nature portfolio

Peer Review File

High-throughput analysis of dendritic and axonal arbors reveals transcriptomic correlates of neuroanatomy

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Review of Gliko et al for Nat comms titled: High-throughput analysis of dendritic and axonal arbors reveals transcriptomic correlates of neuroanatomy

The manuscript by Gliko et al presents a novel machine vision based algorithm for automatically tracing the axons and dendrites of neuronal morphologies collected using the Patch-seq method. The algorithm is based on the application of a number of machine learning approaches, including the application of a convolutional neural network to image stacks collected following biocytin labeling and imaging of neuronal morphologies. The considerable advance of the method is that it provides a 2-order of magnitude increase in time to generate a high-quality reconstruction relative to manual reconstructions. This enables the authors to automatically reconstruct additional sets of neurons from prior Patch-seq based datasets that the Allen Institute has collected previously.

The authors rigorously compare their automated to manually traced reconstructions and demonstrate that for many morphological features they have good accuracy from automated reconstructions relative to manual reconstructions. They mention that for some features, for example, axonal branch order, these are especially not well characterized from their automated compared to their manual reconstructions.

To demonstrate the utility of their automated morphological reconstruction approach in helping answer new scientific questions, they quantify metrics related 2D arbor density representations, and then relate these to gene expression from the same neurons as quantified via Patch-seq. They specifically try to understand how gene expression can explain differences in the amount of innervation of axons in layer 1 of the neocortex. They show that gene expression features can explain cell-to-cell differences within the Sst Calb2 Pdlim5 t-type based on this feature, and moreover, this isn't trivially due to differences in the soma position of these cells in the cortex. Specifically, when repeating this procedure for Lamp5 Lsp1 cells, they show that this prediction is entirely related to the depth of these cells in the cortex, and not L1 innervation per se. I find this demonstration for the scientific utility of automated reconstructions especially convincing and applaud the authors for adding this aspect to the paper.

In general, I find the paper very high quality and acceptable for publication in Nature Communications with some revisions (see below). In particular, the automated yet overall high accuracy approach towards morphological reconstruction presents a major, unprecedented advance.

Major comments and suggestions
1. There are many prior efforts for automating neuronal reconstructions based on microscopy. For example, the bigneuron consortium: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4725298/. How does the effort of Gliko et al fit into this larger framework? Have the authors considered using tools including reference datasets or benchmarking approaches developed via the bigneuron consortium (https://www.nature.com/articles/s41592-023-01848-5)? If not, why not?
2. How general is the pipeline and does it work well when applied to different datasets, for example, those with different imaging or tissue quality? For example, my expectation is that imaging datasets collected from human neocortical slices (e.g.,

https://www.science.org/doi/10.1126/science.adf6484) are likely to be of lower or varying quality relative to the mouse datasets analyzed in the current version of the manuscript. I would strongly encourage the authors trying this comparison (or a related cross-dataset comparison) and report the results, as this will do a lot to demonstrate the cross-lab usability and portability of their tool.
3. I appreciate the authors making their code and ML models accessible for reuse. My expectation is that others who might reuse this code and approach would appreciate a clearer and more sophisticated write up for how this code should be used that is better clearer and more thoroughly explained than what is provided in the shell script here: https://github.com/ogliko/patchseq-autorecon/blob/master/pipeline/example_pipeline.sh. Relatedly, do the authors think that the final trained ML models are likely to work on other lab's datasets without a lot of fine tuning? If not, it would be helpful to be more explicit about this in the discussion.

Minor comments
1. Please break up the results with major section headings
2. In the discussion, please expand on how the presented algorithm can further be improved. What is required for this? More manual reconstructions? How can these be scaled further up to enable this? If a human needs to remain in the loop, in which part is human intervention most needed?
3. Unclear why only 51 neurons used for training set, why not more? Like full 543 set?
4. Unclear based on legend what columns are in arbor density plot in Figure 2C
5. Unclear what data is exactly used in Figure 3. Is this the ADRs for axons? Dendrites? Both?
6. Why are features related to ADRs in Figure 3 are more predictive than features all?
7. Consider including a scatter plot of some automated vs manual data for Figure 2D to illustrate what the R2's are calculated from
8. Please provide more info than what's provided for new automated reconstructions here https://github.com/ogliko/patchseq-autorecon/tree/master/data

Reviewer #1 (Remarks on code availability):

I didn't try running the code but did quickly review major aspects of the code. I have provided a few brief comments related to the code and data as comments to the authors.

Reviewer #2 (Remarks to the Author):

Gliko et al. present an analysis pipeline for reconstructing neuronal morphologies from brightfield image stacks. Overall, this work represents an important step towards solving a long-standing challenge in the field, and I agree with the authors that morphologic reconstruction is the major bottleneck in analyzing multimodal Patch-seq datasets. The data are compelling and clearly presented.

While this certainly sounds like a useful tool for the Allen Institute, my major concern is that the authors have not gone far enough to ensure that it will be useful to the neuroscience

community more broadly. Specifically, the authors perform all training and testing on data generated by their own group using highly standardized protocols, and all from the adult mouse visual cortex. Can this analysis pipeline handle data collected using different imaging platforms, using slight alterations in staining protocols, or from different species and brain regions? Will every lab need to re-train the algorithm using their own data? If so, what sort of computational resources would be required and how long will that take (estimates are given for analysis after training but not for the training itself)? The Github repositories are a nice start but seem more focused on documenting what the authors have done than on helping other labs implement a similar approach for their own data (e.g. re-training the algorithms). If the 100 cells/day estimate is only a two-fold increase over the standard approach with a single anatomist, I worry that implementation of this pipeline won't be feasible or valuable for labs without trained machine learning experts to easily implement and train the pipeline.

