COMMENTARY

BrainLine: An Open Pipeline for Connectivity Analysis of Heterogeneous Whole-Brain Fluorescence Volumes

Thomas L. Athey[1,2] · Matthew A. Wright[3] · Marija Pavlovic[4] · Vikram Chandrashekhar[1] · Karl Deisseroth[4,5,6,7] · Michael I. Miller[1,2,8,9] · Joshua T. Vogelstein[1,2,8,9]

Main

Whole-brain image volumes at the micron scale are helping scientists characterize neuron-level morphology and connectivity, and discover new neuronal subtypes. These volumes require intense computational processing to uncover the rich neuronal information they contain. Currently, however, image acquisition is outstripping the availability and throughput of analysis pipelines. The steps in analyzing these images include registration, axon segmentation, soma detection, visualization and analysis of results. Several tools exist for these individual steps, but are rarely all part of an integrated pipeline and able to facilitate cloud-based collaboration (Tyson et al., 2022; Pisano et al., 2021). Further, many existing machine learning based tools are highly tuned to their training data and perform poorly when they encounter out-of-distribution artifacts or signal levels (Geisa et al., 2021).

To address these challenges, we present BrainLine, an open-source, fully-integrated pipeline that performs registration, axon segmentation, soma detection, visualization, and analysis on whole-brain fluorescence volumes (Figure 1a). BrainLine combines state-of-the-art, already available open-source tools such as CloudReg (Chandrashekhar et al., 2021) and ilastik (Berg et al., 2019) with brainlit, our Python package developed here. The BrainLine pipeline accommodates images that are hundreds of gigabytes in size and uses generalizable machine learning training schemes that adapt to out-of-distribution samples.

To share and interact with data across multiple institutions, BrainLine uses Amazon S3 to store data in precomputed format, so it can be viewed using Neuroglancer (n/a). Specifically, we use CloudReg (Chandrashekhar et al., 2021) for file conversion of the stitched image, and for image registration to the Allen atlas (Wang et al., 2020).

For axon segmentation and soma detection, we sought to leverage recent machine learning advances but experienced two major constraints. First, as generating ground truth image annotations is labor intensive, we wanted the approach to be effective on a small amount of training data. Second, images were provided to us in a sequential manner, and new samples would sometimes have unique artifacts or different levels of image quality (Figure 1b, c, e, f). We therefore sought a learning algorithm that could be quickly retrained on new data. Many learning algorithms assume that all training and testing data come from the same distribution and fail when this is not the case (Quinonero-Candela et al., 2008). However, using our closed-loop training paradigm with ilastik (Berg et al., 2019), we were able to use a single ilastik project for all samples, only occasionally adding training data when difficult samples arose.

We used an ilastik pixel classification workflow for both axon segmentation and soma detection, but in the latter case

