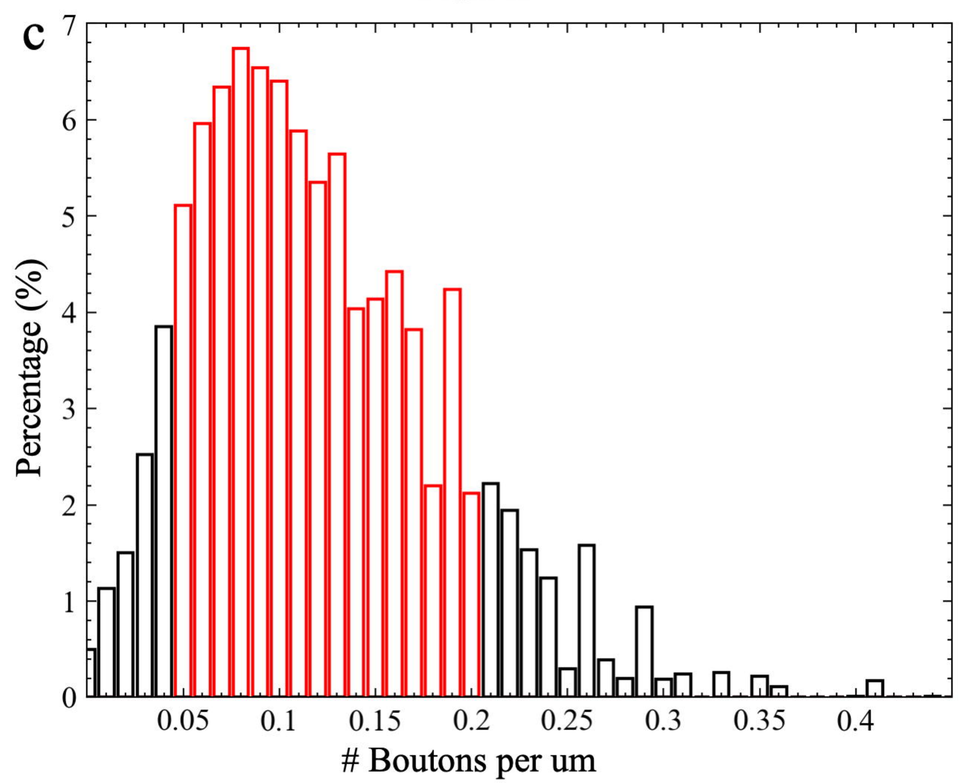
c

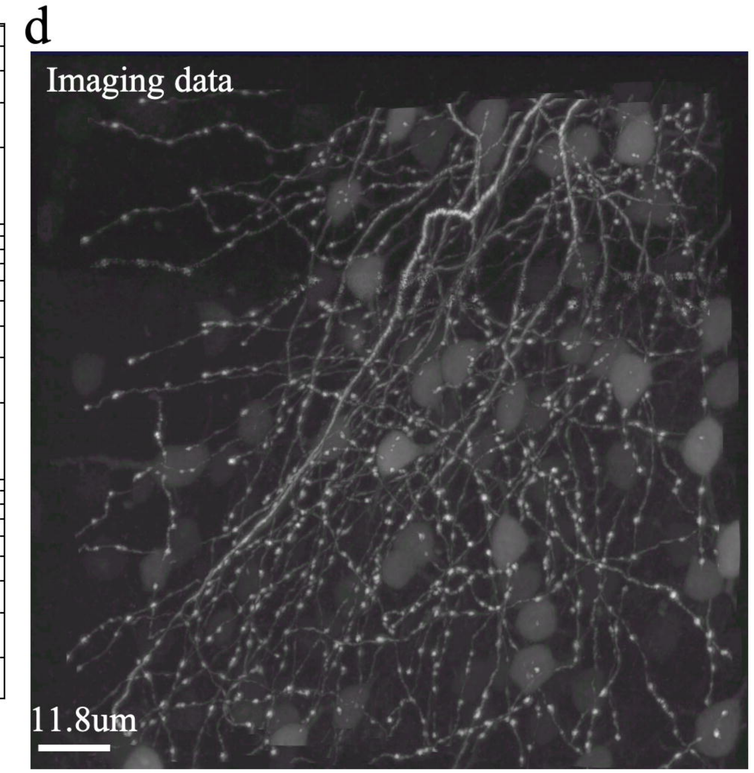


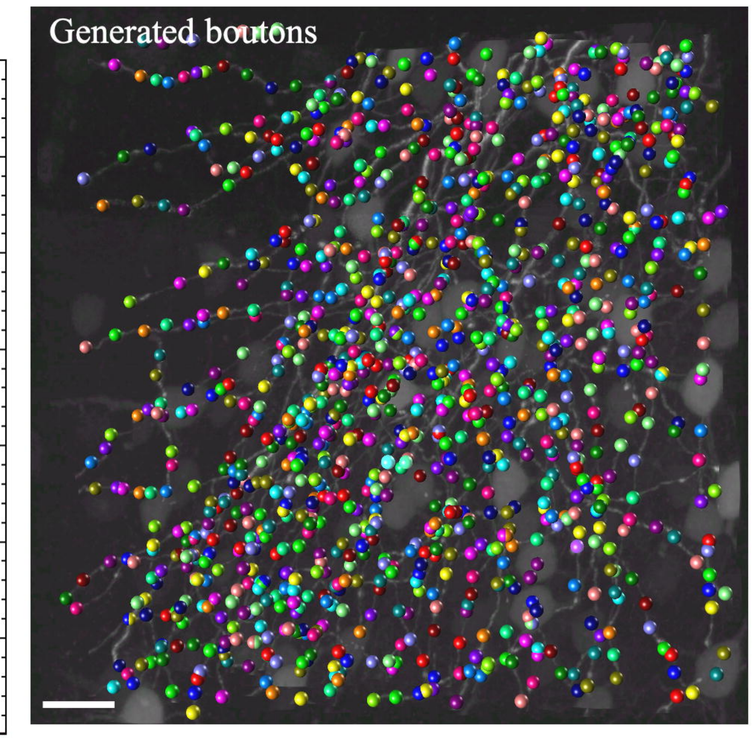

a
y=6x
y=x
L2 length (mm)
L1 length (mm)
Region
STR
TH
CTX
b
y=6x
y=x
# L2 branches
# L1 branches
Region
STR
TH
CTX
c
Overlapping ratios
L1-L2 projection targets (%)
Neurons
596
59
160
89
50
44
36
2
5
4
2
3
d
L1-L2 leveled protocol
One-level strategy
Overall accuracy
Assumed error rate for primitive structures