Additional minor concerns:
1. Line 66: "Martinotti cells (e.g., 24)." Is this a typo? Seems like something besides a reference should come after e.g.
2. Figure 2a – is it possible to reconstruct the axon for excitatory neurons? Is the axon reconstruction pipeline only built using inhibitory neurons? Are the structural characteristics of the inhibitory neuron axon similar enough to the excitatory neuron axon, so you can apply one pipeline to the other?
3. Line 126: any possible reason for such mistakes (i.e. maximum branch order)?

Reviewer #2 (Remarks on code availability):

There seems to be enough information to reproduce what the authors have done (I did not try it myself), but definitely not enough information for others to replicate this type of analysis on their own data (i.e. re-training the model).

Reviewer #3 (Remarks to the Author):

In their paper, Gliko et al. introduce a U-Net architecture to automatically segment brightfield microscopy images from neurons stained during patch-seq. They train their model on manually traced neurons and produce reconstructions of many more, enabling them to further analyze the relationship between morphology/projection patterns and gene expression. A segementation procedure for brightfield microscopy has been long overdue.

Major comments:
- Choice of training set: Why is the training of the algorithm only based on 51 manually traces neurons? The Allen Institute surely has many more examples, maybe without transcriptomic data but still sufficient for training. Would this improve the segmentation quality?
- Speed comparison: This comparison in lines 87ff is slightly odd and should be rewritten. Why compare a single anatomist to 16 GPU cards? Maybe turn around and ask: If an anatomist can reliably trace a single cell a day, how long does a computer on a single GPU take? With multiple GPUs there is obvious parallelism, and you also don't consider 100 anatomists.
- Reconstruction accuracy: It would be interesting to report the reconstruction accuracy separately for axons and dendrites and pyramidal and interneurons. Also, could the authors

provide an estimate of the consistency of two anatomists reconstructing identical cells and how close the algorithm gets?
- Feature accuracy: The analysis presented in Fig. 2d is surprisingly worrisome, given the nice reconstructions shown before. This seems to suggest that a large number of features do not correlate at all between manual and automated reconstructions, and for only 50%, the correlation is above 0.5 (it is a bit hard to judge the colorscale). Some of the ones at the bottom of the figure like dendrite_num_branches I find very surprising. Could the authors double check and maybe add scatter plots across cells to the supplement?
- A radical idea: The authors argue in the paper (lines 169ff) that for much of the analysis related to cell typing can just as well only rely on ADRs (density maps) rather than detailed reconstructions. Could we then not just sidestep the difficult reconstruction problem and straightaway train a unet to obtain ADRs from the raw microscopic images? Would this work?
- The conclusion stated in line 188 in passing that the gene prediction analysis suggests a discrete code is a little too strong for my taste as only selected representatives of the different subclasses are organized. The conclusion e.g. by Scala et al. that cell types form a continuum was made based on the very fine t-types within a cell class and showed that pairs of t-types with more similar transcriptome were also correspondingly more similar in physiology/anatomy, in a way that suggests continuous variation. This is entirely consistent with the finding by the authors, that selected types especially in different subclasses are quite different. The authors could move this to the Discussion with a more cautious tone.
- For the gene selection procedure using lasso regression, was cross validation with inner and outer loop used like in Cadwell et al. 2015? Otherwise, it is possible to obtain quite some overfitting by selecting the hyperparameter alpha and the coefficients w in one go, especially in the n<<d scenario studied here.

Minor comments:
- The sentence starting “In these repetitive…” in line 32 is very hard to parse. Consider rephrasing. Also, the “therefore” at the beginning of the next sentence is a non sequitur for me.
- Line 103: To ensure _the_ integrity
- Line 105: If _the_ manual trace…
- In fact, I have a hard time following the information presented in lines 100-108. Please rewrite.
- The authors may want to point out earlier that the most informative features of neural morphologies are typically anyway the overall shape of the dendritic and axonal projections, not the fine details, so it is not crucial to get every last bit right if the task is cell typing.
- Line 132: develop -> implemented
- Fig. 4: Is Axon mass really in a.u. or rather between 0 and 1? Wouldn’t it make sense to report the fraction of axon mass in L1?
- Fig. 2d: RMS distance is hard to intuitively interpret – could the authors also compute correlation?
- Line 152: “Here an ultimate comparison…” -> consider rephrasing, style
- Line 152: “Boschloo’s test…” -> rephrase to match the style of the rest of the paper, where you describe results instead of tests
- Line 162: “Even if the assignment is incorrect, the dominance…” This statement is wrong, and compares entities on different levels. The first part of the sentence refers to a single datum (and in this case, if the classification is incorrect, it is incorrect), the second half of the sentence to the overall reasonable performance. Rephrase.
- Sparse regression models and correction for some depth: Why is some depth not included

as a feature into the regression but removed separately?

Philipp Berens

Reviewer #3 (Remarks on code availability):

I look at the online repository and superficially browsed the code. It is available, seems well commented and straightforward to parse.

Response to referees' comments

We thank the referees for their insightful and constructive reviews. We are happy to receive their broadly positive and encouraging comments. They have also raised multiple points, some of which are shared across the reviews. We have now edited the manuscript to address them in full and we believe the manuscript has improved significantly as a result. Below, we first respond to those common concerns. Then, we provide a point-by-point response. Throughout, bold font indicates original referee comments.

