Large-scale unsupervised discovery of excitatory morphological cell types in mouse visual cortex

Marissa A. Weis[1,2], Stelios Papadopoulos[4,5], Laura Hansel[1], Timo Lüddecke[1], Brendan Celii[4,5,10], Paul G. Fahey[4,5], J. Alexander Bae[6,8], Agnes L. Bodor[9], Derrick Brittain[9], JoAnn Buchanan[9], Daniel J. Bumbarger[9], Manuel A. Castro[6], Forrest Collman[9], Nuno Maçarico da Costa[9], Sven Dorkenwald[6,7], Leila Elabbady[9], Akhilesh Halageri[6], Zhen Jia[6,7], Chris Jordan[6], Dan Kapner[9], Nico Kemnitz[6], Sam Kinn[9], Kisuk Lee[6,11], Kai Li[6,7], Ran Lu[6], Thomas Macrina[6,7], Gayathri Mahalingam[9], Eric Mitchell[6], Shanka Subhra Mondal[6,7], Shang Mu[6], Barak Nehoran[6,7], Sergiy Popovych[6,7], R. Clay Reid[9], Casey M. Schneider-Mizell[9], H. Sebastian Seung[6,7], William Silversmith[6], Marc Takeno[9], Russel Torres[9], Nicholas L. Turner[6,7], William Wong[6], Jingpeng Wu[6], Wenjing Yin[9], Szi-chieh Yu[6], Jacob Reimer[4,5], Andreas S. Tolias[4,5], and Alexander S. Ecker[1,3,*]

[1]Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Germany
[2]Institute for Theoretical Physics, University of Tübingen, Germany
[3]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany
[4]Department of Neuroscience, Baylor College of Medicine, Houston, USA
[5]Center for Neuroscience and AI, Baylor College of Medicine, Houston, USA
[6]Princeton Neuroscience Institute, Princeton University, USA
[7]Department of Computer Science, Princeton University, USA
[8]Department of Electrical Engineering, Princeton University, USA
[9]Allen Institute for Brain Science, Seattle, WA, USA
[10]Rice University, Houston, TX, USA
[11]Massachusetts Institute of Technology, Cambridge, MA, USA
[*]*Correspondence: ecker@cs.uni-goettingen.de*

Abstract

Neurons in the neocortex exhibit astonishing morphological diversity which is critical for properly wiring neural circuits and giving neurons their functional properties. The extent to which the morphological diversity of excitatory neurons forms a continuum or is built from distinct clusters of cell types remains an open question. Here we took a data-driven approach using graph-based machine learning methods to obtain a low-dimensional morphological "bar code" describing more than 30,000 excitatory neurons in mouse visual areas V1, AL and RL that were reconstructed from a millimeter scale serial-section electron microscopy volume. We found a set of principles that captured the morphological diversity of the dendrites of excitatory neurons. First, their morphologies varied with respect to three major axes: soma depth, total apical and basal skeletal length. Second, neurons in layer 2/3 showed a strong trend of a decreasing width of their dendritic arbor and a smaller tuft with increasing cortical depth. Third, in layer 4, atufted neurons were primarily located in the primary visual cortex, while tufted neurons were more abundant in higher visual areas. Fourth, we discovered layer 4 neurons in V1 on the border to layer 5 which showed a tendency towards avoiding deeper layers with their dendrites. In summary, excitatory neurons exhibited a substantial degree of dendritic morphological variation, both within and across cortical layers, but this variation mostly formed a continuum, with only a few notable exceptions in deeper layers.

1 Introduction

Neurons have incredibly complex and diverse shapes. Ever since Ramón y Cajal, neuroanatomists have studied their morphology [19] and have classified them into different types. From a computational point of view, their dendritic morphology constrains which inputs a neuron receives, how these inputs are integrated and, thus, which computations the neuron and the circuit it is part of can learn to perform.

Less than 15% of neocortical neurons are inhibitory, yet they are morphologically the most diverse and can be classified reliably into well-defined subtypes [3, 7]. The vast majority of cortical neurons are excitatory. Excitatory cells can be divided into spiny stellate and pyramidal cells [16]. Although pyramidal cells have a very stereotypical dendritic morphology, they do exhibit a large degree of morphological diversity. Recent studies subdivide them into 10–20 cell types using manual classification (e.g. [14]) or clustering algorithms applied to dendritic morphological features [6, 10, 15].

Existing studies of excitatory morphologies have revealed a number of consistent patterns, such as the well-known thick-tufted pyramidal cells of layer 5 [8, 6, 10, 14, 15]. However, a commonly agreed-upon morphological taxonomy of excitatory neuron types is yet to be established. For instance, Markram et al. [14] describe two types of thick-tufted pyramidal cells based on the location of the bifurcation point of the apical dendrite (early vs. late). Later studies suggest that these form two ends of a continuous spectrum [10, 6]. Other authors even observe that morphological features overall do not form isolated clusters and suggest an organization into families with more continuous variation within families [21]. There are two main limitations of previous morphological characterizations: First, many rely on relatively small numbers of reconstructed neurons used to asses the morphological landscape. Second, they represent the dendritic morphology using summary statistics such as point counts, segment lengths, volumes, density profiles (so-called morphometrics; [15, 23, 13]) or graph-based topological measures [9]. These features were handcrafted by humans and may not capture all crucial axes of variation.

We here take a data-driven approach using a recently developed unsupervised representation learning approach [26] to extract a morphological feature representation directly from the dendritic skeleton. We apply this approach to a large-scale anatomical dataset [2] to obtain a low-dimensional vector embedding ("bar code") of more than 30,000 neurons in mouse visual areas V1, AL and RL. Our analysis suggests that excitatory neurons' morphologies form a continuum, with notable exceptions such as layer 5 thick-tufted cells, and vary with respect to three major axes: soma depth, total apical and total basal skeletal length. Moreover, we found a number of novel morphological features in the upper layers: Neurons in layer 2/3 showed a strong trend of a decreasing width of their dendritic arbor and a smaller tuft with increasing cortical depth. In layer 4, morphologies showed area-specific variation: atufted neurons were primarily located in the primary visual cortex, while tufted neurons were more abundant in higher visual areas. Finally, layer 4 neurons in V1 on the border to layer 5 showed a tendency towards avoiding layer 5 with their dendrites.

2 Results

2.1 Self-supervised learning of embeddings for 30,000 excitatory neurons from visual cortex

Our goal was to perform a large-scale census of the dendritic morphologies of excitatory neurons without prescribing a-priori which morphological features to use. Therefore, we used machine learning techniques [26] to learn the features directly from the neuronal morphology.

Our starting point was a $1.3 \times 0.87 \times 0.82\,\mathrm{mm}^3$ volume of tissue from the visual cortex of an adult P75–87 mouse, which has been densely reconstructed using serial section electron microscopy [2]. This volume has been segmented into individual cells, including non-neuronal types and more than 54,000 neurons whose soma was located within the volume. From these detailed reconstructions we extracted each neuron's dendritic tree and represented it as a skeleton (Fig. 1B) [1]: each neuron's dendritic morphology was represented as a graph, where each node had a location in 3d space. This means we focused on the location and branching patterns of the dendritic tree, not fine-grained details of spines or synapses (see companion paper [22]), or any subcellular structures (see companion paper [4]).

Our next step was to embed these graphs into a vector space that defined a measure of similarity, such that similar morphologies were mapped onto nearby points in embedding space (Fig. 1B). To do so, we employed a recently developed self-supervised learning method called GRAPHDINO [26] that learns semantic representations of graphs without relying on manual annotations. The idea of this method is to generate two "views" of the same input by applying random identity-preserving transformations such as rotations around the vertical axis, slightly perturbing node locations or dropping sub-branches (Fig. 1B, top and bottom). Then both views are encoded using a neural network. The neural network is trained to map both views onto similar vector embeddings. For model training, the data was split into training, validation and test data to ensure that the model did not overfit (Sec. 4.5). The model outputs a 32-dimensional vector for each neuron that captures the morphological features of the neuron's dendritic tree. Thus, each neuron is represented as a point in this 32-dimensional vector space (Fig. 1C).

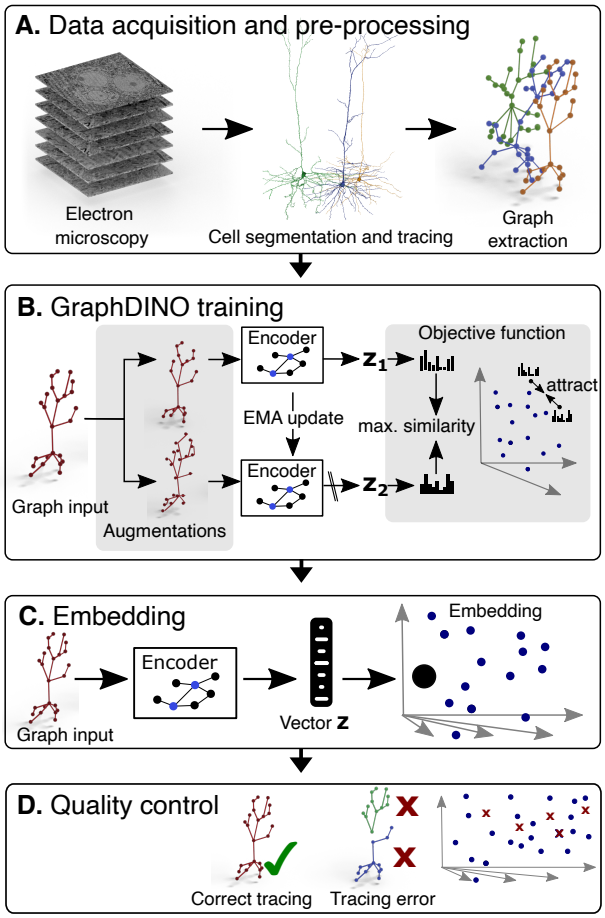

Figure 1: **Pipeline to generate vector embeddings for large scale datasets that capture the morphological features of the neurons' dendritic trees.** **A** Imaging of brain volume via electron microscopy and subsequent segmentation and tracing to render 3D meshes of individual neurons that are used for skeletonization. **B** Self-supervised learning of low dimensional vector embeddings z_1, z_2 that capture the essence of the 3D morphology of individual neurons using GraphDINO. Two augmented "views" of the neuron are input into the network, where the weights of one encoder (bottom) are an exponential moving average (EMA) of the other encoder (top). The objective is to maximize the similarity between the vector embeddings of both views. Vector embeddings of similar neurons are close to each other in latent space. **C** An individual neuron is represented by its vector embedding as a point in the 32-dimensional vector space. **D** Quality control to remove neurons with tracing errors.

