bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Quantitative Census of Local Somatic Features in Mouse Visual Cortex

Authors

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, Forrest Collman

Abstract

Mammalian neocortex contains a highly diverse set of cell types. These types have been mapped systematically using a variety of molecular, electrophysiological and morphological approaches. Each modality offers new perspectives on the variation of biological processes underlying cell type specialization. While many morphological surveys focus on branching patterns of individual cells, fewer have been devoted to sub-cellular structure of cells. Electron microscopy (EM) provides dense ultrastructural examination and an unbiased perspective into the subcellular organization of brain cells, including their synaptic connectivity and nanometer scale morphology. Here we present the first systematic survey of the somatic region of nearly 100,000 cortical cells, using quantitative features obtained from EM. This analysis demonstrates a surprising sufficiency of the perisomatic region to recapitulate many known aspects of cortical organization, while also revealing novel relationships. Parameters of cell size, nuclear infolding and somatic synaptic innervation co-vary with distinct patterns across depth and between types. Further, we describe how these subcellular features can be used to create highly accurate predictions of cell-types across large scale EM datasets. More generally, our results suggest that the shifts in cellular physiology and molecular programming seen across cell types accompany profound differences in the fine-scale structure of cells.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

The different cell types of neocortex vary in their complex dendritic and axonal morphologies,[1–7] along with their physiological properties[8,9] and molecular expression profiles.[10–13] Large scale electron microscopy (EM) data with automated reconstructions[14–17] creates both opportunities and challenges for examining the structural variation of cell types across cortex. EM provides an opportunity to study a host of morphological and subcellular detail that has not historically been studied at a large quantitative scale. Further, large scale EM enables the study of connectivity patterns between diverse cell types. However, in very large scale EM datasets with tens to hundreds of thousands of cells, it is impractical to manually classify all the cells in the dataset. Existing methods for automated cell-typing based on morphology often necessitate nearly complete axonal or dendritic reconstructions.[2,8,18] Such reconstructions currently require manual correction to the segmentation, often referred to as proofreading, which is prohibitively time consuming at scale. Furthermore, cells can be classified at varying degrees of refinement, from high level distinctions (neurons vs non-neuronal cells), to highly specific types (Chandelier cells, or a particular transcriptomic type). Depending on the scientific question and perspective of the investigator, different levels of cell type classification are desirable. This highlights the need for a flexible automated solution which can map differing cell classification schemes at varying levels of resolution.

Focusing on the somatic region of cells provides a unique window for analysis that balances these challenges and opportunities. The probability of a segmentation containing an error grows with the size of the segmentation, so analysis that is restricted to a local region is less sensitive to such errors. The soma is a unique local region in that there is only one per cell, and of obvious importance to the biology of the cell. For example, because the soma is the hub for dendritic branches and sits close to the site of action potential initiation, inputs to this region have a great influence on the firing of the cell. Moreover, all the machinery for translating DNA into proteins sits within this cellular compartment, and structural correlates of this machinery are all visible within the EM imagery. Despite this, several basic properties of this compartment have not been measured quantitatively and at a large scale and may vary between types. First, individual cells have widely varying sizes and corresponding demands for gene expression and metabolic load,[10] which might induce shifts in their nuclear and cytosolic size. Second, there are anecdotal reports of differences in nucleus infoldings varying across cortical cell types and modulated by activity.[19,20] Third, in excitatory cells, somatic input has been noted to be dominated by inhibitory sources.[21] Although this is less well characterized for inhibitory subtypes, some have larger fractions of excitatory input.[22,23] In addition, different types of inhibitory axons show preference or avoidance of the somatic compartment.[21,22] Therefore, even a coarse measurement such as the total somatic synaptic density might reflect differences in cell-type specific connectivity[23–25] and thus vary across types. Fourth, smaller scale features of cellular morphology have been described for specific cell types,[23] but little quantitative work has been done in this area across all the full diversity of cortical cells. The local somatic region contains both somatic membrane and proximal dendritic shafts thereby enabling the examination of properties like the size and distribution of spines.[4,26]

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Here we report the first large scale survey of somatic properties of cells as measured with EM across the cortical depth. From the analysis of these properties we describe some novel structural patterns of cortical organization. We integrate our analysis with a systematic anatomical study of a columnar region of cortex, where an anatomical based classification of all cells has been performed. The concept of a cell type is a categorical label which implies that all the cells within that category have stereotyped patterns to all its relevant attributes. There is not broad agreement about what attributes are relevant or what methods define stereotypy, and so cell type definitions can be controversial. We use the term “cell type” in this manuscript to refer to the abstract concept of stereotyped variation across cells. However, when talking about particular categories, we will use the term “class” and “subclass” here to refer to levels of distinctions within this particular categorical scheme. This classification describes cell classes at the level of neurons and non-neuronal classes, and sub-classes within the excitatory and inhibitory populations. We demonstrate that there is enough variation in features amongst cells to build quantitative cellular classification models that predict cell classes and subclasses with high accuracy and enable dataset-wide predictions. This both improves the ability to find many examples of cells of a specific category and allows better measurements of neuronal connectivity with less proofreading (in preparation), solving one of the key problems to exploring large scale EM datasets. These results demonstrate that cortical cells have profound differences in the anatomical structure of their somatic region and provides a novel platform to understand how specialized biological demands affect somatic and nuclear properties.