Common points:

1. Testing/demonstrating automated segmentation on other datasets (R1, R2)
   The referees inquired about the generality of the pipeline, beyond the dataset studied in the manuscript. To address this concern, we ran our pipeline without any further training or fine-tuning on four different example image stacks for which we have access to corresponding manual segmentations:

   - image of a human cortical neuron (tissue preparation was performed at Gabor Tamas's lab, imaging was done at the Allen Institute)
   - image of a macaque subcortical neuron obtained by the Allen Institute's Patch-seq setup
   - image of a mouse subcortical neuron obtained by the Allen Institute's Patch-seq setup
   - image of the part of a local axon of a cat cortical excitatory cell (This image is part of the publicly available DIADEM dataset. It was originally obtained $\sim$19 years ago by Judith Hirsch's lab.)

   Across all four examples, the automated pipeline produced compelling traces of the underlying neurons, demonstrating the robustness of the method, and suggesting that it can be useful to scientists by providing automated initial traces out of the box. This can dramatically shorten the time needed for a final manual tracing/inspection stage. Naturally, in the presence of training examples on the anatomy of interest, transfer learning (e.g., fine-tuning) should further boost the performance.

   We have now generated a new supplementary figure (Supp. Fig. 13) displaying maximum intensity projections of these four example images, and their manual and automated traces. We have also edited the main text to refer to this new figure in the context of generalizability and transfer learning (lines 90-93).

2. Choice of training set (R1, R3)

   The reviewers asked questions on both the size of the training set and the methodology used in choosing those training examples.

   We had three main objectives in choosing the training set:

   - The training set should ideally be representative of the underlying image space. This argues for a large and balanced set of training images.
   - As few of the images in the dataset as possible should be used in the training set to facilitate the downstream cross-validation studies (e.g., quantification of cell type prediction accuracy based on automatically generated traces) via the availability of cells that are not used during training. This argues for a minimal set of training images.

- As the training set grows, storage and efficient retrieval of the raw and label image patches become practical engineering concerns. While solutions exist, such deployment is beyond the scope of this paper and our computational resources. We note that training was performed by a single desktop computer with one GPU. (Inference was performed with 16 GPUs.)

We balanced these concerns by choosing a set of 51 images that we considered as representative both in terms of image quality (e.g., including images that are noisy or otherwise considered hard to segment) and neuronal identity (i.e., including images of neurons with a variety of t-type assignments).

Once the merits of the approach are established on test images not used during training (e.g., via this manuscript), a training routine that uses more of the available manual segmentations is likely to further improve performance. However, using all the images in a straightforward way may still be suboptimal because some cell types and/or image types may be over-represented, potentially resulting in poorer generalization.

To address the reviewers' question, we have now edited the main text to explain our methodology and objectives in choosing the training images (lines 76-80).

3. Use of code base by other scientists (R1, R2)

To improve the usability of the code base by other scientists (potentially with minimal computational experience), we have taken two steps:

- We have provided more guidance on running the different aspects of our github code base by (i) adding new sample notebooks on various analysis and postprocessing steps, (ii) adding new README files under individual folders, and (iii) added more comments to better explain the different modules.
- We have set up an interactive repository on CodeOcean (facilitated by the journal) to demonstrate some of the main aspects of the code base. We believe this repository is shared with the reviewers by the journal. This interactive repository enables users to test certain aspects of the pipeline without downloading or installing anything, and directly from their web browsers. Not all aspects are available under CodeOcean due to space and compute restrictions. We will add the link to this CodeOcean repository under the Code Availability section once it is made available to us.

Point-by-point response:

Reviewer #1 (Remarks to the Author):

Review of Gliko et al for Nat comms titled: High-throughput analysis of dendritic and axonal arbors reveals transcriptomic correlates of neuroanatomy

The manuscript by Gliko et al presents a novel machine vision based algorithm for automatically tracing the axons and dendrites of neuronal morphologies collected using the Patch-seq method. The algorithm is based on the application of a number of machine learning approaches, including the application of a convolutional neural network to image stacks collected following biocytin labeling and imaging of neuronal morphologies. The considerable advance of the method is that it provides a 2-order of magnitude increase in time to generate a high-quality reconstruction relative to manual reconstructions. This enables the authors to automatically reconstruct additional sets of neurons from prior Patch-seq based datasets that the Allen Institute has collected previously.

The authors rigorously compare their automated to manually traced reconstructions and demonstrate that for many morphological features they have good accuracy from automated reconstructions relative to manual reconstructions. They mention that for some features, for example, axonal branch

order, these are especially not well characterized from their automated compared to their manual reconstructions.

To demonstrate the utility of their automated morphological reconstruction approach in helping answer new scientific questions, they quantify metrics related 2D arbor density representations, and then relate these to gene expression from the same neurons as quantified via Patch-seq. They specifically try to understand how gene expression can explain differences in the amount of innervation of axons in layer 1 of the neocortex. They show that gene expression features can explain cell-to-cell differences within the Sst Calb2 Pdlim5 t-type based on this feature, and moreover, this isn't trivially due to differences in the soma position of these cells in the cortex. Specifically, when repeating this procedure for Lamp5 Lsp1 cells, they show that this prediction is entirely related to the depth of these cells in the cortex, and not L1 innervation per se. I find this demonstration for the scientific utility of automated reconstructions especially convincing and applaud the authors for adding this aspect to the paper.

In general, I find the paper very high quality and acceptable for publication in Nature Communications with some revisions (see below). In particular, the automated yet overall high accuracy approach towards morphological reconstruction presents a major, unprecedented advance.

We thank the reviewer for their encouraging comments. We think the above summary is both informative and accurate.