At this stage, we performed another quality control step: Using the learned embeddings as a similarity metric between neurons, we clustered the neurons into 100 clusters and manually inspected the resulting clusters. We found a non-negligible fraction of neurons whose apical dendrite left the volume or was lost during tracing (see Methods for details). We removed neurons whose somata are in close proximity to the imaged volume boundary (Fig. 2A). Additionally, we used the clusters containing fragmented neurons as examples for broken neurons and trained a classifier to predict whether a neuron has reconstruction errors using the learned morphological embeddings as input features (Fig. 2B, Fig. A.2A, B). We then removed all neurons from the dataset that were classified as erroneous. Also, at this point we removed all interneurons from the dataset since we focused on excitatory neurons in this paper (Fig. 2C, Fig. A.2B, C).

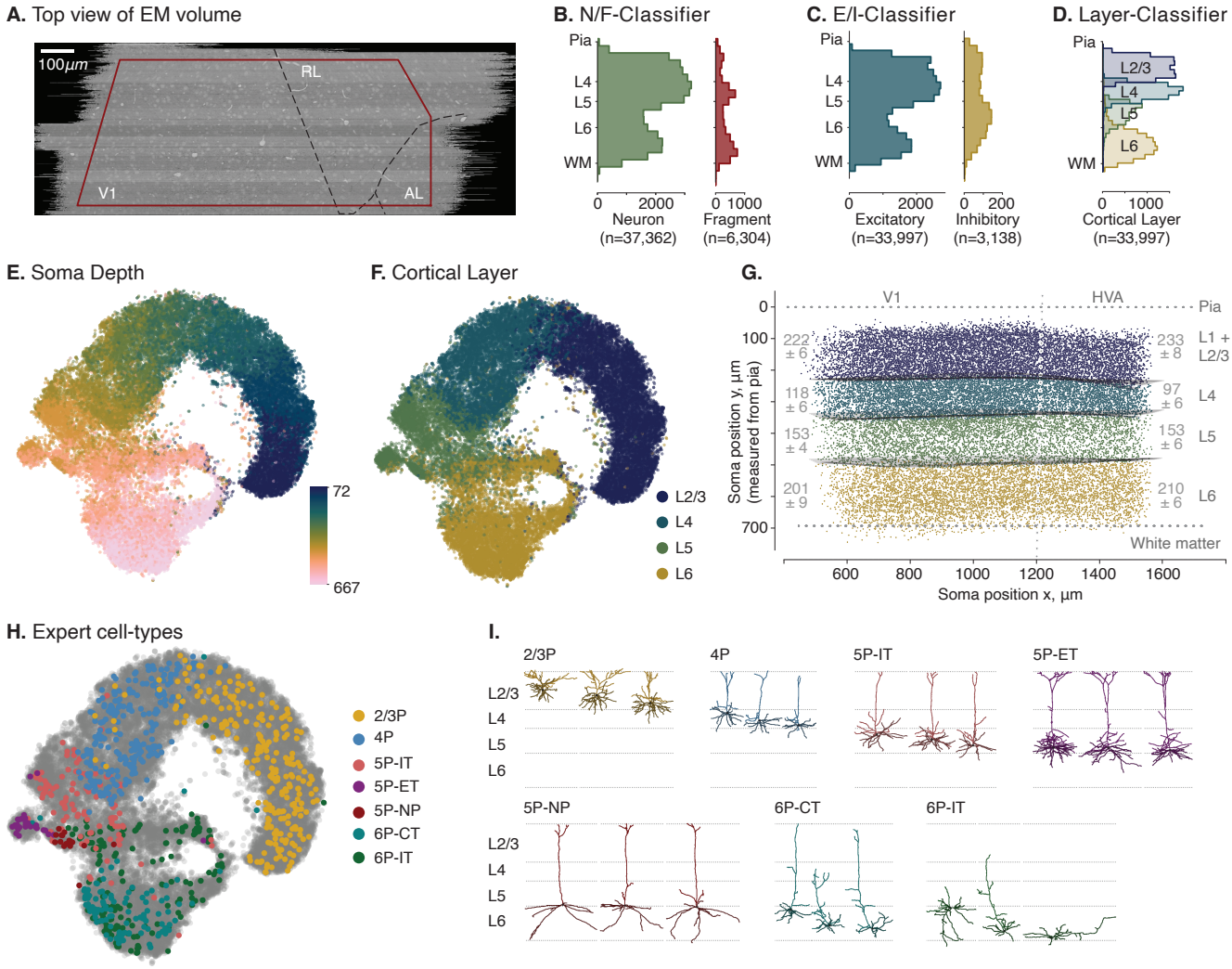

Figure 2: **Visualization of soma depths and cortical layer assignments of excitatory neuronal morphologies showing mostly a continuum with distinct clusters only in deeper layers. A.** Top view of the EM volume with approximate visual areas indicated. All neurons with their soma origin within the red boundary were used for analysis. **B.** Distribution of complete neurons and fragments along cortical depth as determined by our classifier based on the morphological embeddings. **C.** Distribution of excitatory neurons and interneurons along cortical depth. **D.** Classifier prediction for cortical layer origin based on learned the morphological embeddings. **E.** t-SNE embedding (perplexity 300) of the vector embeddings of excitatory neuronal morphologies colored by the respective soma depth of the neurons relative to the pia ($n = 33{,}997$). **F.** t-SNE embedding colored by cortical layer assignments as predicted by a cross-validated classifier trained on the morphological embeddings as features and a subset of manually labeled excitatory neurons ($n = 922$). **G.** Cross-section of the brain volume depicting soma positions of neurons colored by their assigned cortical layer. Cortical layer thicknesses for primary visual cortex (V1) (left) and higher visual areas (HVA) (right) given as mean $\pm$ standard deviation. **H.** t-SNE embedding of excitatory neuronal morphologies colored by expert-defined cell types. **I.** Example morphologies of the expert-defined cell types.

2.2 Dendritic morphologies form mostly a continuum with distinct clusters only in deeper layers

We computed the vector embeddings of all excitatory neurons in our volume, which spanned the mouse visual areas V1, RL and AL (Fig. 2A). Dendritic morphology followed mostly a continuum that tracked the cortical depth of the soma from the pia, in counter-clockwise direction in Fig. 2E from layer 2/3 to layer 6. Note that the soma location within the cortex was not provided to the model, but the soma was centered on the origin of the coordinate system. Cells were mostly organized along a continuum and only a few distinct clusters were visible in the deeper layers. Therefore we decided to not assume the existence of clusters a-priori, as many previous studies [6, 15, 10, 14], but instead investigated the major axes of variation within the morphological embedding space.

The learned embedding space also reflects known cell types (Fig. 2H, I) that were assigned by expert neuroanatomists [22] using cortical origin of the somata and their long-range projection type (IT: inter-telencephalic or inter-cortical; ET: extra-telenceophalic or subcortical projecting, NP: near projecting, and CT: Corticothalamic). Note that both sources of information were not provided to the model, showing that dendritic morphology alone is sufficient to infer these broad cell type classes. One exception are the 6P-CT and 6P-IT cells, who were partly intermingled in embedding space. 6P-IT cells show a high variance in their dendritic morphology which in some cases are indistinguishable from 6P-CT cells when no information about the projection type is used (Fig. 2H, I).

Since an organization into cortical layers is well established, we separated cells by cortical layer to study the morphological rules of organization. We determined the layer boundaries by training a classifier using our 32-dimensional embeddings and a set of 922 neurons manually assigned to layers by experts (Fig. 2D, F, G). As expected, the inferred layer boundaries indicated that layer 4 was approximately 20% thicker in V1 than in higher visual areas RL and AL (Fig. 2G; mean $\pm$ SD: $118 \pm 6\,\mu\text{m}$ in V1 vs. $97 \pm 6\,\mu\text{m}$ in HVA), the difference being compensated for by layers 2/3 and 6 each being approximately $10\,\mu\text{m}$ thinner. In the following we proceed by assigning neurons to layers based on their soma location relative to these inferred boundaries.

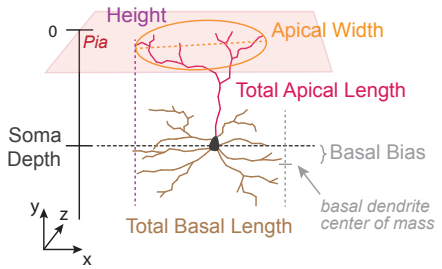

Figure 3: **Schematic of morphometric descriptors computed from neuronal skeletons and their labeled compartments.** SOMA DEPTH. Depth of the centroid of the soma relative to the pia. HEIGHT. Extent of the cell in y-axis. TOTAL APICAL LENGTH. Total length of the skeletal branches of the apical dendrites. APICAL WIDTH. Maximum extent of the apical dendritic tree in the xz-plane. TOTAL BASAL LENGTH. Total length of the skeletal branches of the basal dendrites. BASAL BIAS. Depth in y-axis of center of mass of basal dendrites relative to the soma.

To visualize the main axes of morphological variation within each layer, we performed nonlinear dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE; [24]) and identified a number of morphological features that formed major axes of variation within the two-dimensional space (Fig. 4).