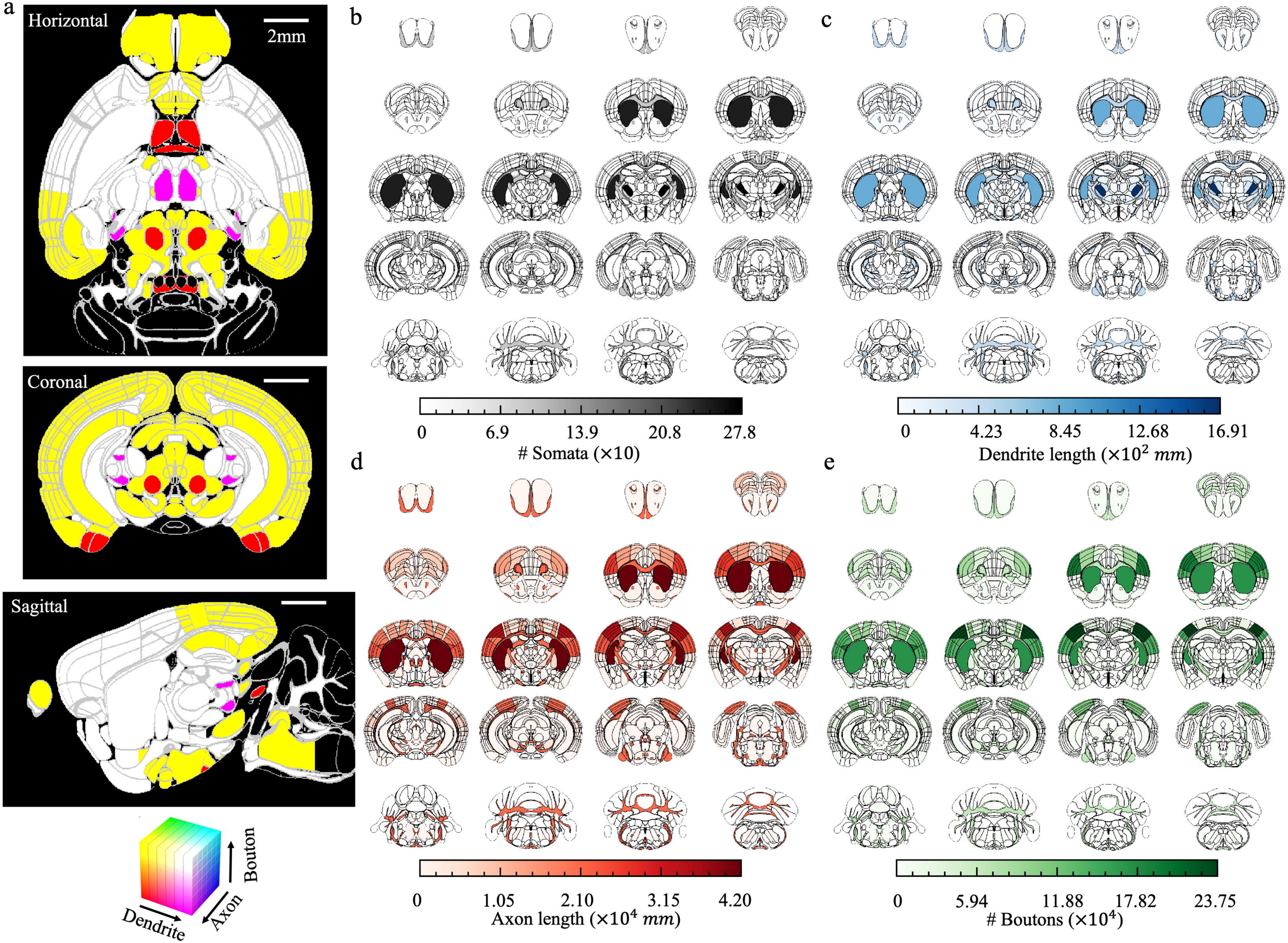

a
Brain ID
236174
191804
18872
18869
18868
18867
18866
18865
18864
18862
18861
18860
18472
18471
18470
18469
18468
18467
18466
18465
18464
18463
18462
18461
18459
18458
18457
18455
18454
18453
18452
182712
18053
18052
18049
18047
17788
17787
17786
17785
17783
17782
17781
17545
17544
17543
17542
17541
17539
17304
17302
17301
17300
17109
17052
Somata(Total = 50,233)
Dendrites(Total = 11,322)
0
1000
2000
3000
Somata or # dendrites