Major comments and suggestions

1. There are many prior efforts for automating neuronal reconstructions based on microscopy. For example, the bigneuron consortium: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4725298/`**. How does the effort of Gliko et al fit into this larger framework? Have the authors considered using tools including reference datasets or benchmarking approaches developed via the bigneuron consortium (**`https://www.nature.com/articles/s41592-023-01848-5`**)? If not, why not?**

The reviewer makes an interesting observation. Indeed, the shortcomings of these efforts on brightfield images (because their focus is on fluorescent, optical-sectioning microscopy) have been a starting point for us. We have now tested the performance of the APP2 automated segmentation plug-in of the Vaa3D software, which is the main software tool associated with the BigNeuron consortium, on four randomly chosen example images in our dataset.

To address the reviewer's questions, we generated a new supplementary figure (Supp. Fig. 12) displaying maximum intensity projections of those four example images, the corresponding manual segmentations, the segmentations generated automatically by Vaa3D, and the segmentations generated by our automating tracing pipeline. Broadly, we observe that Vaa3D/APP2 struggles to segment these images, despite our efforts to tune APP2's parameters to maximize its performance. As mentioned above, this may be due to these tools' focus on fluorescent, optical-sectioning microscopy. We have also edited the main text to refer to this new supplementary figure (lines 88-90).

2. How general is the pipeline and does it work well when applied to different datasets, for example, those with different imaging or tissue quality? For example, my expectation is that imaging datasets collected from human neocortical slices (e.g., `https://www.science.org/doi/10.1126/science.adf6484`**) are likely to be of lower or varying quality relative to the mouse datasets analyzed in the current version of the manuscript. I would strongly encourage the authors trying this comparison (or a related cross-dataset comparison) and report the results, as this will do a lot to demonstrate the cross-lab usability and portability of their tool.**

We thank the reviewer for this suggestion. We have taken their advice and tested our pipeline on images obtained from different species, brain regions, and imaging setups. Please refer to Common points/1 above.

Our experience is that thin axons represent the most challenging part of the neuronal anatomy in terms of

reconstruction, independent of the brain region and species. In all cases in the new Supp. Fig. S13, the automated traces seem to provide good starting points to delineate the arbor and manually reconstruct its fine morphology, without any fine-tuning/training.

Finally, the model that is used in the Sorensen, Gouwens, Wang et al preprint, for which some example traces are shown in Figure 2 (and new Supplementary Figure S14), was initialized with the model presented here. This is now clarified in the Neural network architecture and training subsection of the Methods section.

3. I appreciate the authors making their code and ML models accessible for reuse. My expectation is that others who might reuse this code and approach would appreciate a clearer and more sophisticated write up for how this code should be used that is better clearer and more thoroughly explained than what is provided in the shell script here: `https://github.com/ogliko/patchseq-autorecon/blob/master/pipeline/example_pipeline.sh`**. Relatedly, do the authors think that the final trained ML models are likely to work on other lab's datasets without a lot of fine tuning? If not, it would be helpful to be more explicit about this in the discussion.**

Please refer to Common points/3 above for the steps we have taken to address the referee's concern on the usability of the code base. Please refer to our response to Common points/1 above for the steps we have taken to address the referee's concern on the applicability of our method to other labs' datasets without much fine-tuning. The experimental steps other than imaging for one of the images in Supp. Fig. 13 were performed by a lab in Hungary. For another brightfield image in that figure, the whole experiment was performed by a different lab more than 15 years ago, with a different experimental setup.

To better address the referee's concern, we have now emphasized that while our method can be useful without fine-tuning on setups and neurons that differ in significant ways, optimal performance will be achieved by transfer learning on a (presumably smaller) training set for the study of interest (lines 90-93).

Minor comments

1. Please break up the results with major section headings

Done as suggested.

2. In the discussion, please expand on how the presented algorithm can further be improved. What is required for this? More manual reconstructions? How can these be scaled further up to enable this? If a human needs to remain in the loop, in which part is human intervention most needed?

We thank the reviewer for this suggestion. We have now expanded the relevant part of the Discussion to address these questions (lines 304-314).

3. Unclear why only 51 neurons used for training set, why not more? Like full 543 set?

Please refer to Common points/2 for an expanded response. Briefly, one key reason is to exclude as many cells as possible from training to be able to report trustworthy test results for downstream tasks (e.g., classification).

4. Unclear based on legend what columns are in arbor density plot in Figure 2C

We have now added axis labels to Figure 2c. We thank the reviewer for catching this.

5. Unclear what data is exactly used in Figure 3. Is this the ADRs for axons? Dendrites? Both?

Arbor density representations of both the axon and dendrites are used for classification. We have now clarified this in the figure caption. We thank the reviewer for catching this.

6. Why are features related to ADRs in Figure 3 are more predictive than features all?

We believe there was a confusion due to our shorthand headers in the figure. We have now clarified these to replace "features all" with "features, all cells". We also edited the headers of all the other panels in a similar way.

7. Consider including a scatter plot of some automated vs manual data for Figure 2D to illustrate what the R2's are calculated from

Thank you for this suggestion. We have now added a new figure (Figure S16) displaying the scatter plots for each morphometric feature. We have also added another new figure (Figure S17) displaying the scatter plots for precision and recall values for the node matching-based comparison. Lastly, we have added another new figure (Figure S18) displaying scatter plots for root-mean-square error and correlation values based on the manual vs automatically generated ADRs.

We now report Pearson's correlation instead of R^2 in Fig. 2d, following R3's suggestion. We also fixed two mistakes: (i) the values on the scale bar were larger than they should be, making the rms error look larger, (ii) six training cells were included in Figs. 2b,d. Both mistakes are corrected in the revised manuscript.