Results

We analyzed the larger portion of a 1.4mm x 800um x 800um volumetric serial section EM dataset from mouse visual cortex,[14] that contains a dense segmentation of cells along with a nucleus segmentation and large scale synapse detection.[15,27] The quality of the automated segmentation and amount of proofreading varies across the dataset and different structures, but in general the quality of the nucleus segmentation is high for cells away from the edges of the dataset, and the quality of the cellular segmentation is high near the somas of cells. This dataset includes 82,454 high quality nuclear detections fully enclosed within the boundaries of the volume (see methods) and spans cortical Layer 1 through to the white matter (Fig. 1a). For each cell we analyzed the nucleus and somatic region of the 3D segmentation (Fig. 1b) within a 15um radius around the nuclear center point. We also included synapse information within this region (Fig. 1b,c-e). For a subset of 1909 cells in a 100 um columnar region of VISp (referred here to as the column), we had manual labels for cellular classes and neuronal subclasses (Excitatory: Layers 2/3,4,6, and Layer 5 inter-telencenphalic (IT), near-projecting (NP) and extra-telencenphalic (ET) and Inhibitory: Martinotti cell (MC), Basket cell (BC), Bi-polar cell (BPC) and neurogliaform cell (NGC), Nonneurons: astrocyte, oligo-precursor cell (OPC), oligodendrocyte, microglia, pericyte) (Fig. 1b). For all neurons within this column, we also analyzed the nano-scale structure of the postsynaptic compartments, what we are terming a “post synaptic shape”(PSS) (see methods) within 60um of the nucleus center (Fig. 1g) .

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

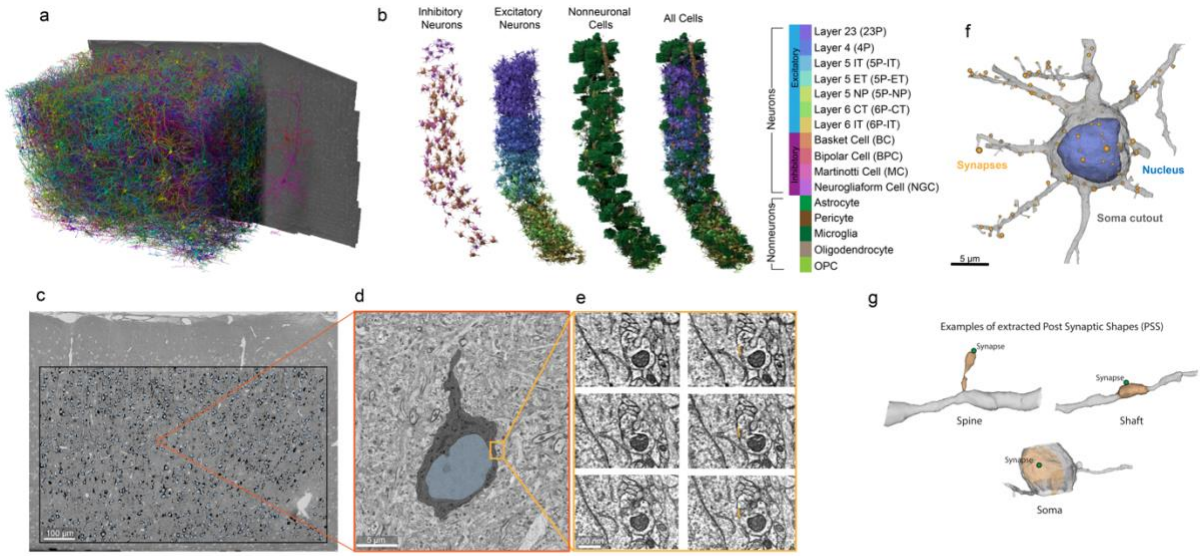


Fig. 1) Systematic evaluation of soma and nucleus features from a large scale EM dataset. a) Rendering of a small fraction of neurons from a large scale dataset covering all layers of cortex and multiple visual areas, with 1207 rendered and then cutaway to reveal the full morphology of 2 selected neurons on the right portion of the dataset. b) rendering of all the neuronal cells from a 100 um column of data for which manual cell type labels were given. Cell class and subclass labels and color schemes used in the paper are shown to the right. c) a 2d view of a single EM section of data with nuclear detections in light blue and somatic regions in gray. d) an enlarged view of the portion of c (red box) showing a single somatic segmentation with the nucleus in blue and the cytoplasm of the cell in gray. e) a series of enlarged 2d views of a single synapse onto the somatic region in d (orange box) showing a single synapse onto that cell. The left column shows the raw EM image, and the left column includes the synapse prediction overlaid in orange f) a 3d mesh rendering of the somatic region of a cell including its cellular membrane (gray), its nucleus segmentation (blue), and the location of synapses as orange spheres (size indicates voxel size of synapse prediction). g) Examples of fine scale morphologies of postsynaptic shapes.

Nucleus Features

We began by systematically analyzing geometric properties of the nucleus, including their volume, surface area, and depth from the pial surface. In addition, we quantified the fraction of nucleus membrane area that was within an infolding (Fig. 2a-c) (see methods). The nucleus has anecdotally been described as having different degrees of infolding in different cell types,[20,28] but a systematic quantification has not been done across cortical types.