What do these axes of variation in the two-dimensional t-SNE embeddings mean in human-interpretable terms? To answer this question, we looked for morphological metrics that form gradients within the t-SNE embedding space. Based on visual inspection, we found the following six morphological metrics to account well for a large fraction of the dendritic morphological diversity in our dataset (see Fig. 3 for an illustration): (1) depth of the soma relative to the pia, (2) height of the cell, (3) total length of the apical dendrites, (4) width of the apical dendritic tree, (5) total length of the basal dendrites, (6) the location of the basal dendritic tree relative to the soma ("basal bias").

2.3 Layer 2/3: Width and length of apical dendrites decrease with depth

We start with layer 2/3 (L2/3) where we found a continuum of dendritic morphologies that formed a gradient from superficial to deep, with deeper neurons (in terms of soma depth) becoming thinner and less tufted (Fig. 4A1, A2, A3). The strongest predictors of the embeddings were the depth of the soma relative to the pia and the total height of the cell ($R^2 > 0.9$; Fig. 4B). These two metrics were also strongly correlated (Spearman's rank correlation coefficient, $\rho = 0.93$; Fig. 4C), since nearly all L2/3 cells had an apical dendritic tree that reached to the pial surface (see example morphologies in Fig. 4A, top). L2/3 cells varied in terms of their degree of tuftedness: both total length and width of their apical tuft decreased with the depth of the soma relative to the pia (Fig. 4A2, A3). L2/3 cells also varied along a third axis: the skeletal length of their basal dendrites (Fig. 4A4), but this property was not strongly correlated with either soma depth or shape of the apical dendrites (Fig. 4C).

2.4 Layer 4: Small or no tufts and some cells' basal dendrites avoid layer 5

The dendritic morphology of layer 4 (L4) was again mostly a continuum and appeared to be a continuation of the trends from L2/3: The skeletal length of the apical dendrites was shorter, on average, than that of most L2/3 cells (Fig. 4A2) and approximately 20% of the cells were untufted. Within L4 the total apical skeletal length was not correlated with the depth of the soma ($\rho = 0.0$; Fig. 4C), suggesting that it forms an independent axis of variation. There was also quite some variability in terms of the total length of the basal dendritic tree, but – as in L2/3 – it was not correlated with any of the other properties.

Our data-driven embeddings revealed another axis of variation that had previously not been considered important: the location of the basal dendritic tree relative to the soma ("basal bias"; Fig. 3). We found that many L4 cells avoided reaching

into L5 with their dendrites (Fig. 4A3). As a result, the depth of the basal dendrites was anticorrelated with the depth of the soma ($\rho = -0.29$; Fig. 4A3 and Fig. 4C). We will come back to this observation later.

2.5 Layer 5: Thick-tufted cells stand out

Layer 5 (L5) showed a less uniformly distributed latent space than L2/3 or L4. Most distinct was the cluster of well-known thick-tufted pyramidal tract (PT) cells [8, 6, 10, 14, 15] on the bottom right (Fig. 4A4, light green points). These cells accounted for approximately 17% of the cells within L5 (based on a classifier trained on a smaller, manually annotated subset of the data; see methods). They were restricted almost exclusively to the deeper half of L5 (Fig. 4A1, A4 inset 2, C inset top right). They have the longest skeleton for all three dendritic compartments: apical, basal and oblique.

Another morphologically distinct type of cell was apparent: the near-projecting (NP) cells [11, 6] with their long and sparse basal dendrites (Fig. 4A4, inset 3). These cells accounted for approximately 4% of the cells within L5. They tended to send their dendrites deeper (relative to the soma), had little or no obliques and tended to have small or no apical tufts. However, the dendritic morphology of this cell type appeared to represent the extreme of a continuum rather than being clearly distinct from other L5 cells.

The remaining roughly 80% of the cells within L5 varied continuously in terms of the skeletal length of the different dendritic compartments. While there was a correlation between apical and basal skeletal length (apical vs. basal: $\rho = 0.43$; Fig. 4C), there was also a significant degree of diversity. Within this group there was no strong correlation of morphological features with the location of the soma within L5 (depth vs. apical length $\rho = 0.2$, depth vs. basal $\rho = 0.06$; Fig. 4C).

In upper L5 we found a group of cells that resembled the L4 cells whose dendrites avoid L5 (Fig. 4A4, inset 1). This type of cell was restricted to the uppermost portion of L5 and morphologically resembled L4 cells by being mostly atufted. We refer to these cells as displaced L4 cells. The presence of these cells suggests there are no precise laminar boundaries based on morphological features of neurons, but instead different layers blend into one another, a finding also observed by other authors [15, 4].

2.6 Layer 6: Long and narrow, oblique and inverted pyramidal neurons

Dendritic morphology in layer 6 (L6) also formed a continuum with a large degree of morphological diversity. The dominant feature of L6 was the large variety of cell heights ($R^2 > 0.9$; Fig. 4B). Overall, the height of a cell was not strongly correlated with its soma's location within L6 ($\rho = -0.13$; Fig. 4C). Unlike other layers, where the apical dendrites usually reach all the way up to layer 1, many cells in L6 had shorter apical dendrites. However, due to tracing errors, our analysis overestimates the number of such short cells. We therefore manually inspected 183 putative untufted early-terminating neurons within L6 and found that, among those, 45 % were incompletely traced, whereas 55 % were true untufted cells whose apical dendrite terminated clearly below L1.

As described previously [6, 15], the dendritic tree of L6 cells was narrower than in the layers above. Also consistent with previous work, we also found a substantial number of horizontal and inverted pyramidal neurons, where the apical dendrite points sideways or downwards (Fig. 4A4, inset 1 & 6).

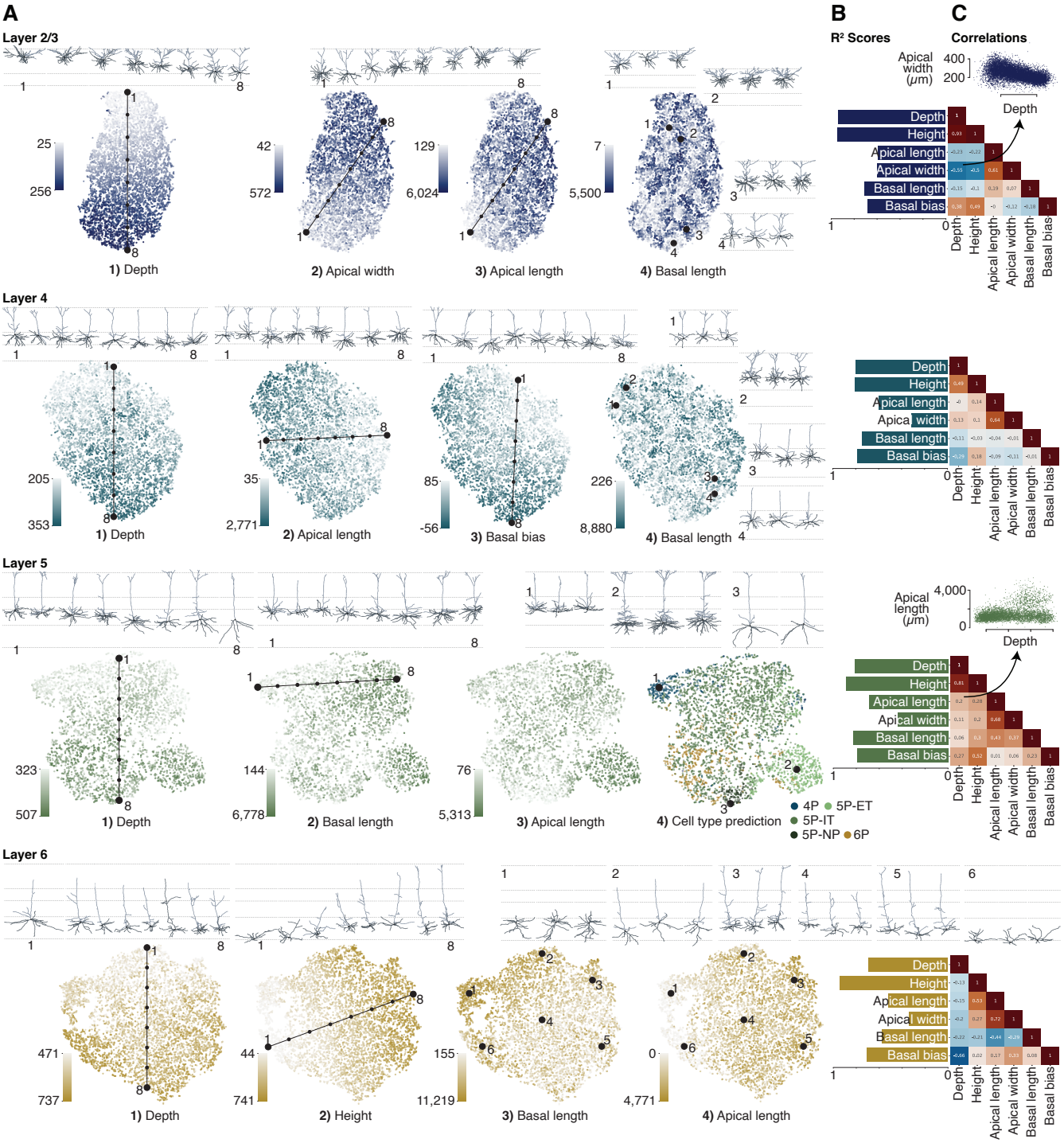

Figure 4: **t-SNE visualization of vector embeddings per cortical layer reveal axis of variation in neuronal morphologies. A.** t-SNE embeddings per layer colored by percentiles of various morphometric descriptors with example neuronal morphologies along the axis of variation displayed above the embedding. **B.** R^2 scores of the six morphometric descriptors (see Fig. 3) per layer showing the strength as predictors of the 32d embeddings. **C.** Spearman's rank correlation coefficient between morphometric descriptors per layer. **Layer 2/3** (blue) Continuum of dendritic morphologies with thinner and less tufted neurons in increasing distance to the pia. **Layer 4** (turquoise) Continuation of L2/3 trends with shorter apical dendrites and more untufted cells. Many cells avoid reaching dendrites into L5 (basal bias). **Layer 5** (green) Clustering of thick-tufted PT and NP cells. Upper L5 cells resemble L4 cells that avoid reaching into L5, indicating too strict laminar borders. **Layer 6** (orange) Continuum with a large morphological diversity e.g. in cell heights, and existence of horizontal and inverted pyramidal neurons.