8. Please provide more info than what's provided for new automated reconstructions here `https://github.com/ogliko/patchseq-autorecon/tree/master/data`

Done as suggested. Please refer to Common points/3 for an expanded response.

Reviewer #1 (Remarks on code availability):

I didn't try running the code but did quickly review major aspects of the code. I have provided a few brief comments related to the code and data as comments to the authors.

We thank the referee for those suggestions. We believe the usability of our code base has improved. Please refer to Common points/3 to review the steps we have taken.

Reviewer #2 (Remarks to the Author):

Gliko et al. present an analysis pipeline for reconstructing neuronal morphologies from brightfield image stacks. Overall, this work represents an important step towards solving a long-standing challenge in the field, and I agree with the authors that morphologic reconstruction is the major bottleneck in analyzing multimodal Patch-seq datasets. The data are compelling and clearly presented.

We thank the reviewer for their encouraging comments.

While this certainly sounds like a useful tool for the Allen Institute, my major concern is that the authors have not gone far enough to ensure that it will be useful to the neuroscience community more broadly. Specifically, the authors perform all training and testing on data generated by their own group using highly standardized protocols, and all from the adult mouse visual cortex. Can this analysis pipeline handle data collected using different imaging platforms, using slight alterations in staining protocols, or from different species and brain regions?

We thank the referee for this comment, which was useful in improving the manuscript. In addition to our response here, please also refer to Common points/1 above for the steps we have taken to address the referee's concerns.

We have now added a new figure (Figure S13) demonstrating the degree of robustness of the method against changes to the underlying brain region, species, and imaging setup. Overall, this figure demonstrates that the performance of the pre-trained network, *out-of-the-box*, deteriorates relatively mildly in the face of major changes to the experimental setup and tissue. We think this pre-trained network will be useful in a few ways:

- If only a delineation of the morphology is needed, the pre-trained network captures it in most cases without further training, potentially with an increased rate of axon vs. dendrite confusions.
- In case a precise depiction of the morphology, hence manual tracing, is needed, the tracing output of this pre-trained model serves as a good starting point, decreasing the time needed for manual tracing.

- When possible, some fine-tuning by the user should optimize the model's performance on the dataset of interest. However, this fine-tuning does not have to be a full-fledged training session with a large training set. The pre-trained network can provide a warm start in this scenario, decreasing the number of training epochs and the number of cells traced to train this fine-tuning stage. This is sometimes referred to as transfer learning in the machine learning literature.

To address the reviewer's concern, we have now reflected a summary of the above discussion to the manuscript (lines 90-93), in addition to the above-mentioned new figure to support these statements.

Will every lab need to re-train the algorithm using their own data? If so, what sort of computational resources would be required and how long will that take (estimates are given for analysis after training but not for the training itself)?

As mentioned above, we expect the need to re-train to depend on the particular scientific application. In almost all cases, we believe training from scratch with a large training set will not be necessary. It took us $\sim$3 weeks to train the model using a single GPU. We expect fine-tuning to require a fraction of this time. The training time, along with other details on training, is provided under the Neural network architecture and training subsection of the Methods section.

The Github repositories are a nice start but seem more focused on documenting what the authors have done than on helping other labs implement a similar approach for their own data (e.g. re-training the algorithms).

Please refer to Common points/3 above for the steps we have taken to improve the usability of our code base by other labs.

If the 100 cells/day estimate is only a two-fold increase over the standard approach with a single anatomist, I worry that implementation of this pipeline won't be feasible or valuable for labs without trained machine learning experts to easily implement and train the pipeline.

We would like to address a potential confusion: our estimate is that $\sim$100 cells/day corresponds to two orders-of-magnitude improvement in speed, not two-fold. That is, we estimate that, on average, it takes an anatomist $\sim$1 day per cell. So, we estimate that the improvement is closer to 100-fold. Please also refer to our response above on the scenarios that, we think, would either not require re-training or require minimal resource use for fine-tuning.

Additional minor concerns:

1. Line 66: "Martinotti cells (e.g., 24)." Is this a typo? Seems like something besides a reference should come after e.g.

We have now edited inside the parentheses - thanks for catching it.

2. Figure 2a – is it possible to reconstruct the axon for excitatory neurons? Is the axon reconstruction pipeline only built using inhibitory neurons? Are the structural characteristics of the inhibitory neuron axon similar enough to the excitatory neuron axon, so you can apply one pipeline to the other?

We have now added a new figure to the manuscript (Figure S14) demonstrating that the model can reconstruct the local axon for excitatory neurons when it is captured in the image, and referred to this ability and the figure in the main text (lines 100-102). This figure was generated using our excitatory model, showcased in Fig. 2a (right) for apical and basal dendrites, which is also available in the github repository.

We did not attempt to run the inhibitory model on excitatory neurons to study axon detection. We observe that the axon initial segments of excitatory neurons look different from peripheral axonal branches.

3. Line 126: any possible reason for such mistakes (i.e. maximum branch order)?

The model was trained with a voxel-based loss function for computational efficiency. Therefore, flipping the value of a voxel where it creates a spurious connection or splits a branch is penalized equally with flipping a

voxel where it just changes the thickness of a branch. Moreover, in these filamentous structures, the topology is very fragile because a mistake in a single voxel is likely to change the connectivity. This is why the arbor density representation is more robust although it doesn't preserve fine details of morphology. To address the reviewer's concern, we have now added this perspective to the Discussion (lines 309-314). We thank the referee for this comment.

Reviewer #2 (Remarks on code availability):

There seems to be enough information to reproduce what the authors have done (I did not try it myself), but definitely not enough information for others to replicate this type of analysis on their own data (i.e. re-training the model).