Nucleus features of excitatory neurons have a striking laminar organization, wherein the borders between layer ⅔ (L23), layer 4 (L4), layer 5 (L5), and layer 6 (L6) are all clearly demarcated by shifts in the distribution of nucleus volumes. (Fig. 2d). Furthermore, the fraction of membrane inside an infolding also varies widely and systematically depending on depth (Fig. 2b). Layer ⅔ neurons have largely smooth nuclear membranes. There is a clear gradient of infolding within layer 4. All layer 5 excitatory cells have high degrees of infolding, despite the notable diversity of cell types and sizes within that population, which is reflected in the notable

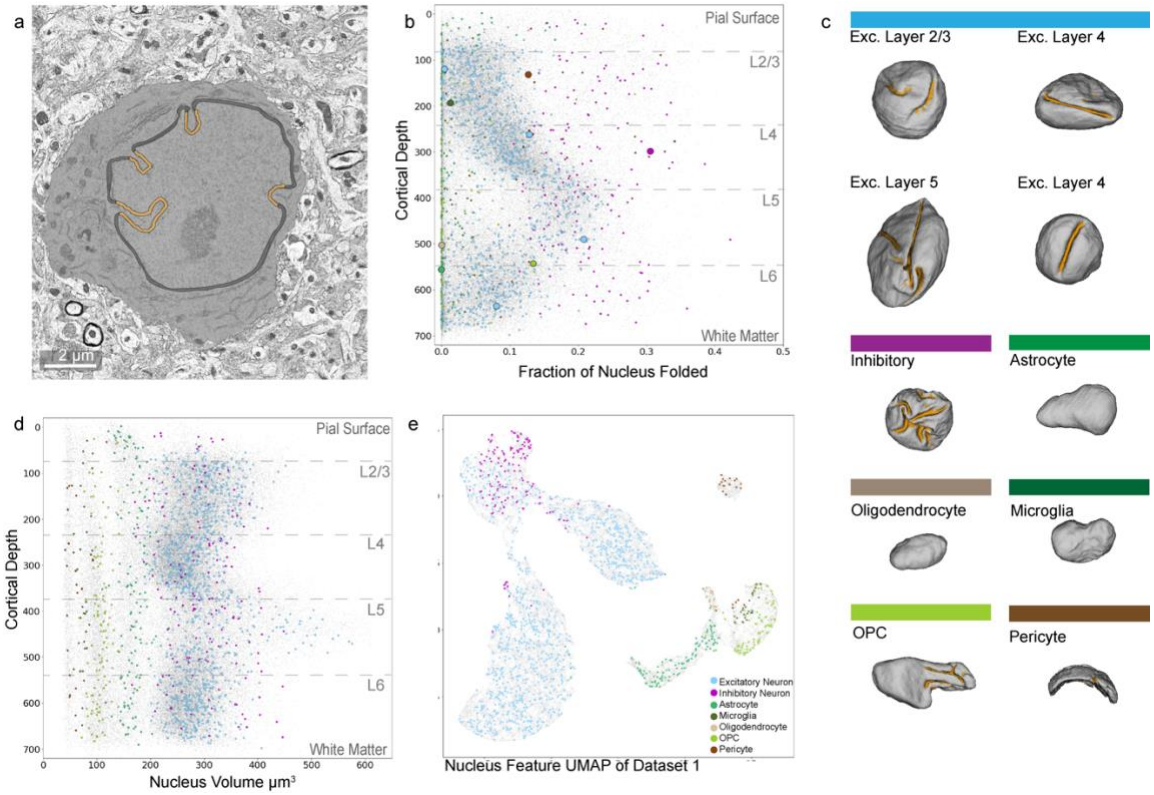
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

increased variation of nucleus volume in that layer.[29] Infolding drops sharply again in layer 6 (Fig. 2c).

Inhibitory cells on the other hand have less striking laminar patterns, but with more overall variation. They have a wider variation of nucleus volumes, which are overlapping with excitatory cells, with the exception of the larger layer 5 neurons. They have 15 to 30% of their membrane within an infolding, regardless of their position within cortex. This makes them quite distinct from excitatory neurons in layer 1, 2/3, 4 and 6 of cortex, but highly overlapping in layer 5 (Fig. 2b).

Non-neuronal cells had generally smaller nuclei compared to neuronal cells, though astrocytes overlap in this distribution with the smallest neurons. Each non-neuronal cell class exhibited a distinct range and consistency in their nucleus volume across the layers of cortex. Non-neuronal cells generally did not have infoldings, though microglia, OPCs and oligodendrocytes had less spherical and convex shapes. Perictyes had the least reliable nucleus segmentations in the dataset, but when segmented well they had the smallest overall volumes with shapes dominated by their close apposition to the vascular walls (Fig. 2c).

Nucleus features were extracted for all cells in the dataset and a 2-dimensional embedding was computed. This low-dimensional space is plotted in Figure 2e with manual class labels in color and all the unlabelled cells in gray (column =1,639, unlabelled=80,815)). The distinct clustering of different cell classes suggests that neuronal and non-neuronal cells can be separated using these metrics alone, and that most non-neuronal classes have distinct nuclear shapes (Fig. 2c).



bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fig. 2) Nuclear shape and infolding features show laminar and cell-type distinctions. a) 2D view of a single EM section of data with the cell body in lightgray, the nuclear membrane in black with the portion of the membrane considered in a fold highlighted in orange. b) Fraction of the nuclear membrane within a fold plotted against distance from the pial surface. Human labeled cell-types are represented in color (n=1,639) and unlabelled examples in light gray (n=80,815). Cortical layer boundaries are noted by the dashed lines. c) 3D mesh renderings of representative nuclei from different neuronal and nonneuronal cell-types with the folded surface area highlighted in orange. d) Nuclear volume in um^3 plotted against distance from the pial surface, colored as in b). e) 2D UMAP embedding of all neuronal and nonneuronal cells inferred from nuclear size features, infolding features, and cortical depth (n=82,454).

Somatic Features

We then calculated similar geometric properties of the somatic region of cells (see methods), including the total volume, surface area, cytosolic volume, the ratio of the nucleus volume to the soma volume, and the average distance from the centroid of the nucleus to the centroid of the soma. We also measured the number and density of synapses detected on the somatic region of the cell. Together these somatic and nucleus features represent a feature space that can be extracted automatically from most cells in the segmentation (75% of nuclei). We filtered out soma regions that represented direct soma to soma merges (n=3,852), which naturally fall out as statistical outliers (see methods, Extended Data figure 1a). Plotting just the metrics of the nucleus volume versus the surface area of the cell within 15 microns of the center of the nucleus, reveals a surprisingly striking separation between the major cell classes found in the brain (Fig. 3a).