2.7 Pyramidal neurons are less tufted in V1 than in higher visual areas

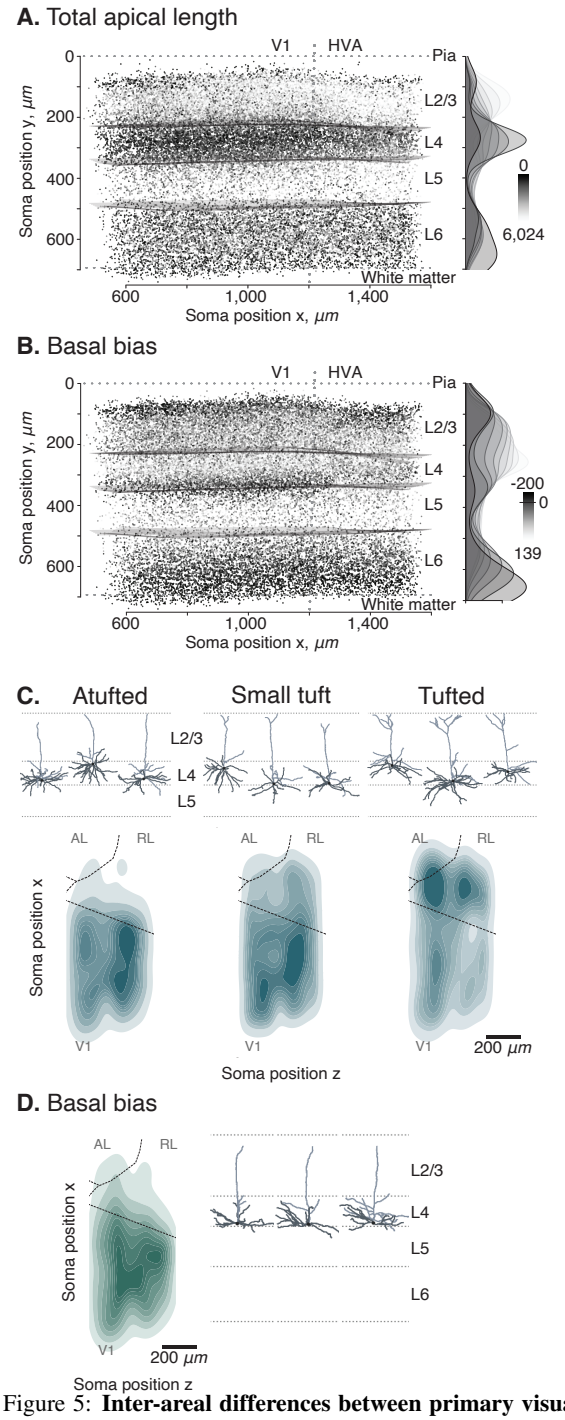

Figure 5: **Inter-areal differences between primary visual cortex (V1) and higher visual areas (HVAs). A.** Side view showing apical skeletal length, color-coded by percentiles (dark=short, bright=long). Projection from the side orthogonal to the V1/HVA border after a 14 degree rotation around y-axis (vertical dashed line); top: pia; bottom: white matter. **B.** Side view showing basal bias (as in A) (dark=negative basal bias: center of mass of basal dendrites is above the soma; bright=positive basal bias: center of mass of basal dendrites is below soma). **C.** Top view showing density of untufted (left), small tufted (middle) and tufted (right) L4 cells. Atufted neurons are mostly confined to V1, while tufted neurons are more abundant in HVA. Dashed lines: area borders between primary visual cortex (V1), anterolateral area (AL) and rostrolateral area (RL), estimated from reversal of the retinotopic map measured using functional imaging. **D.** Top view (as in C) showing horizontal distribution of L4 cells whose dendrites avoid reaching into L5 and who are mostly located in V1.

After our layer-wise survey of excitatory neurons' morphological features, we next asked whether there are inter-areal differences between primary visual cortex (V1) and higher visual areas (HVAs). The total length the apical dendrites of neurons in V1 was significantly shorter than for neurons in HVA (Fig. 5A): For L2/3, neurons in V1 had on average 16% shorter apical branches than in HVA (mean $\pm$ SD: $1{,}423 \pm 440\,\mu\text{m}$ in V1 vs. $1{,}688 \pm 554\,\mu\text{m}$ in HVA; *t-test*: $p < 0.0025$, Cohen's $d = 0.53$). Similarly, in L4, neurons in V1 had on average 16% shorter apical branches than L4 neurons in HVA ($851 \pm 264\,\mu\text{m}$ vs. $1{,}019 \pm 313\,\mu\text{m}$; $p < 0.0025$, $d = 0.58$). In L5, neurons in V1 had on average 14% shorter apical branches than L5 neurons in HVA ($1{,}326 \pm 661\,\mu\text{m}$ vs. $1{,}549 \pm 745\,\mu\text{m}$; $p < 0.0025$, $d = 0.32$). While the trend continued in L6, the difference in apical length between V1 and HVA neurons was smaller. There was only a 4% increase in apical length in HVA compared to V1 ($1{,}112 \pm 383\,\mu\text{m}$ vs. $1{,}159 \pm 397\,\mu\text{m}$; $p < 0.0025$, $d = 0.12$). For this analysis, only neurons with identified apical dendrites were taken into account (see companion paper; Celii et al. [1]).

Upon closer inspection, we observed that L4 contained substantially more untufted neurons than higher visual areas RL and AL (Fig. 5A). We clustered each layer's morphological embeddings into 15 clusters using a Gaussian Mixture Model and looked for clusters that were restricted to particular brain areas. Clusters that were clearly confined to V1 or HVAs were primarily found in L4. When classifying (manually, at the cluster-level) L4 neurons into untufted, small tufted and tufted, we observed that untufted neurons were almost exclusively located in V1, while tufted neurons were more frequent in HVAs (Fig. 5C).

2.8 Layer 4 cells avoiding layer 5 are located primarily in primary visual cortex

We observed a second area difference, which was related to the novel morphological cell type in L4. Recall that these cells' dendrites avoid reaching into L5. Interestingly, these cells were located in a very narrow strip of around $50\,\mu\text{m}$ above the L4–L5 border (Fig. 5B). Moreover, they were also untufted and almost exclusively located in V1 (Fig. 5D).

3 Discussion

In summary, our data-driven unsupervised learning approach identified the known morphological features of excitatory cortical neurons' dendrites and enabled us to make four novel observations: (1) Superficial L2/3 neurons are wider than deep ones; (2) L4 neurons in V1 are less tufted than those in HVAs; (3) we discovered an untufted L4 cell type that is specific to V1 whose basal dendrites avoid reaching into L5; (4) excitatory cortical neurons form mostly a continuum with respect to dendritic morphology, with some notable exceptions.

First, our finding that superficial L2/3 neurons are wider than deeper ones is clearly visible in the data both qualitatively and quantitatively. A similar observation has been made recently in concurrent work [25].

Second, the trend of deeper neurons being less tufted continues into L4 where a substantial number of cells are completely untufted. Here we see a differentiation with respect to brain areas: completely untufted cells are mostly restricted to V1 while HVA neurons in L4 tend to be more tufted. Why would V1 neurons be less tufted than those in higher visual areas? V1 – as the first cortical area for visual information processing – and L4 – as the input layer, in particular – might be less modulated by feedback connections than other layers and higher visual areas. Therefore, these neurons might sample the feedback input in L1 less than other neurons.

Third, we found that some neurons in L4 of V1 avoid reaching into L5 with their dendrites. To our knowledge, this morphological pattern has not been described before in the visual cortex. Retrospectively, it can be observed in Gouwens and colleagues' data: their spiny m-types 4 and 5, which are small- or untufted L4 neurons, show a positive basal bias (assuming their "basal bias y" describes the same property; Gouwens et al. [6]; Suppl. Fig. 15). What function could this avoiding L5 have? Similarly to the nonexisting tuft of these neurons, avoiding L5 could support these neurons in focusing on the thalamic input (which targets primarily L4) and, thus, represent and distribute the feedforward drive within the local circuit. It is therefore tempting to speculate that these untufted, L5-avoiding L4 neurons might be precursors of spiny stellate cells, which are nearly absent in the mouse visual cortex [20], but exist only in somewhat more developed sensory areas like barrel cortex or in cat and primate V1.

Fourth, except from the well-known L5 thick-tufted pyramidal-tract (PT) neurons that form a cluster in L5, our data and methods suggest that excitatory neurons in the mouse visual cortex form mostly a continuum with respect to dendritic morphology. This result does not rule out the possibility that there are in fact distinct types; it simply suggests that features beyond dendritic morphology need to be taken into account to clearly identify them. For instance, the results of [22] suggest that the 5P-NP cells can be separated from other layer 5 pyramidal neurons by considering the class of interneurons that target them. It is also not guaranteed that our data-driven method identifies all relevant morphological features. Every method has (implicit or explicit) inductive biases. We tried to stay clear of explicit human-defined features, but by choosing a graph-based input representation we provide different inductive biases than, for instance, a voxel-based representation or one based on point clouds. However, the fact that we could reconcile known morphological features, discover novel ones and achieve excellent classification accuracy on an annotated subset of the data suggests that our learned embeddings indeed contain a rich and expressive representation of a neuron's dendritic morphology.