Thank you for this comment. Please refer to Common points/3 above for the steps we have taken to improve the usability of our code base by other labs.

Reviewer #3 (Remarks to the Author):

In their paper, Gliko et al. introduce a U-Net architecture to automatically segment brightfield microscopy images from neurons stained during patch-seq. They train their model on manually traced neurons and produce reconstructions of many more, enabling them to further analyze the relationship between morphology/projection patterns and gene expression. A segementation procedure for brightfield microscopy has been long overdue.

We thank the reviewer for pointing out this gap in the literature that the present manuscript aims to address.

Major comments:

- Choice of training set: Why is the training of the algorithm only based on 51 manually traces neurons? The Allen Institute surely has many more examples, maybe without transcriptomic data but still sufficient for training. Would this improve the segmentation quality?

Please refer to Common points/2 above for an explanation of our strategy in curating the training set and the steps we have taken. Briefly, for cells with transcriptomic characterization, while a larger training set may improve performance, it would also leave too few samples for downstream cross-validation studies, considering that some of the t-types have only a few samples that are manually traced.

The referee makes an interesting suggestion regarding cells with manual traces but without transcriptomic characterization. Indeed, 2 out of 51 cells in the training set do not have transcriptomic characterization, as the supplemental gallery indicates. Here, we focused on cells from the Patch-seq pipeline with transcriptomic characterization, to be able to place the training set within the context of t-types. (The supplemental reconstruction gallery is organized by transcriptomic types and indicates the cells belonging to the training set.)

We would like to mention two further considerations: (i) The manual work to admit a traced neuron to the training set is not insignificant: choosing the correct brightness parameter for topology-preserving volumetric inflation and ensuring that every branch in view is traced (e.g., tracers may ignore the axon initial segment of excitatory cells, they typically only trace the neuron of interest even if branches from other neurons are in view.) requires on the order of a few hours per cell. (ii) As mentioned in Common points/2 above, storage and retrieval of 3d raw and label images in an efficient way so as to facilitate neural network training became practical concerns for our training setup, which comprises a single desktop computer and a GPU card.

We agree that the referee's idea will be useful for a future model (that also trains on the majority of the cells with transcriptomic profiles, having already performed cross-validation studies here). We now mention this idea

in the Discussion section (lines 304-309). Finally, we consider the existing repository as live and we aim to add more capable models to it if/when we manage to train them. We thank the referee for this idea.

- Speed comparison: This comparison in lines 87ff is slightly odd and should be rewritten. Why compare a single anatomist to 16 GPU cards? Maybe turn around and ask: If an anatomist can reliably trace a single cell a day, how long does a computer on a single GPU take? With multiple GPUs there is obvious parallelism, and you also don't consider 100 anatomists.

While that comparison describes our setup and provides all relevant numbers in two consecutive sentences (so that we think the chance of confusion is very slim), we decided to rewrite those sentences to address the referee's concern (lines 93-97):

```
The overall pipeline produces neuron reconstructions in the commonly-used swc
format [32] from raw image stacks at a rate of ~6 cells/day with a single
GPU card (Methods). Our setup uses 16 cards to achieve two orders of magnitude
improvement in speed over semi-manual segmentation [5] with one anatomist.
```

We would also like to mention that the obvious parallelism and automation is part of the point: (i) while the computer can work 24/7, the anatomist cannot, (ii) while the anatomist may find the repetitive nature of tracing boring, the computer will not, (iii) while the anatomist may make mistakes due to decreased attention (e.g., due to the repetitive nature of tracing), the computer will not, (iii) while it is obvious and trivial to scale up to a large number of GPUs, it is much harder to scale up to a large number of anatomists (it may not even be an option).

- Reconstruction accuracy: It would be interesting to report the reconstruction accuracy separately for axons and dendrites and pyramidal and interneurons. Also, could the authors provide an estimate of the consistency of two anatomists reconstructing identical cells and how close the algorithm gets?

We thank the reviewer for this suggestion. To address the reviewer's concern, we have added a new figure (Figure S18) reporting the arbor density prediction error in Fig. 2d separately for axons and dendrites, at the single cell level, via a scatter plot. We have realized that the original scale bar associated with Fig. 2d-right was incorrect, making the RMS error appear much larger than it actually is. We have now corrected this mistake.

Fig. 2b and Table S1 show the reconstruction accuracy separately for axons and dendrites based on the node correspondence study. Fig. 2d calculates the feature accuracies separately for axons and dendrites. The newly added Table S2 (next paragraph) now reports the accuracy separately for axons and dendrites. We have not reported results separately for pyramidal vs interneurons because we have only shown qualitative results from a few pyramidal neurons throughout the manuscript.

We have added another new figure (Figure S15) and a new table (Table S2) that compares inter-human tracing accuracy with that of the automated approach, based on one cell for which we had three different manually generated traces from three anatomists (lines 136-137).

- Feature accuracy: The analysis presented in Fig. 2d is surprisingly worrisome, given the nice reconstructions shown before. This seems to suggest that a large number of features do not correlate at all between manual and automated reconstructions, and for only 50%, the correlation is above 0.5 (it is a bit hard to judge the colorscale). Some of the ones at the bottom of the figure like dendrite_num_branches I find very surprising. Could the authors double check and maybe add scatter plots across cells to the supplement?

We would like to remind that the original Fig. 2d shows R^2 values, not correlations. To avoid confusions, we have now regenerated Fig. 2d-left, to report the correlation values.

We repeat that, thanks to the referee's inquiry, we spotted a mistake in the scale bar of Fig. 2d-right, and corrected it.

As the referee suggested, we have now added scatter plots for all features (new Figure S16), which demon-

strates strong correlations for many features.