Similar to their nucleus features, excitatory neurons showed laminar shifts in their somatic volume and surface area (Extended Data Fig. 5) The excitatory neurons show a consistent synapse density that slightly varies in a linear fashion with depth through the cortical volume, with a notable increase in variation in layer 5 that correlates with the 3 sub-classes found there with ET cells having larger synapse densities, NP cells with low synapse densities and IT cells in between (Fig. 3b, d).

Neurons identified as inhibitory showed less laminar variation in somatic size, however they are on average smaller than excitatory cells in the same cortical layer (Extended Data Fig. 5). On the other hand, inhibitory cells have much larger density of somatic innervation than excitatory cells, but also have a much wider degree of variation, reflecting their rich diversity of sub-classes (Fig. 3b,d).

Based on nucleus volume and soma cutout area alone (Fig. 3a), microglia, oligodendrocytes, OPCs, astrocytes, and pericytes occupy largely distinct portions of this 2-dimensional space. The notably large surface area measurement for astrocytes is explained by the high density of their processes near the soma. Moreover, the high prevalence of segmentation mergers of pericytes with cortical vasculature results in variability in their soma size features as represented by the range in soma surface area in Figure 3a. As expected, all nonneuronal cells had very low soma synapse counts and thus are clearly distinct from neurons across laminar layers (Fig. 3b).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Following the same framework as in Figure 2d, soma features were extracted for all cells in the dataset. A new low dimensional embedding was computed based on somatic features in addition to the previously reported nucleus features. Consistent with the diversity observed in individual features, low dimensional embeddings of the feature space show variation that correlates with the manually labeled cell classes (Fig. 3c). As before, nonneuronal cell classes occupy distinct areas of the feature space. Note that despite the variability in pericytic somatic features, they are still well separated in the low dimensional embedding space, likely due to the distinguishability from nuclear features alone. On the other hand, adding the soma features improves the separation between inhibitory and excitatory cells, with two distinctly inhibitory clusters and finer separation between excitatory neurons in different layers. However, some distinctions between neuronal classes are not as clear. For example, a subset of inhibitory cells overlap with the Layer 5 excitatory cells due to similarities in somatic size and fraction of nuclear infolding. We therefore searched for additional quantitative features that we could extract from the local region around cell bodies to aid in separating these neuronal classes further.

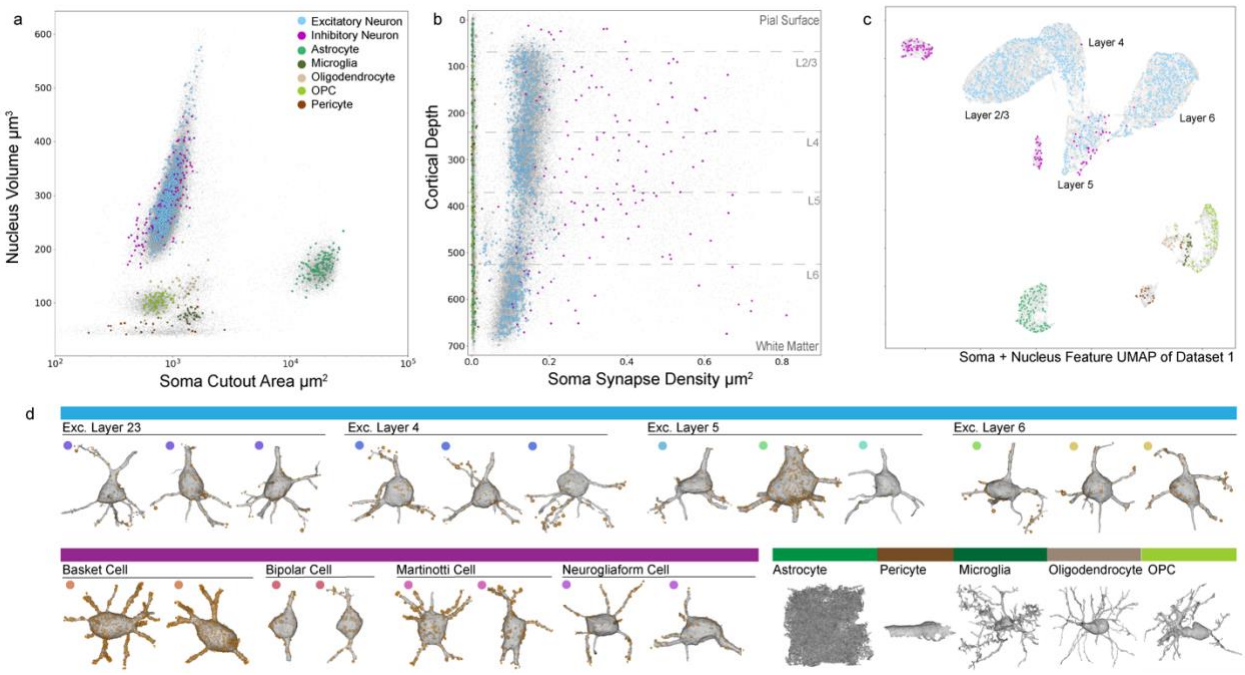


Fig. 3) Somatic shape and synapse features further distinguish neuronal and nonneuronal cell class types. a) Somatic surface area in um^2 (within 15um from the nuclear center) plotted against nuclear volume um^3. b) Somatic synapse density um2 plotted by distance from the pial surface. c) 2D UMAP embedding of all neuronal and nonneuronal cells inferred from somatic features, nuclear features and cortical depth. a-c) Human labeled class types in color and unlabelled cells in gray. d) 3D mesh renderings examples of 15 um somatic cutouts in gray and somatic synapses in orange. Size of individual spheres relative to the size of the post-synaptic density.