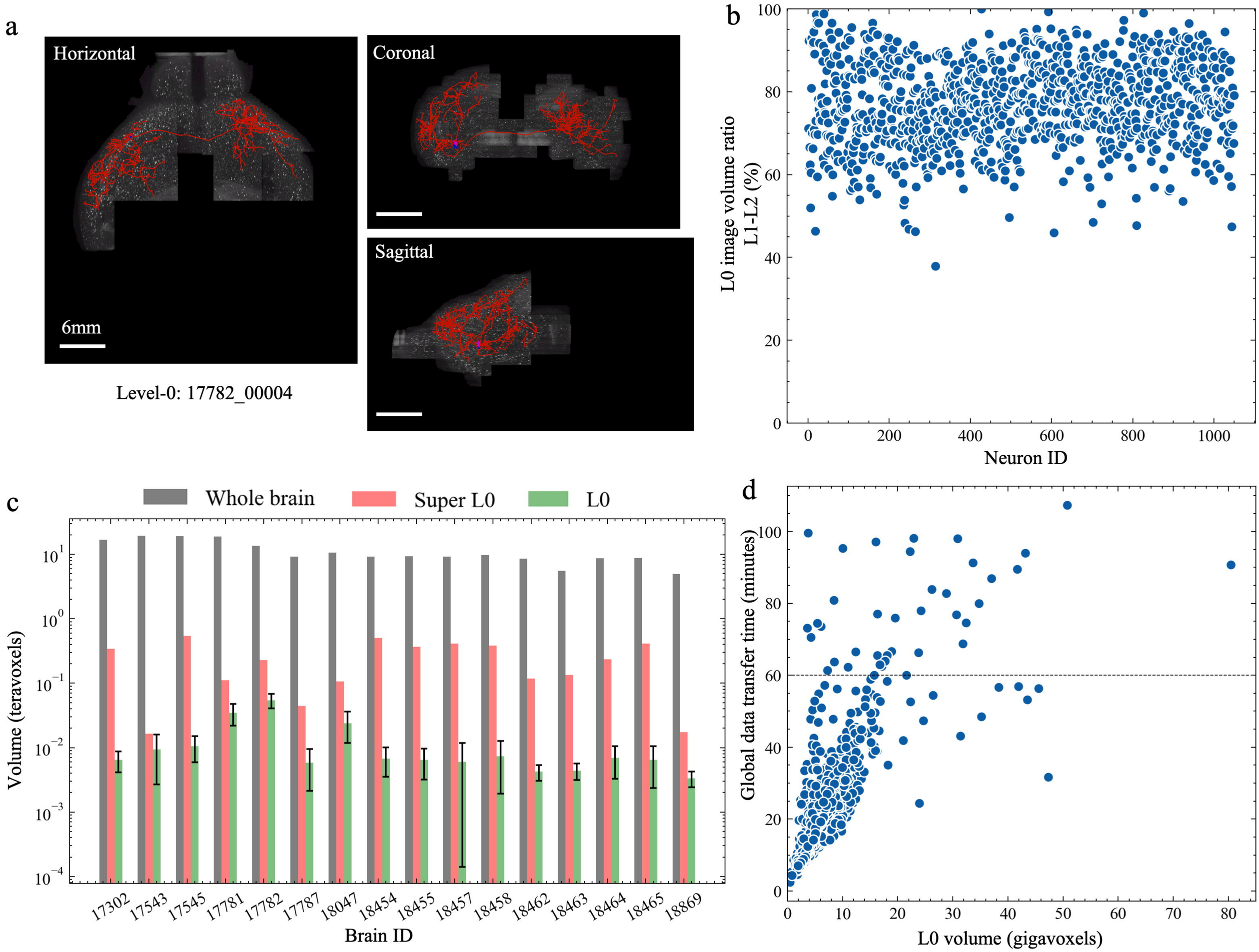

b
Boutons
10^5
10^4
17302
17543
17545
17781
17782
17787
18047
18454
18455
18457
18458
18462
18463
18464
18465
18869
Brain ID


c
Percentage (%)
0
1
2
3
4
5
6
7
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Boutons per um


d
Imaging data
11.8um


Generated boutons


a
Horizontal
2mm
Coronal
Sagittal
Bouton
Axon
Dendrite
b
0
6.9
13.9
20.8
27.8
Somata (×10)
c
0
4.23
8.45
12.68
16.91
Dendrite length (×10² mm)
d
0
1.05
2.10
3.15
4.20
Axon length (×10⁴ mm)
e
0
5.94
11.88
17.82
23.75
Boutons (×10⁴)


a
Horizontal
Coronal
Sagittal
6mm
Level-0: 17782_00004
b
L0 image volume ratio
L1-L2 (%)
Neuron ID
c
Whole brain
Super L0
L0
Volume (teravoxels)
Brain ID
17302
17543
17545
17781
17782
17787
18047
18454
18455
18457
18458
18462
18463
18464
18465
18869
d
Global data transfer time (minutes)
L0 volume (gigavoxels)