We have also added two more new figures: (i) new Figure S17 shows scatter plots of precision and recall values for the node correspondence study, (ii) new Figure S18a shows scatter plots of axon vs dendrite RMS error of automatically generated ADRs, (iii) new Figure S18b shows scatter plots of axon vs dendrite Pearson correlation values of automatically generated ADRs and ADRs generated from manual traces.

- A radical idea: The authors argue in the paper (lines 169ff) that for much of the analysis related to cell typing can just as well only rely on ADRs (density maps) rather than detailed reconstructions. Could we then not just sidestep the difficult reconstruction problem and straightaway train a unet to obtain ADRs from the raw microscopic images? Would this work?

This idea sounds interesting and reasonable. In an older paper (Ref. [38]), we explored a related idea in a sparse imaging setting in the retina where the arbor density was generated from the volumetric reconstruction (e.g., "denoised" image) instead of the (fragile) skeleton. Within the confines of that study, we observed that this idea worked well with slight degradations.

From a practical perspective, the right amount of low-pass filtering along the different axes to obtain ADRs will depend on the region, species, and the number of cell types in the experimental setup. Therefore, reconstructing the neuron, if doable, would allow for quick, post-hoc adjustments to ADR generation rather than training from scratch. Naturally, other studies utilizing the arbor morphology beyond cell typing would also benefit from detailed reconstructions.

To address the reviewer's suggestion, we now mention this perspective in the Discussion (lines 275-276). We thank the referee for their suggestion.

- The conclusion stated in line 188 in passing that the gene prediction analysis suggests a discrete code is a little too strong for my taste as only selected representatives of the different subclasses are organized. The conclusion e.g. by Scala et al. that cell types form a continuum was made based on the very fine t-types within a cell class and showed that pairs of t-types with more similar transcriptome were also correspondingly more similar in physiology/anatomy, in a way that suggests continuous variation. This is entirely consistent with the finding by the authors, that selected types especially in different subclasses are quite different. The authors could move this to the Discussion with a more cautious tone.

We thank the referee for this suggestion. We have edited that statement and added a cautionary remark against conclusions on the discreteness of the cell type landscape (lines 206-209).

- For the gene selection procedure using lasso regression, was cross validation with inner and outer loop used like in Cadwell et al. 2015? Otherwise, it is possible to obtain quite some overfitting by selecting the hyperparameter alpha and the coefficients w in one go, especially in the n«d scenario studied here.

Yes, as described under the Sparse feature selection analysis subsection, the LassoCV command in the scikit-learn library was used. For every cross-validation fold, first the alpha parameter is inferred, then the model is fit again to infer the sparse regression coefficients. Finally, 10 genes with the absolute largest coefficients are chosen, and test R^2 values are calculated based on the held-out test set.

Minor comments:

- The sentence starting "In these repetitive…" in line 32 is very hard to parse. Consider rephrasing. Also, the "therefore" at the beginning of the next sentence is a non sequitur for me.

Done as suggested.

- Line 103: To ensure _the_ integrity

Done as suggested.

- Line 105: If _the_ manual trace...

Done as suggested.

- In fact, I have a hard time following the information presented in lines 100-108. Please rewrite.

Done as suggested.

- The authors may want to point out earlier that the most informative features of neural morphologies are typically anyway the overall shape of the dendritic and axonal projections, not the fine details, so it is not crucial to get every last bit right if the task is cell typing.

While we agree with this perspective, we decided that it would require multiple additional sentences to make this point properly in light of recent studies showing that fine details can *also* carry information [Elabbady et al., bioRxiv, 2024, `https://www.biorxiv.org/content/10.1101/2022.07.20.499976v2`], potentially distracting the reader.

- Line 132: develop -> implemented

While the 2D *radial* ADR is not a main claim of novelty for the manuscript, it is also different from existing representations. Therefore, we thought 'develop' would be more suitable. Throughout the main text, we use simple present tense.

- Fig. 4: Is Axon mass really in a.u. or rather between 0 and 1? Wouldn't it make sense to report the fraction of axon mass in L1?

Done as suggested.

- Fig. 2d: RMS distance is hard to intuitively interpret – could the authors also compute correlation?

Done as suggested. We have also added a new figure (Figure S18) reporting the associated correlation values as a scatter plot. Moreover, as mentioned above, the referee's inquiry regarding Fig. 2d has led us to catch a mistake in the scale bar of this figure. We thank the referee for their careful examination - we have now corrected that mistake.

We also corrected one more mistake: In Figs. 2b,d, six training cells were included by mistake. The revised manuscript shows the corrected figures.

- Line 152: "Here an ultimate comparison..." -> consider rephrasing, style

Done as suggested.

- Line 152: "Boschloo's test..." -> rephrase to match the style of the rest of the paper, where you describe results instead of tests

Done as suggested.

- Line 162: "Even if the assignment is incorrect, the dominance..." This statement is wrong, and compares entities on different levels. The first part of the sentence refers to a single datum (and in this case, if the classification is incorrect, it is incorrect), the second half of the sentence to the overall reasonable performance. Rephrase.

We have rephrased this statement to the best of our understanding of the reviewer's concern (lines 177-180).

Please note the original text is "Even when...", not "Even if".

- Sparse regression models and correction for some depth: Why is some depth not included as a feature into the regression but removed separately?

(i) The main character of our study is to predict morphological features from transcriptomic features. Including soma depth as an independent variable would have gone against this. (ii) Including soma depth as an independent variable would have further increased the correlations between features (i.e., soma depth and soma depth-predicting genes).