Post Synaptic Shape Features

Proximal processes of cells display large variations in fine morphology. For example, some excitatory dendritic branches are covered with a high density of short stubby spines, others have long wispy spines or filopodia. Inhibitory branches vary from being very smooth and uniform, to those that vary in caliper, and some are covered in smaller spine-like protrusions.[23,30] These

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

differences are visible on the dendritic arbor even in close proximity to the soma (Fig. 4). We developed a method to summarize the statistics of the proximal arbor within 60 microns of the nucleus center. This feature combines the patterns of synaptic innervation with the fine morphological structure of the post-synaptic neuron surrounding any given synapse (Fig. 4).[31] Briefly, this method computationally segments the compartment on the post-synaptic side of a synapse, which we refer to as the post-synaptic shape (PSS). This shape typically represents either a portion of the soma, the shaft of a dendrite, or a spiney protrusion, though it can also be onto an axon or axon-initial segment. We then used a PointNet auto-encoder to summarize a latent space of PSSs (Fig. 4a) (see methods). This resulted in a continuous latent space where PSS objects of similar morphological character were closer together (Extended Data Fig. 2). This representation recapitulates well known distinctions between PSSs such as dendritic shafts versus spines versus somatic compartments, but also more fine grained distinctions, such as variations in the size and shape of spines (Extended Data Fig. 3).

We then summarized a cell's distribution of PSSs by binning the PSS space into 30 shape bins and 4 distance bins based upon the distance from the soma (see methods), resulting in a 120 dimensional vector, similar in spirit to a multi-dimensional Sholl analysis.[32] Examining individual example cells of different types suggest that this feature effectively quantifies differences in the fine scale synaptic distribution of different cell types (Fig. 4c). For example, the layer 5 ET and basket cell both contain large numbers of somatic synapses, but the layer 5 ET cell has a much higher density of spines. The Layer 5 NP cell has a lower number of spines, but also the distribution of its spines is less concentrated in the PSS shape bins that correspond to smaller shorter spines, compared to the L5 ET cell. Some inhibitory cells contain counts in the bins associated with spines, consistent with the fact that some inhibitory dendrites do have protrusions. However, these protrusions most often lack the classic spine neck and bulbous head morphology of most spines on excitatory cells. Because they are similar in shape, this analysis groups these inhibitory "thorn" like protrusions with the smaller stubby spines found on excitatory cells.[23] Consistent with that, the bipolar example shown here contains an elevated number of such small protrusions,[26] including in the somatic region, whereas the smoother dendritic shafts of this basket cell example contain virtually no counts in these bins. (Fig. 4c). This distance dependent PSS histogram was calculated for all neurons within the column. We then recomputed a low dimensional embedding by aggregating nucleus, somatic and PSS features (Fig. 4d). Note that the addition of the PSS-based features results in a near complete distinction between inhibitory and excitatory neurons within this space.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