Our observation that morphologies formed mostly a continuum is in line with a recent study in motor cortex examining the relationship between transcriptomic and morphological cell types [21]. These authors found a substantial degree of (continuous) morphological variation within transcriptomically defined cell types. Moreover, they found that morphological and transcriptomic features correlated, suggesting a more fine-grained organization of neurons into a relatively small number of distinct and broad "families," each of which exhibits substantial continuous variation among its family members. Our analysis support this notion: excitatory cells can be mostly separated by layers into roughly a handful of families, each of which contains a substantial degree of variation in terms of morphology, which might also co-vary with other modalities.

4 Methods

4.1 Dataset

The dataset consists of a $1.3 \times 0.87 \times 0.82\ mm^3$ volume of tissue from the visual cortex of an adult P75-87 mouse, which has been densely reconstructed using serial section electron microscopy (EM) [2]. We here use the subvolume 65, which covers approcimately $1.3 \times 0.56 \times 0.82\ mm^3$. It covers all layers of cortex and spans primary visual cortex (V1) and two higher visual areas, anterolateral area (AL) and rostrolateral area (RL). We refer to the original paper on the dataset [2] for details on the identification and morphological reconstruction of individual neurons.

4.2 Skeletonization and cell compartment label assignment

The EM reconstructions yield neuronal meshes. These meshes might be incomplete or exhibit different kinds of errors including merges of other neuronal or non-neuronal compartments onto the neurons. Therefore an automatic proof-reading pipeline that results in neuronal skeletons was executed (companion paper; Celii et al. [1]).

For the skeletal detection from the reconstructed meshes, the meshes were first downsampled to 25% of their resolution and made water tight. Then glia and nuclei meshes were identified and removed. For the remaining meshes the locations of the soma was identified using a soma detection algorithm [27]. Each neurite submesh was then skeletonized using a custom skeletonization algorithm which transformed axonal and dendritic processes into a series of line segments to obtain the skeleton (companion paper; Celii et al. [1]). For each skeleton, the highest probability axon subgraph was determined and all other non-soma nodes were labeled as dendrites. A final heuristic algorithm classifies subgraphs of dendritic nodes into compartments, such as apical trunks generally projecting from the top half of somas and with a general upward trajectory and obliques as projections off the apical trunks at an approximate 90 degree angle. For further details on the compartment label assignment please see companion paper [1].

4.3 Coordinate transformations

The EM volume is not perfectly aligned. First, the pial surface is not a horizontal plane parallel to the (x, z) axis, but instead slightly tilted. Second, the thickness of the cortex varies across the volume such that the distance from pia to white matter is not constant. Without any pre-processing, an unsupervised learning algorithm would pick up these differences and, for instance, find differences of layer 6 neurons across the volume simply because in some parts of the volume they tend to be located deeper than in others and their apical dendrites that reach to layer 1 tend to be larger. Using *relative* coordinates can solve such issues if pia and white matter correspond to planes (approximately) parallel to the (x, z) plane. To transform our coordinate system in such standardized coordinates, we first applied a rotation about the z-axis of 3.5 degrees. This transformation removes the systematic rotation with respect to the native axes (Fig. A.1B). To standardize measurements across depth (y axis) and to account for differential thickness of the cortex, we estimated the best linear fit for both pial surface and white matter boundary by using a set of manually placed points, which were located on a regular grid along (x, z) with a spacing of $25\mu\mathrm{m}$. Then for each (x, z) coordinate, the y coordinate was normalized such that the pia's z coordinate corresponds to the average depth of the pia and the same for the white matter. This transformation results in an approximation of the volume where pia and white matter boundaries are horizontal planes orthogonal to the y axis and parallel to the (x, z) plane. Fig. A.1C shows example neurons before and after normalization. All training and subsequent analysis were performed on this pre-processed data.

4.4 Expert cell type labels

For a subset of the neurons in the volume experts labeled neurons according the following cell types: layer 2/3 and 4 pyramidal neurons, layer 5 near-projecting (NP), extra-telencenphalic (ET) and inter-telencenphalic (IT) neurons, layer 6 inter-telencenphalic (IT) and cortico-thalamic (CT) neurons, Martinotti cells (MC), basket cells (BC), bipolar cells (BPC) and neurogliaform cells (NGC). Cell types were assigned based on visual inspection of individual cells taking into account morphology, synapses and connectivity, nucleus features and their (x, y, z) location. All neurons were taken from one $100\,\mu\mathrm{m}$ column in the primary visual cortex (see companion paper, Schneider-Mizell et al. [22]). We did not use neurons with expert labels to train GRAPHDINO, but used them only for evaluation.

4.5 Morphological feature learning using GRAPHDINO

For learning morphological features in an unsupervised, purely data-driven way, we use a recently developed machine learning method called GRAPHDINO [26]. GRAPHDINO maps the skeleton graph of a neuron onto a 32-dimensional feature vector, which we colloquially refer to as the neuron's "bar code." For training GRAPHDINO, each neuron's skeleton is represented as an undirected graph $G = (V, E)$. V is the set of nodes $\{v_i\}_{i=1}^{N}$ and E the set of undirected edges $E = \{e_{ij} = (v_i, v_j)\}$ that connect two nodes v_i, v_j. Each node has a feature vector attached to it that holds the 3D Cartesian coordinate of the node, relative to the soma of the neuron, which has the coordinate $(0, 0, 0)$, i.e. is at the origin of the coordinate system. Because axons are not well reconstructed in the data yet, we focus on the dendritic skeleton only and remove segments labeled as axon. We train GRAPHDINO on a subset of the dataset, retaining 5,113 neurons for validation and 2,941 neurons for testing. The test set is chosen to contain the 1,011 neurons that were labeled by expert anatomists into morphological cell types (Sec. 4.4; [22]). The training and validation sets are i.i.d. sampled from the remaining neurons with a 90%-10% split (Fig. A.3).

GRAPHDINO is trained by generating two "views" of the same input graph by applying random identity-preserving transformations (described below). These two views are both encoded by the same neural network. The training objective is to maximize the similarity between the embeddings of these two views. To obtain the two views of one input graph, we subsampled the graph, randomly rotated it around the y-axis (orthogonal to pia), dropped subbranches and perturbed node locations. When subsampling the graph, we randomly dropped all but 200 nodes, always retaining the branching points. Rotations around the y-axis were uniformly distributed around the circle. During subbranch deletion we removed $n = 5$ subbranches. For node location jittering we used $\sigma = 1$. In addition the entire graph was randomly translated with $\sigma = 1$. For further details on the augmentation strategies, see Weis et al. [26].

The ADJACENCY-CONDITIONED ATTENTION network architecture we used had seven AC-ATTENTION layers with four attention heads each. The dimensionality of the latent representation z was set to 32 and the dimensionality of the projection p was $5,000$. All other architecture details are as described in the original paper [26]. For training we used the Adam optimizer [12] with a batch size of 128 for $50,000$ iterations. The learning rate increased linearly to 10^{-3} during the first 1,000 iterations and then decayed using a exponential schedule with a decay rate of 0.5.

4.6 Morphological clustering

For qualitative inspection of the data and the analyses in Fig. 5C+D we clustered the neurons using the learned bar code of each neuron's morphological features. We fit a Gaussian Mixture model (GMM) with diagonal covariance matrix using scipy [17] on the whole dataset as well as per cortical layer using 60 clusters and 15 clusters, respectively. As we found no evidence that these clusters (or any other clustering with fewer or more clusters) represent distinct cell types, we do not use this clustering to define cell types, but rather think of them as modes or representing groups of neurons with similar morphological features.

4.7 Data quality control steps

The dataset was generated by automatic segmentation of EM images and subsequent automatic processing into skeletons. As a consequence, not all cells were reconstructed perfectly. There was a significant fraction of wrongly merged or incompletely segmented cells. We used a combination of our learned GRAPHDINO embeddings and supervised classifiers trained on a subset of the neurons ($n = 1,011$) which were manually proofread and annotated by experts (see Sec. 4.4 and companion paper, Schneider-Mizell et al. [22]). Our quality control pipeline was as follows: First, we computed GRAPHDINO embeddings on the full dataset of 54,192 neurons (including both excitatory and inhibitory neurons). Next, we removed neurons which were close to the boundaries of the volume, as these neurons were only partly reconstructed. After this step we were left with 43,666 neurons. Within this dataset we identified neurons which were incorrectly reconstructed using a supervised classifier described in the next section, reducing the dataset to 37,362 neurons. Subsequently, we identified interneurons using a supervised classifier described in the next section, reducing the dataset to 33,997 excitatory neurons. Finally, on this dataset we manually proofread around 480 atufted neurons. As a result, we identified and removed another set of 2,684 neurons whose reconstructions were incomplete, leaving us with a final sample size of 31,313 putative excitatory and correctly reconstructed neurons for our main analyses.

4.8 Supervised classifiers

To identify reconstruction errors and interneurons, we used a subset of the dataset ($n = 1,011$) that was manually proofread and annotated with cell type labels by experts (see Sec. 4.4 and companion paper, Schneider-Mizell et al. [22]). Based on these and additional neurons we identified, we trained classifiers to detect segmentation errors, inhibitory cells and cortical

layer membership using our learned 32-dimensional bar codes of the neurons' skeletons (see Sec. 4.5). In our subsequent analysis, we focused on neurons that were identified as complete and excitatory by our classifier and used the inferred cortical layer labels to perform layer-specific analyses.