✉ Thomas L. Athey
   tathey1@jhu.edu

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

2 Institute of Computational Medicine, Johns Hopkins University, Baltimore, MD, USA

3 Department of Psychiatry, Weill Cornell Medicine, New York, NY, USA

4 Department of Bioengineering, Stanford University, Stanford, CA, USA

5 CNC Program, Stanford University, Stanford, CA, USA

6 Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

7 Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA

8 Center for Imaging Science, Johns Hopkins University, Baltimore, MD, USA

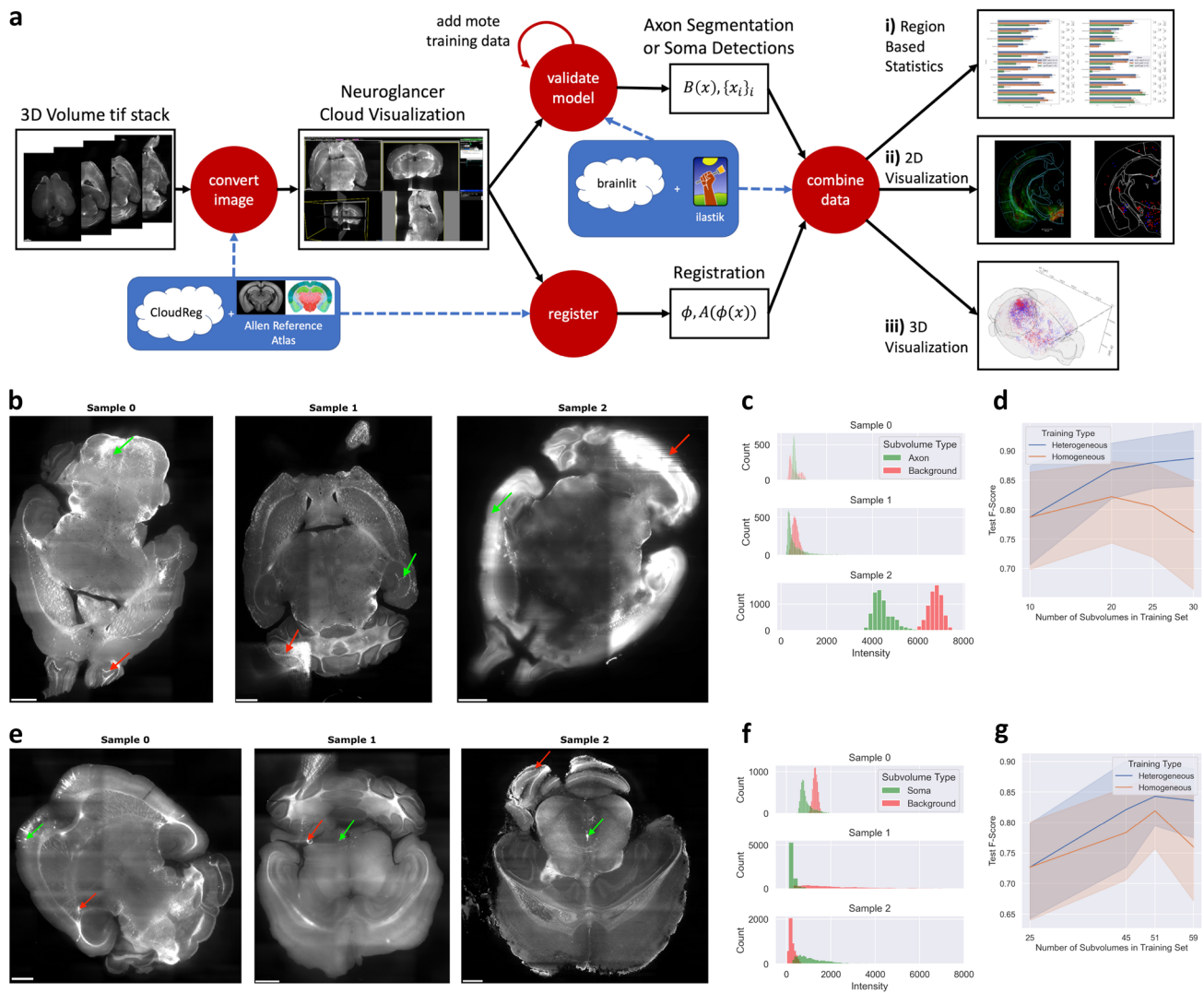9 Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD, USA

Fig. 1 BrainLine allows for efficient processing of heterogeneous whole brain fluorescence volumes. **a** BrainLine combines CloudReg (Chandrashekhar et al., 2021), ilastik (Berg et al., 2019) and our package, brainlit, to produce results in both quantitative (**a.i**) and visual (**a.ii-a.iii**) formats. **b** Example images with fluorescently labeled axon projections and arrows pointing to regions with (green) and without (red) labeled axons. **c** Intensity histograms of 20x20x20 voxel subvolumes located at the arrows in **b**. **d** Comparison between axon segmentation performance after training on subvolumes from different samples (heterogeneous) or the same sample (homogeneous). **e** Example images with fluorescently labeled cell bodies and arrows pointing to regions with (green) and without (red) labeled cell bodies. **f** Intensity histograms of 20x20x20 voxel subvolumes located at the arrows in **e**. **g** Comparison between soma detection performance after training on subvolumes from different brain samples (heterogeneous) or a single brain sample (homogeneous)