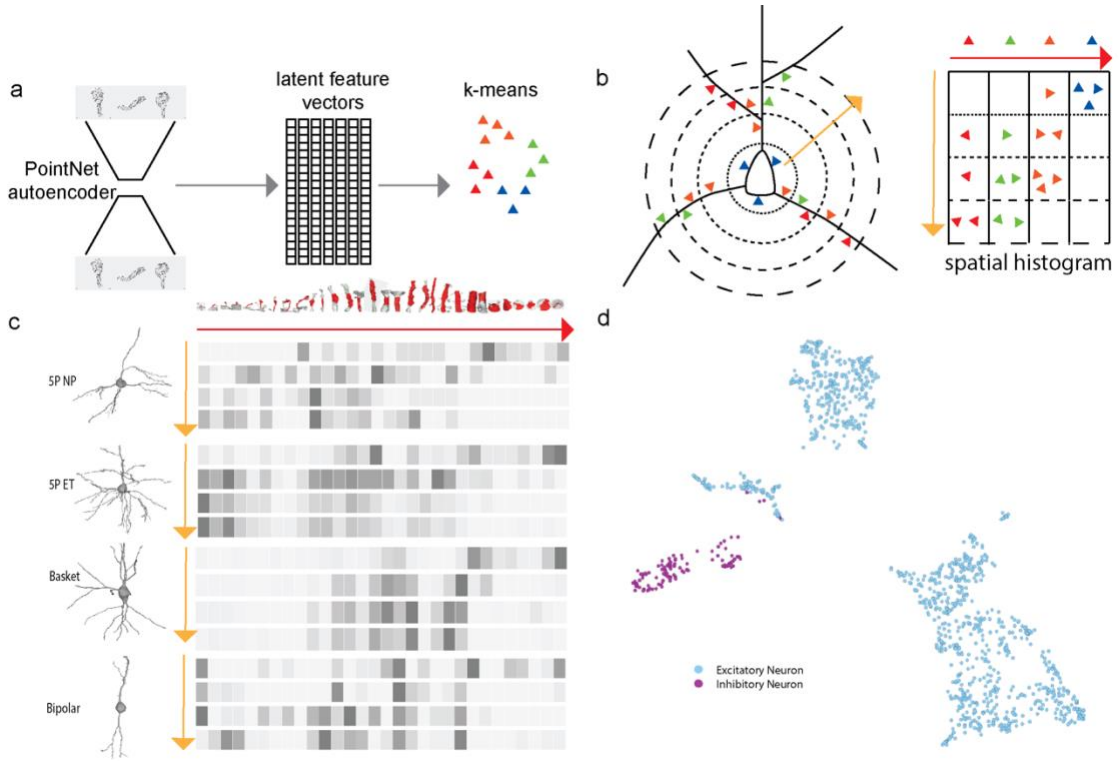


Fig. 4) Post Synaptic Shape (PSS) Features. a) Procedure for building a PSS dictionary model. The set of shapes is used to train a PointNet autoencoder which learns a latent feature vector of a fixed size (1024). This autoencoder is then applied to all shapes in the dictionary to generate a set of latent feature vectors. K-means with K = 30 is applied to this to obtain a set of cluster centers for binning the shapes. b) Schematic showing procedure for creating the PSS histogram. The PSS are binned by shape type (30 bins calculated from a) and distance from the soma (4 bins) from 0 - 60 microns with 15 micron bin sizes. For each synapse, its shape type is determined by finding the closest cluster center in the latent feature space and its distance bin is determined by using the location of the synapse center. The resulting histogram is a 2D histogram shown above with the shapes in the x direction and distances in the y direction. c) Examples of 60 micron cutouts of 2 different types of cells with their spatial histograms shown as heatmaps. The top row shows the shape of the cluster center of each of the 30 clusters. In each heat map, darker boxes indicate higher values. Observe that the pattern of the overall histogram is very different between the two cell types. d) 2D UMAP of all neurons in the column inferred after concatenating nucleus, soma and PSS features. Observe the separation of inhibitory and excitatory neurons.

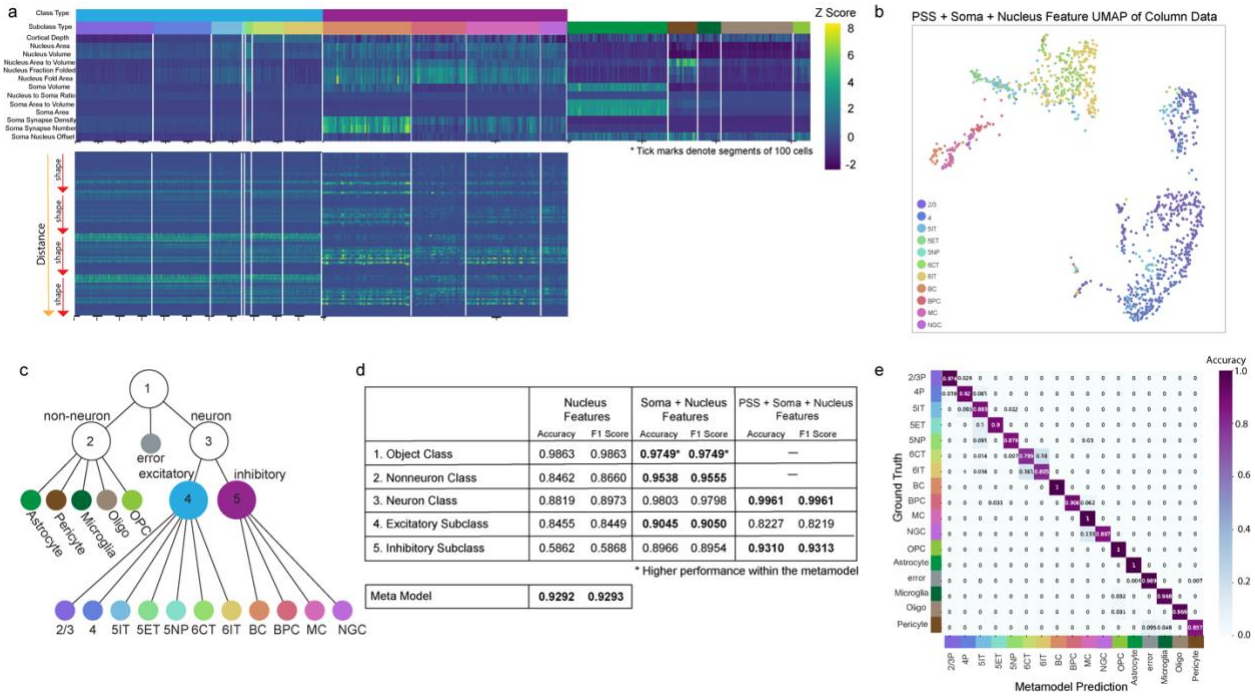
Classifiers

For the cells in the column, we collected three sets of features: nucleus and soma features for all cells, and PSS features for the neurons. The cell classes and sub-classes defined across the column appear to have distinct fingerprints when visualized by the z-scored feature heatmaps (Fig. 5a). A low dimensional embedding of the combined features of neurons colored by sub-class supports that perspective (Fig. 5b). However, in order to quantitatively compare how well different feature spaces separate cells at different levels of distinction, we fit a collection of classification models that could be applied to all the cells in the dataset (see methods). We took a hierarchical approach (Fig. 5c), developing a cascade of classifiers to sort cells at increasingly finer

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

distinctions. At each step (1-5) in the hierarchy we trained a series of classifiers using the different feature sets available for that step (Fig. 5d).

We then integrated the steps into a meta-model, where individual cells are sequentially sorted down the hierarchical tree. We found an optimal combination of classifiers which predicted cell types with a 92% overall accuracy (Fig. 5d) (see methods). This performance is high across different portions of the cell-type space, including between the inhibitory sub-classes whose definitions historically have come from their targeting patterns or dendritic morphology. The largest single confusion between types here was between adjacent layers of similar pyramidal classes, where arguably the strict laminar boundaries found in the manual labels might not be biologically accurate (Fig. 5e). This demonstrates that these features are indeed useful for separating cell-types based on local somatic reconstructions of cortical cells.