For all classifiers, we used ten-fold cross-validation on a grid search to find the best hyperparameters. We tested logistic regression with the following hyperparameters: type of regularization (none, L1, L2 or elastic net), regularization weight ($C \in 0.5, 1, 3, 5, 10, 20, 30$) and whether to use class weights that are inversely proportional to class frequencies or no weights. In addition, we tested support vector machines (SVMs) with the following hyperparameters: type of kernel (Linear, RBF or polynomial), L2 regularization weight ($C \in 0.5, 1, 3, 5, 10, 20, 30$) and degree of polynomial ($d \in 2, 3, 5, 7, 10, 20$ for the polynomial kernel and whether to use class weights or no weights. After having determined the optimal hyperparameters using cross-validation, we retrained the classifier using the optimal hyperparameters on its entire training set.

Removal of fragmented neurons. To remove fragmented neurons prior to analysis, we trained a classifier to differentiate between the manually proofread neurons from all layers ($n = 1,011$) and fragmented cells ($n = 240$). We identified fragmented cells using our clustering of the bar codes of the whole dataset without boundary neurons ($n = 43,666$) into 25 clusters per layer and manually identifying clusters that contained fragmented cells (2–3 clusters per layer). We then sampled 60 fragmented cells per layer as training data for our classifier.

We trained a support vector machine (SVM) using cross-validation as described above. Its cross-validated accuracy was 95%. The best hyperparameters were: polynomial kernel of degree 4 and $C = 3$. We used those hyperparameters to retrain on the full training set of 1,251 neurons. Using this classifier, we inferred whether a neuron is fragmented for the entire dataset ($n = 43,666$). We then removed cells predicted to be fragmented ($n = 6,304$) from subsequent analyses.

To validate the classification into fragmented and whole cells, we manually inspected ten neurons that were not in "fragmented" clusters before classification, but were flagged as fragmented by the classifier. Nine out of the ten had missing segments due to segmentation errors or due to apical dendrites leaving the volume.

Removal of inhibitory neurons. Analogously, we trained a classifier to predict whether a neuron was excitatory or inhibitory by using the manually proofread and annotated neurons ($n = 1,011$) (Sec. 4.4). As input features to the classifier we used our learned bar codes and additionally two morphometric features: synaptic density on apical shafts (number of synapses per micrometer of skeletal length except those located on spines) and spine density (number of spines per micrometer of skeletal length). These two features have been shown to separate excitatory from inhibitory neurons well in previous work (see companion paper, Celii et al. [1]). The annotated dataset contained 922 excitatory and 89 inhibitory neurons.

We trained a logistic regression. Its cross-validated accuracy was 92%. The best hyperparameters were: L2 regularization ($C = 5$) and using class weights. We used those hyperparameters to retrain on the full training set of 1,011 neurons. Using this classifier, we inferred whether a neuron was excitatory or inhibitory for the entire dataset after removing fragmented cells and after removal of 227 neurons that did not have spine and synapse densities available ($n = 37,135$). We then removed all inhibitory cells from subsequent analyses ($n = 3,138$).

Inference of cortical layers. To determine cortical layer labels for the entire dataset, we followed a two-stage procedure. First, we inferred the layer of each neuron using a trained classifier. Then we determined anatomical layer boundaries based on the optimal cortical depth that separates adjacent layers.

We first trained a support vector machine classifier for excitatory cells on the 922 manually annotated excitatory neurons by pooling the cell type labels per layer. Its cross-validated balanced accuracy was 89%. The best hyperparameters were: polynomial kernel of degree 5, $C = 3$. Using this classifier, we inferred the cortical layer of all excitatory neurons ($n = 33,997$; Fig. 2).

The spatial distribution of inferred layer assignments was overall well confined to their respective layers. As expected, there was some spatial overlap of labels at the boundaries, since layer boundaries are not sharp. We nevertheless opted for assigned neurons to layers based on their anatomical location rather than their inferred label. To do so, we determined the optimal piecewise linear function that separates two consecutive layers. At the end, the layer assignments were purely based on the soma depth of each neuron relative to the inferred layer boundaries – not on the classifier output.

Inference of coarse cell type labels. In Fig. 4 we show cell type labels for layer 5. These were determined by training a support vector machine classifying cell type labels for excitatory cells using the 922 manually annotated neurons. The cross-validated balanced accuracy of this classifier was 85%. The best hyperparameters were: polynomial kernel of degree 2, $C = 20$, using class weights. Using this classifier, we inferred cell type labels for all excitatory neurons (n=33,997).

4.9 Manual validation of apical skeletons

We found a significant fraction of atufted neurons across layers 4–6. To determine the extent to which these cells are actually atufted or an artifact of incomplete reconstructions, we manually inspected ca. 480 neurons in Neuroglancer [5] with respect to the validity of their apical termination. During manual inspection, we annotated neurons' reconstruction as "naturally terminating," "out-of-bounds," "reconstruction issue" or "unsegmented region." Reconstruction issues are the case where the EM slice was segmented correctly, but the tracing missed to connect two parts of the same neuron. Unsegmented regions are the case where one or multiple EM images or parts thereof were not segmented correctly and therefore the neuron could not be traced correctly. In addition, we classified the neurons as either "atufted," "small tufted" or "tufted," both before validation and after correcting reconstruction errors.

For layer 4, we inspected 120 atufted neurons. Of those, 64% had missing segments on their apical dendrites and 36% had a natural termination. Note, however, that 74% of the neurons had a consistent tuft before and after validation. Even though parts of the apical dendrite were missing, qualitatively the degree of tuftedness did not change. For atufted neurons this means that their apical dendrite merely terminated early, but this reconstruction error did not change their classification as atufted. In layer 4, neurons with a natural termination ended more superficially than neurons with missing segments. We therefore excluded L4 neurons from the analysis whose apicals ended more than 154 micrometers below the pia to exclude neurons with reconstruction errors from our analysis. This threshold was selected such that the F1-score was maximized, i.e. retaining as many atufted neurons with natural termination, while removing as many neurons with missing segments as possible. The threshold was computed on the 120 validated neurons. This process excluded 557 neurons from layer 4.

For layer 5, we inspected 176 neurons with early-terminating apical dendrites. Of those, 59 showed a natural apical termination, while 117 had reconstruction issues or left the volume. We found no clear quantitative metric like the depth of the apical to exclude neurons with unnatural terminations. Therefore, we excluded neurons based on their cluster membership from further analysis if the cluster contained more than 50% of neurons with unnatural terminations. Of the 15 clusters, we excluded four, corresponding to 1,258 out of 5,858 L5 neurons.

For layer 6, we inspected 183 neurons with early terminating apicals. Of those, 100 showed a natural apical termination, while 83 had reconstruction issues or left the volume. Due to the slant of the volume, long, narrow L6 cells near the volume boundary had a high likelihood of leaving the boundary with their apical dendrite. Therefore, we excluded all L6 neurons whose apical dendrite leaves the volume ($n = 867$) prior to our analysis. We considered a neuron as leaving the volume if the most superficial point of its apical tree was within a few micrometers of the volume boundary.

Overall, we excluded 2,684 neurons as a result of this manual validation step, resulting in a final sample size of 31,313 neurons used in our analysis (Figs. 4+5.

4.10 Cortical area boundaries

Cortical area boundaries were manually drawn from retinotopic maps of visual cortex taken before EM imaging. For further details see companion paper [2].

4.11 Dimensionality reduction

For visualization of the learned embeddings, we reduced the dimensionality of the 32d embedding vector to 2d using t-distributed stochastic neighbor embedding (t-SNE; [24]) using the openTSNE package [18] with cosine distance and a perplexity of 30 for t-SNE plots for individual cortical layers and a perplexity of 300 for the whole dataset.

4.12 Morphometric descriptors

We computed morphometrics based on the neuronal skeletons for analysis of the learned latent space. Morphometrics are not used for learning the morphological vector embeddings. We computed morphometrics based on compartment labels: soma, apical dendrites, basal dendrites and oblique dendrites (Sec. 4.2). They are visualized in Fig. 3. TOTAL APICAL LENGTH is defined as the total length of all segments of the skeletons that are classified as apical dendrites. TOTAL BASAL LENGTH is computed analogously. DEPTH refers to the depth of the soma centroid relative to the pia after volume normalization (Sec. 4.3), where pia depth is equal to zero. HEIGHT is the absolute difference between the highest and the lowest skeleton node of a neuron in y-direction. APICAL WIDTH refers to the widest extent of apical dendrites in the xz-plane. BASAL BIAS describes the difference between the soma depth and the center of mass of the basal dendrites along the y-axis.

4.13 Statistics

Apical lengths in Sec. 2.7 were compared between V1 and HVA per laminar layer with four independent two-tailed Student's t-tests. The single-test significance level of 0.01 was corrected for multiple tests using Bonferroni correction to 0.0025. Only neurons that had any nodes labeled as apical were taken into account for this analysis. In L2/3 $6,760$ neurons were taken into account from V1 and $3,436$ from HVA; for L4 $n = 5,217$ (V1) and $n = 2,534$ (HVA); for L5 $n = 3,708$ (V1) and $n = 1,924$ (HVA); and for L6 $n = 3,959$ (V1) and $n = 2,618$ (HVA).

Acknowledgements

M.A.W. was supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS), Tübingen. A.S.E. received funding for this project from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101041669). The authors thank David Markowitz, the IARPA MICrONS Program Manager, who coordinated this work during all three phases of the MICrONS program. We thank IARPA program managers Jacob Vogelstein and David Markowitz for co-developing the MICrONS program. We thank Jennifer Wang, IARPA SETA for her assistance. The work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003, D16PC00004, and D16PC0005. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government. We also thank the Allen Institute for Brain Science founder, Paul G. Allen, for his vision, encouragement and support. This work was also supported by the National Institute of Mental Health under Award Numbers R01 MH109556, P30EY002520 and the NSF NeuroNex program through grant NSF-1707400. A.T. acknowledges support from National Institute of Mental Health and National Institute of Neurological Disorders And Stroke under Award Number U19MH114830 and National Eye Institute/National Institutes of Health Core Grant for Vision Research T32-EY-002520-37.