we applied a size threshold to the connected components following segmentation. In both cases, the training approach was the same. For each new whole-brain volume, we identified a set of subvolumes (99^3 voxels for axons, 49^3 for somas) across a variety of brain regions, and annotated only a few slices (three for axons, five for somas) in each subvolume for our validation set. This strategy is similar to that employed in Friedmann et al. (2020). If our model could not achieve a satisfactory f-score on this validation dataset, we would annotate more subvolumes from the sample and add them to the training set until satisfactory performance was achieved.

We observed that this heterogeneous training procedure (i.e. training on multiple brain samples) often improved performance on other samples as well. In an experiment where we controlled the number of subvolumes used for training, this approach was at least as good as a homogeneous approach, where all training subvolumes came from a single brain sample (Figure 1d, g).

The pipeline can display the axon segmentation and soma detection results in a variety of ways, including brain-region-based bar charts accompanied by statistical tests (Fig. 1a.i), 2D plots with the atlas borders (Fig. 1a.ii), and

3D visualizations using brainrender (Fig. 1a.iii) (Claudi et al., 2021). Since every experimental design is unique, we designed our pipeline in a modular way, so investigators can pick and choose which components they want to incorporate in their own analyses. We also leverage existing software and file formats to facilitate interoperability (Tyson et al., 2022).

BrainLine enables accelerated analysis of brain-wide connectivity through parallel programming, the use of cloud-compliant file formats, and a machine learning training scheme that generalizes across brain samples. As a result, BrainLine alleviates the need for investigators to build custom analysis pipelines from scratch, helping them characterize the morphology and connectivity profiles of neurons, and discover new neuronal subtypes. BrainLine is available as a set of thoroughly documented notebooks and scripts in our Python package brainlit: http://brainlit. neurodata.io/.

Declarations

Conflicts of Interest M.I.M. owns a significant share of Anatomy Works with the arrangement being managed by Johns Hopkins University in accordance with its conflict of interest policies. V.C. owns a significant share of Neurosimplicity, LLC, which is a medical device and technology company focusing on medical image processing. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Berg, S., Kutra, D., Kroeger, T. et al. (2019). Ilastik: interactive machine learning for (bio)image analysis. *Nature methods*, 16(12):1226–1232.

Chandrashekhar, V., Tward, D. J., Crowley, D. et al. (2021). Cloudreg: automatic terabyte-scale cross-modal brain volume registration. *Nature methods*, 18(8):845–846.

Claudi, F., Tyson, A. L., Petrucco, L., et al. (2021). Visualizing anatomically registered data with brainrender. *eLife*, 10:e65751.

Friedmann, D., Pun, A., Adams, E. L., et al. (2020). Mapping mesoscale axonal projections in the mouse brain using a 3d convolutional network. *Proceedings of the National Academy of Sciences*, 117(20):11068–11075.

Geisa, A., Mehta, R., Helm, H. S., et al. (2021). Towards a theory of out-of-distribution learning. arXiv preprint arXiv:2109.14501

Neuroglancer. (nd). https://github.com/google/neuroglancer

Pisano, T. J., Dhanerawala, Z. M., Kislin, M., et al. (2021). Homologous organization of cerebellar pathways to sensory, motor, and associative forebrain. *Cell reports*, 36(12):109721.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press.

Tyson, A. L., & Margrie, T. W. (2022). Mesoscale microscopy and image analysis tools for understanding the brain. *Progress in Biophysics and Molecular Biology*, 168:81–93.

Wang, Q., Ding, S., Li, Yang, R., et al. (2020). The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell, 181*(4), 936–953.

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