	Nucleus Features		Soma + Nucleus Features		PSS + Soma + Nucleus Features	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
1. Object Class	0.9863	0.9863	**0.9749***	**0.9749***	—	
2. Nonneuron Class	0.8462	0.8660	**0.9538**	**0.9555**	—	
3. Neuron Class	0.8819	0.8973	0.9803	0.9798	**0.9961**	**0.9961**
4. Excitatory Subclass	0.8455	0.8449	**0.9045**	**0.9050**	0.8227	0.8219
5. Inhibitory Subclass	0.5862	0.5868	0.8966	0.8954	**0.9310**	**0.9313**

* Higher performance within the metamodel

Meta Model	**0.9292**	**0.9293**

Fig. 5) A combination of nucleus, soma, and PSS features predict neuronal and nonneuronal subclasses for unlabelled cells. ai) depth, nucleus and soma feature matrix depicting z score values for all cells within the manually labeled column. Cells ordered by manual subclass labels, class and subclass colors noted in the bars at the top. Some sub-classes are compressed relative to others in order to better illustrate the variation across classes. Dashed marks along the x axis denote segments of 100 cells (1127 excitatory neurons, 141 inhibitory neurons, 325 nonneurons) aii) pss feature matrix for all neurons within the manually labeled column. Cell ordering matches the matrix in ai. b) 2D UMAP embedding inferred from depth, nucleus, soma, and pss features of all neurons within the column dataset colored by human labeled subclass. c) diagram of hierarchical model framework to predict nonneuronal class and neuronal subclass using a set of 5 models. Initial objects are outputs of a segmentation based nucleus detection model (in preparation) and include neuronal nuclei, nonneuronal nuclei and errors. Cells are subsequently sorted and pushed through the hierarchical framework based on predictions from higher level models. d) Accuracy and F1 scores for all trained models at each level of the hierarchical framework in b. Bolded values denote which model was used in the metamodel due to highest performance. Overall metamodel performance based on the confusion matrix in c is noted in the lower right corner. e) Confusion matrix results for all cells within the column based on subclass predictions from the hierarchical metamodel.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Generalization to other datasets

To test whether the features we extracted were generalizable to other datasets from cortex, the same metrics were collected from a second dataset (see methods). The second dataset covers all layers of cortex, but is only 50um thick and so most cells are severely cut off. The same nucleus, soma and PSS features were extracted for all cells. Cells in the second dataset occupy similar areas when transformed into the low embedding space fit by the features extracted from the first dataset (Fig 6b-d). Similar to the first dataset, as more features are added, the separation between cell class and neuronal subclass increases. Further, the feature to feature correlations between the soma and nucleus features are highly similar between the two datasets (Fig. 6e). Because the PSS feature used a 60um cutout the PSS spatial histograms were often highly truncated for some cells, causing systematic differences between the datasets. We could however, use a meta-model constructed solely with nucleus and somatic features to predict sub-class in the second dataset in a way that aligns with what we expected based upon performance tests in the larger dataset (Fig. 6f, Extended Data Fig. 1b). This suggests these features are generally useful and reliable metrics that can be used to identify cells in experiments that are not large enough to capture morphology beyond the soma.