Reviewer #3 (Remarks on code availability):

I look at the online repository and superficially browsed the code. It is available, seems well commented and straightforward to parse.

Thank you. We have further improved the presentation and usability of the code base.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I applaud the authors for sufficiently responding and revising their manuscript in light of mine and the other reviewer's comments. I have no further comments and feel the manuscript is acceptable in published form for Nature Communications.

I am signing my review for the purpose of transparency: Shreejoy Tripathy

Reviewer #1 (Remarks on code availability):

I've provided a very brief glance at the code and between the updated code and my prior comments on the last version, feel that the authors have sufficiently addressed our prior comments on code availability and documentation.

Reviewer #2 (Remarks to the Author):

The authors have adequately addressed all of my concerns. I think the manuscript is a good fit for the audience of Nature Communications.

Reviewer #3 (Remarks to the Author):

I thank the authors for their responses to my comments and the revised manuscript. Unfortunately, I am not quite satisfied with their answers.

For example, I still don't understand why the authors didn't use a larger training set. For this, you don't need Patch-Seq'ed neurons, you could just use any brightfield imaged neuron, and I am sure the Allen Institute has thousands. The stated reasons - representative for the underlying space, remaining neurons needed for downstream analysis, computational power - seem like made up excuses (e.g. I am sure the Allen Institute has access or could buy access to more than one (!) desktop GPU.

The authors still refer to their 51 neurons as a "large training set", something it is clearly not. It remains unclear how the performance of the model would scale. If the authors are concerned of systematic bias by overrepresenting certain types, if anything a larger training set should help, and it would be for the authors to show these kind of effects.

The reply to the question for demonstrating generalization abilities by R1 and R2 is also insufficient, as the authors only show four examples, instead of a quanitative evaluation on a different dataset. Similarly, in response to my question on consistency, the authors produced one example, where they simply plot the automated trace and the traces by the three tracers. A quantitative analysis would be needed in both cases. I realize tracing neurons is hard work, but it should be possible to obtain traces for 4-5 neurons from 2-3 anatomists.

Response to referees' comments

Reviewer #1 (Remarks to the Author):

I applaud the authors for sufficiently responding and revising their manuscript in light of mine and the other reviewer's comments. I have no further comments and feel the manuscript is acceptable in published form for Nature Communications.

I am signing my review for the purpose of transparency: Shreejoy Tripathy

We are happy that this reviewer finds our manuscript suitable for publication. We thank the reviewer for their constructive and insightful comments throughout, which have helped to improve our manuscript significantly.

Reviewer #1 (Remarks on code availability):

I've provided a very brief glance at the code and between the updated code and my prior comments on the last version, feel that the authors have sufficiently addressed our prior comments on code availability and documentation.

We are happy that this reviewer finds our publicly available code base and its documentation adequate.

Reviewer #2 (Remarks to the Author):

The authors have adequately addressed all of my concerns. I think the manuscript is a good fit for the audience of Nature Communications.

We are happy that this reviewer finds our manuscript suitable for publication. We thank the reviewer for their constructive and insightful comments throughout, which have helped to improve our manuscript significantly.

Reviewer #3 (Remarks to the Author):

I thank the authors for their responses to my comments and the revised manuscript. Unfortunately, I am not quite satisfied with their answers.

We thank the reviewer for their constructive and insightful comments throughout, which have helped to improve our manuscript significantly.

For example, I still don't understand why the authors didn't use a larger training set. For this, you don't need Patch-Seq'ed neurons, you could just use any brightfield imaged neuron, and I am sure the Allen Institute has thousands. The stated reasons - representative for the underlying space, remaining neurons needed for downstream analysis, computational power - seem like made up excuses (e.g. I am sure the Allen Institute has access or could buy access to more than one (!) desktop GPU.

The authors still refer to their 51 neurons as a "large training set", something it is clearly not. It remains unclear how the performance of the model would scale. If the authors are concerned of systematic bias by overrepresenting certain types, if anything a larger training set should help, and it would be for the authors to show these kind of effects.

These 3-d training sets include $\sim$1 Teravoxel of data and labels, including both foreground and background voxels ($\sim$0.5 Teravoxels of raw data for the inhibitory model and $\sim$0.8 Teravoxels for the excitatory model).

We also emphasize that the manual work to admit one already traced neuron to the training set takes a few hours per cell. This includes choosing the correct brightness parameter for topology-preserving volumetric inflation and ensuring that every branch in view is traced. (e.g., tracers may have ignored the axon initial segment of excitatory cells. Also, they typically only trace the neuron of interest even if branches from other neurons are in view.)

To address the reviewer's concern, we edited the main text to discuss these points more explicitly (lines 305–314).

The reply to the question for demonstrating generalization abilities by R1 and R2 is also insufficient, as the authors only show four examples, instead of a quanitative evaluation on a different dataset. Similarly, in response to my question on consistency, the authors produced one example, where they simply plot the automated trace and the traces by the three tracers. A quantitative analysis would be needed in both cases. I realize tracing neurons is hard work, but it should be possible to obtain traces for 4-5 neurons from 2-3 anatomists.

We address a potential omission by the reviewer: in response to their question on consistency, we also included a new table (Table S2) quantifying the reconstruction accuracies of multiple traces of one test neuron. We do not disagree that more manual tracing would not be futile. We remind that manual tracing of one neuron takes a few days. We edited the corresponding text to clarify this point (lines 136–138).

The newly added examples from different datasets in response to R1 and R2's comments are meant to provide anecdotal evidence that the method does not simply collapse under significantly different settings. We do not have quantitative claims on datasets from different species, etc when the trained model is applied as is. We edited the relevant text to avoid any confusion (lines 94–96).