References

[1] Brendan Celii, Stelios Papadopoulos, Zhuokun Ding, Paul G. Fahey, Eric Wang, Christos Papadopoulos, Alexander Kunin, Saumil Patel, J. Alexander Bae, Agnes L. Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J. Bumbarger, Manuel A. Castro, Erick Cobos, Sven Dorkenwald, Leila Elabbady, Akhilesh Halageri, Zhen Jia, Chris Jordan, Dan Kapner, Nico Kemnitz, Sam Kinn, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Gayathri Mahalingam, Eric Mitchell, Shanka Subhra Mondal, Shang Mu, Barak Nehoran, Sergiy Popovych, Casey M. Schneider-Mizell, William Silversmith, Marc Takeno, Russel Torres, Nicholas L. Turner, William Wong, Jingpeng Wu, Szi-chieh Yu, Wenjing Yin, Daniel Xenes, Lindsey M. Kitchell, Patricia K. Rivlin, Victoria A. Rose, Caitlyn A. Bishop, Brock Wester, Emmanouil Froudarakis, Edgar Y. Walker, Fabian H. Sinz, H. Sebastian Seung, Forrest Collman, Nuno Maçarico da Costa, R. Clay Reid, Xaq Pitkow, Andreas S. Tolias, and Jacob Reimer. Neurd: A mesh decomposition framework for automated proofreading and morphological analysis of neuronal em reconstructions. *bioRxiv*, 2023. doi: 10.1101/2023.03.14.532674.

[2] MICrONS Consortium, J. Alexander Bae, Mahaly Baptiste, Agnes L. Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J. Bumbarger, Manuel A. Castro, Brendan Celii, Erick Cobos, Forrest Collman, Nuno Maçarico da Costa, Sven Dorkenwald, Leila Elabbady, Paul G. Fahey, Tim Fliss, Emmanouil Froudarakis, Jay Gager, Clare Gamlin, Akhilesh Halageri, James Hebditch, Zhen Jia, Chris Jordan, Daniel Kapner, Nico Kemnitz, Sam Kinn, Selden Koolman, Kai Kuehner, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Gayathri Mahalingam, Sarah McReynolds, Elanine Miranda, Eric Mitchell, Shanka Subhra Mondal, Merlin Moore, Shang Mu, Taliah Muhammad, Barak Nehoran, Oluwaseun Ogedengbe, Christos Papadopoulos, Stelios Papadopoulos, Saumil Patel, Xaq Pitkow, Sergiy Popovych, Anthony Ramos, R. Clay Reid, Jacob Reimer, Casey M. Schneider-Mizell, H. Sebastian Seung, Ben Silverman, William Silversmith, Amy Sterling, Fabian H. Sinz, Cameron L. Smith, Shelby Suckow, Marc Takeno, Zheng H. Tan, Andreas S. Tolias, Russel Torres, Nicholas L. Turner, Edgar Y. Walker, Tianyu Wang, Grace Williams, Sarah Williams, Kyle Willie, Ryan Willie, William Wong, Jingpeng Wu, Chris Xu, Runzhe Yang, Dimitri Yatsenko, Fei Ye, Wenjing Yin, and Szi-chieh Yu. Functional connectomics spanning multiple areas of mouse visual cortex. *bioRxiv*, 2021. doi: 10.1101/2021.07.28.454025.

[3] Javier DeFelipe, Pedro L. López-Cruz, Ruth Benavides-Piccione, Concha Bielza, Pedro Larrañaga, Stewart Anderson, Andreas Burkhalter, Bruno Cauli, Alfonso Fairén, Dirk Feldmeyer, et al. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Reviews Neuroscience*, 14(3):202–216, 2013.

[4] Leila Elabbady, Sharmishtaa Seshamani, Shang Mu, Gayathri Mahalingam, Casey Schneider-Mizell, Agnes Bodor, J. Alexander Bae, Derrick Brittain, JoAnn Buchanan, Daniel J. Bumbarger, Manuel A. Castro, Erick Cobos, Sven Dorkenwald, Paul G. Fahey, Emmanouil Froudarakis, Akhilesh Halageri, Zhen Jia, Chris Jordan, Dan Kapner, Nico Kemnitz, Sam Kinn, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Eric Mitchell, Shanka Subhra Mondal, Barak Nehoran, Stelios Papadopoulos, Saumil Patel, Xaq Pitkow, Sergiy Popovych, Jacob Reimer, William Silversmith, Fabian H. Sinz, Marc Takeno, Russel Torres, Nicholas Turner, William Wong, Jingpeng Wu, Wenjing Yin, Szi-chieh Yu, Andreas Tolias, H. Sebastian Seung, R. Clay Reid, Nuno Maçarico Da Costa, and Forrest Collman. Quantitative census of local somatic features in mouse visual cortex. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.499976.

[5] Jeremy Maitin-Shepard et al. google/neuroglancer, 2021. URL `https://github.com/google/neuroglancer`.

[6] Nathan Gouwens, Staci Sorensen, Jim Berg, Changkyu Lee, Tim Jarsky, Jonathan Ting, Susan Sunkin, David Feng, Costas Anastassiou, Eliza Barkan, Kris Bickley, Nicole Blesie, Thomas Braun, Krissy Brouner, Agata Budzillo, Shiella Caldejon, Tamara Casper, Dan Castelli, Peter Chong, and Christof Koch. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature Neuroscience*, 22, 2019.

[7] Xiaolong Jiang, Shan Shen, Cathryn R. Cadwell, Philipp Berens, Fabian Sinz, Alexander S. Ecker, Saumil Patel, and Andreas S. Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350 (6264):aac9462, November 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aac9462.

[8] Brian E. Kalmbach, Rebecca D. Hodge, Nikolas L. Jorstad, Scott Owen, Trygve E. Bakken, Rebecca de Frates, Anna Marie Yanny, Rachel Dalley, Lucas T. Graybuck, Tanya L. Daigle, Cristina Radaelli, Matt Mallory, Medea McGraw, Nick Dee, Philip R. Nicovich, C. Dirk Keene, Ryder P. Gwinn, Daniel L Silbergeld, Charles Cobbs, Jeffrey G Ojemann, Andrew L Ko, Anoop P Patel, Richard G. Ellenbogen, Staci A. Sorensen, Kimberly Smith, Hongkui Zeng, Bosiljka Tasic, Christof Koch, Ed S. Lein, and Jonathan T. Ting. Signature morpho-electric, transcriptomic, and dendritic properties of extratelencephalic-projecting human layer 5 neocortical pyramidal neurons. *bioRxiv*, 2020. doi: 10.1101/2020.11.02.365080.

[9] Lida Kanari, Pawel Dlotko, Martina Scolamiero, Ran Levi, Julian C. Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16:3 – 13, 2017.

[10] Lida Kanari, Srikanth Ramaswamy, Ying Shi, Sebastien Morand, Julie Meystre, Rodrigo Perin, Marwan Abdellah, Yun Wang, Kathryn Hess, and Henry Markram. Objective morphological classification of neocortical pyramidal cells. *Cerebral Cortex*, 29(4):1719–1735, 2019.

[11] Euiseok J. Kim, Ashley L. Juavinett, Espoir M. Kyubwa, Matthew W. Jacobs, and Edward M. Callaway. Three types of cortical layer 5 neurons that differ in brain-wide connectivity and function. *Neuron*, 88(6):1253–1267, 2015. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2015.11.002.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015.

[13] Sophie Laturnus, Dmitry Kobak, and Philipp Berens. A Systematic Evaluation of Interneuron Morphology Representations for Cell Type Discrimination. *Neuroinform*, 18(4):591–609, October 2020. ISSN 1559-0089. doi: 10.1007/s12021-020-09461-z.

[14] Henry Markram, Eilif Muller, Srikanth Ramaswamy, Michael Reimann, Marwan Abdellah, Carlos Aguado, Anastasia Ailamaki, Lidia Alonso-Nanclares, Nicolas Antille, Selim Arsever, Atenekeng Kahou Guy Antoine, Thomas K Berger, Ahmet Bilgili, Nenad Buncic, Athanassia Chalimourda, Giuseppe Chindemi, Jean-Denis Courcol, Fabien Delalondre, Vincent Delattre, and Felix Schürmann. Reconstruction and simulation of neocortical microcircuitry. *Cell*, 163:456–492, 2015.

[15] Marcel Oberlaender, Christiaan P. J. de Kock, Randy M. Bruno, Alejandro Ramirez, Hanno S. Meyer, Vincent J. Dercksen, Moritz Helmstaedter, and Bert Sakmann. Cell Type–Specific Three-Dimensional Structure of Thalamocortical Circuits in a Column of Rat Vibrissal Cortex. *Cerebral Cortex*, 22(10):2375–2391, 2012.

[16] James L. O'Leary. Structure of the area striata of the cat. *Journal of Comparative Neurology*, 75(1):131–164, 1941. ISSN 1096-9861. doi: 10.1002/cne.900750107.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12:2825–2830, 2011.

[18] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *bioRxiv*, 2019. doi: 10.1101/731877.

[19] Santiago Ramón y Cajal. *Histologie du système nerveux de l'homme et des vertébrés.* 1911.

[20] Federico Scala, Dmitry Kobak, Shen Shan, Yves Bernaerts, Sophie Laturnus, Cathryn Rene Cadwell, Leonard Hartmanis, Emmanouil Froudarakis, Jesus Ramon Castro, Zheng Huan Tan, et al. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. *Nature communications*, 10(1):4174, 2019.

[21] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn Cadwell, Jesus Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, Shalaka Mulherkar, Zheng Tan, Zizhen Yao, Hongkui Zeng, Rickard Sandberg, Philipp Berens, and Andreas Tolias. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, 598:1–7, 2021.

[22] Casey M Schneider-Mizell, Agnes Bodor, Derrick Brittain, JoAnn Buchanan, Daniel J. Bumbarger, Leila Elabbady, Daniel Kapner, Sam Kinn, Gayathri Mahalingam, Sharmishtaa Seshamani, Shelby Suckow, Marc Takeno, Russel Torres, Wenjing Yin, Sven Dorkenwald, J. Alexander Bae, Manuel A. Castro, Paul G. Fahey, Emmanouil Froudakis, Akhilesh Halageri, Zhen Jia, Chris Jordan, Nico Kemnitz, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Eric Mitchell, Shanka Subhra Mondal, Shang Mu, Barak Nehoran, Stelios Papadopoulos, Saumil Patel, Xaq Pitkow, Sergiy Popovych, William Silversmith, Fabian H. Sinz, Nicholas L. Turner, William Wong, Jingpeng Wu, Szi-chieh Yu, , Jacob Reimer, Andreas S. Tolias, H Sebastian Seung, R Clay Reid, Forrest Collman, and Nuno Maçarico da Costa. Cell-type-specific inhibitory circuitry from a connectomic census of mouse visual cortex. *bioRxiv*, 2023. doi: 10.1101/2023.01.23.525290.

[23] Ruggero Scorcioni, Sridevi Polavaram, and Giorgio A Ascoli. L-measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. *Nature protocols*, 3(5):866–876, 2008.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008.

[25] Simon Weiler, Drago Guggiana Nilo, Tobias Bonhoeffer, Mark Hübener, Tobias Rose, and Volker Scheuss. Orientation and direction tuning align with dendritic morphology and spatial connectivity in mouse visual cortex. *Current Biology*, 32(8):1743–1753.e7, April 2022. ISSN 09609822. doi: 10.1016/j.cub.2022.02.048.

[26] Marissa A. Weis, Laura Pede, Timo Lüddecke, and Alexander S. Ecker. Self-supervised representation learning of neuronal morphologies, 2021.

[27] Ilker O. Yaz and Sébastien Loriot. Triangulated surface mesh segmentation. In *CGAL User and Reference Manual*. CGAL Editorial Board, 5.5.1 edition, 2022. URL `https://doc.cgal.org/5.5.1/Manual/packages.html#PkgSurfaceMeshSegmentation`.

A Appendix

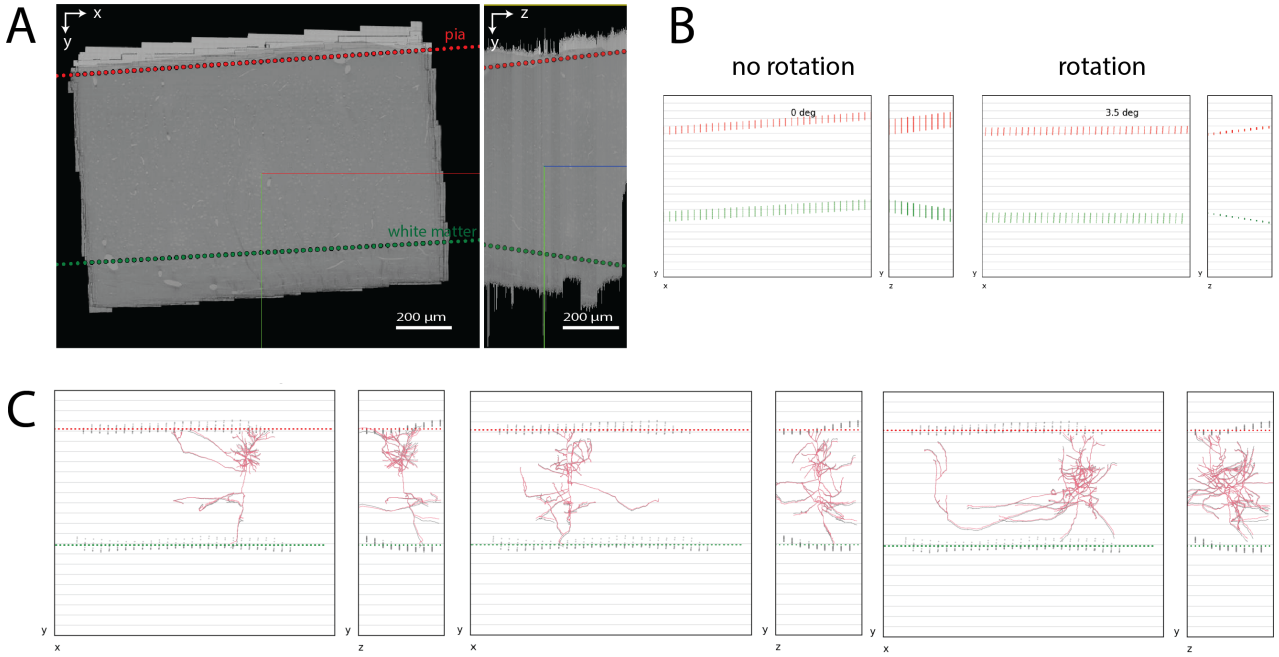

Figure A.1: Volume Pre-processing. A. x-y (left) and y-z (right) 2D cross-sectional views of the EM volume as seen in Neuroglancer. Red scatter points - linear model of pia, Green scatter points - linear model of L6 - white matter boundary. B. Pia and white matter boundary models shown with (right) and without (left) rotating the volume by 3.5 degrees about the z-axis. C. Three example excitatory neuronal skeletons shown from two 2D projections (x-y) and (y-z) after rotation and depth normalization to the mean pia and white matter depths. Red scatter points - pia model after normalization. Green scatter points - white matter boundary after normalization. Gray shadow - pia, white matter and neuronal skeleton after rotation but before normalization.

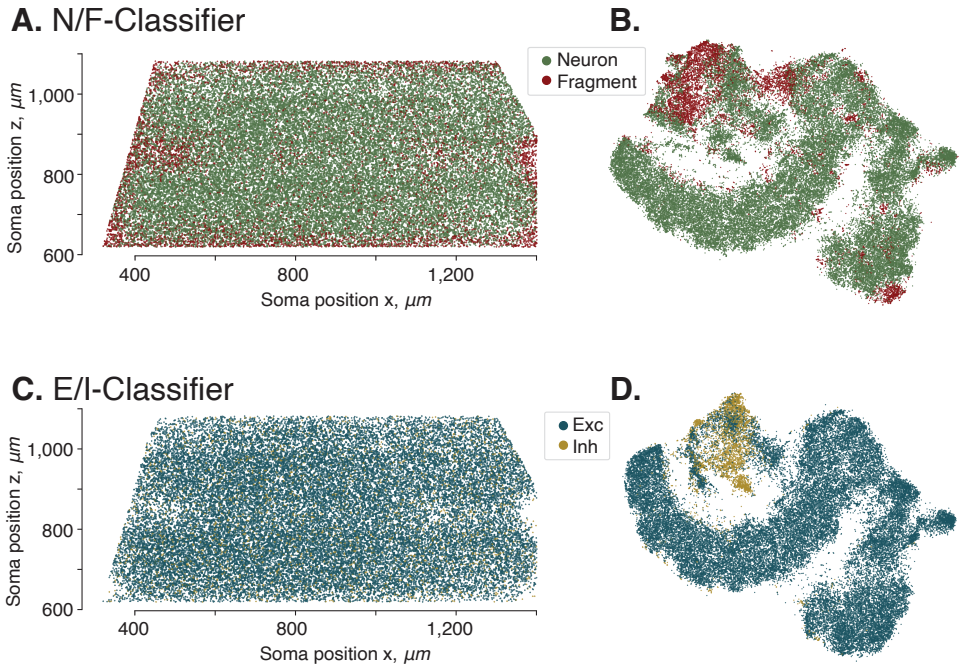

Figure A.2: **A.** Top view of the volume showing the distribution of complete neuron versus fragments predictions by our classifier based on the learned morphological embeddings. Density of fragmented neurons is high at the volume borders, since neurons have a high likelihood of leaving the imaged volume with their dendrites. Additionally, we see a high number of fragmented neurons in areas where we know there have been issues during the imaging process, proving that the classifier works as intended. **B.** t-SNE embedding of neuronal morphologies colored by neuron-fragment predictions. **C.** Top view of the volume showing a uniform distribution of excitatory versus inhibitory neurons across the volume. **D.** t-SNE embedding of neuronal morphologies colored by excitatory-inhibitory predictions.

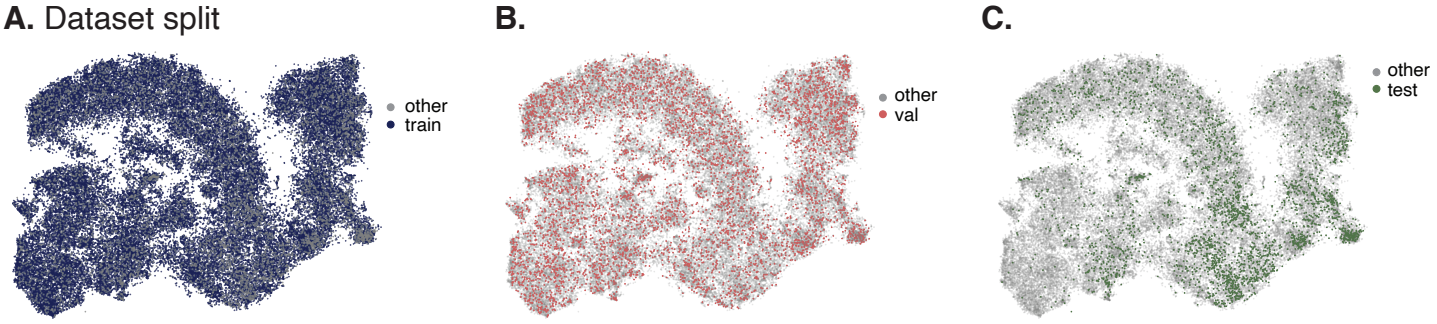

Figure A.3: t-SNE embedding of neuronal morphologies colored by the dataset split: training (**A.**), validation (**B.**) and test set (**C.**) as used for GRAPHDINO training.