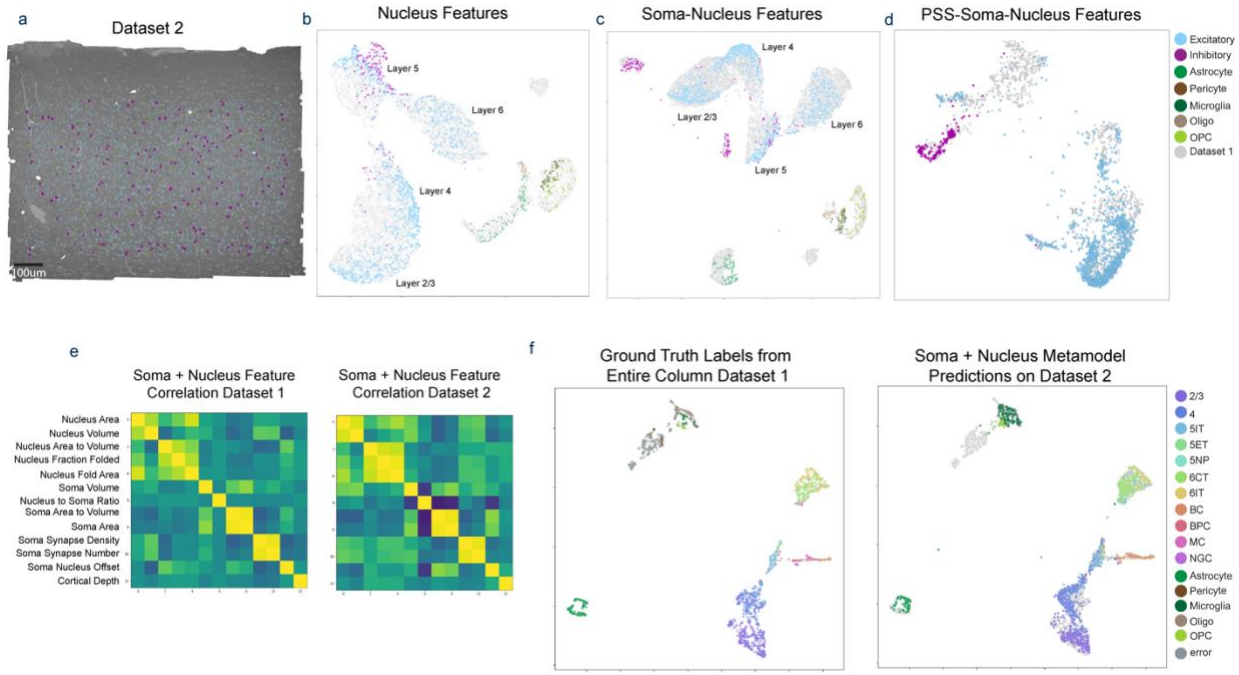


Fig. 6) Application of soma and nucleus features to second dataset. A) An overview of the dataset that covers a mm^2 of cortex but is only 50um thick. Rendering of all the nuclei analyzed within this dataset, colored by manual neuronal and nonneuronal class labels. B) A UMAP of the nucleus features from this dataset, colored by human labeled cell class. The UMAP is fit based on dataset 1, and the points from dataset 1 are included in this plot in gray. C) a UMAP of the nucleus and soma features, again colored by human labeled cell classes and dataset 1 points in gray. D) A UMAP plot of all features, where the neurons are colored by their class. E) The canonical correlation analysis of soma and nucleus features across the two datasets. F) To the left, a 2D UMAP embedding inferred from soma and nucleus features displaying all cells from the column in dataset 1, including errors. Colored by manual subclass labels. On the right, the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

same embedding as before except dataset 1 is in gray. Overlaid is dataset 2 now colored by metamodel subclass predictions.

Discussion

Here we report the first large scale quantitative study of the somatic region of cortical brain cells using electron microscopy. We've shown that different cell-types have systematic differences in their local somatic features that reveal interesting biological patterns and allow cell-classes to be quantitatively distinguished.

For example, nuclear infoldings have been observed in a wide variety of cell types, neuronal and non-neuronal, both in live cells and fixed material.[28,33,34] In dataset 1, the degree of nuclear infoldings varied starkly by excitatory sub-class with different patterns of variation within laminar layers. Specifically, we report the existence of a spatial gradient within excitatory cells in layer 4, which might reflect previously described gradients in gene expression.[10] Alternatively, the overall pattern of nuclear infolding which is shared across layer 5 types, might suggest a common mechanism despite their broad differences in gene expression, electrophysiological properties and connectivity.[8,10] These gradients in excitatory cells could be related to a developmental mechanism, given that cortex develops in a laminar fashion, while interneurons, which don't exhibit laminar shifts in nuclear infolding, migrate in from other parts of the brain. [35]

There is some evidence that in some cell types that nucleus infoldings are distinct calcium microdomains.[27] However, their functional importance has not been clearly established. Infolding of the nuclear membrane should reduce the intranuclear diffusion time to the nuclear membrane and increase the total surface area. A natural hypothesis would thus predict that cells with a higher demand for producing RNA out of the nucleus, should therefore have more infolding. However, our observations do not support this hypothesis because all subclasses of layer 5 excitatory cells (which span a wide range of sizes and overall levels of gene expression) have remarkably consistent levels of infolding. On the other hand, we observe a remarkably consistent scaling of the volume of the nucleus and the volume of the cytoplasm of cells across these subclasses, which supports the view that cells scale their overall size in relation to demands of gene expression.

A mechanism of infolding has been proposed whereby laminar associated domains (LADs) on the heterochromatin attach to the nuclear envelope, and a force is exerted that brings the heterochromatin under tension which causes the membrane to fold.[34] If LAD accessibility and binding to the lamin is driving nuclear infolding, then these data would suggest that measurements of these phenomena should exhibit clear spatial patterns across the cortical depth. Further, genomic analyses should show consistency of LAD accessibility across the layer 5 pyramidal sub-classes, despite large differences in gene expression. Examining the developmental origins of infolding within these varied layer 5 types could produce insight into the mechanisms underlying this phenomena.

Our results reveal some novel properties of distinct cell classes. Near projecting neurons were first described as a set of Layer 5 cells expressing Slc17a8, which did not project to

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

contralateral cortex or any of several sub-cortical structures.[36] They have also been characterized as being enriched in a genetic line, and having a small round somatic shape.[29,36] Since then, transcriptomics and retro-seq experiments has confirmed[4] that they are a rare, but highly conserved subclass of neurons across several species.[37] Morphologically they have been characterized as having sparse long straight basal dendrites and distinct electrophysiological properties.[8] Our description here adds some important facts to their functional profile. They appear to have one of the lowest densities of synaptic input of all cell sub-classes we identified, on both shafts, somas and spines. What synapses they do have on spines are often on uncharacteristically long and spindle-like spines. This suggests that NP cells might be an intriguing target to understand the role of spine shape in cortical computation.[38–40]

Our observation that spine shapes vary systematically across cell types could relate to differences in the average age or nature of spines across those types,[41] as at least VIP dendrites have been shown to have much more motile spines than other interneurons,[26] and some spiny SST+ interneurons have highly multi-synaptic spine like protrusions.[23] We observed some clear but previously unquantified differences in features of different inhibitory sub-classes in cortex. Basket cells are known to have a high rate of connectivity to pyramidal cells and vice versa, on a per cell basis, but our data shows that this correlates with a higher overall density of synapses onto those cells compared to other subtypes (Extended Data Fig 5).[24,42,43] Furthermore, the martinotti subtypes have some notable variation of synaptic density on their soma, and neurogliaform cells have a strikingly lower density of synapses on their somatic region.

It should be noted explicitly that the level of description of cell classes and sub-classes presented here is not meant to reflect a final refined definition of cell types in the mouse visual cortex. In particular, the 100 um column is small enough that it doesn't fully sample the space of inhibitory cells. All the inhibitory cells were not given a label and some known types are missing or only single examples are found. For example, there are no chandelier cells found in that dataset. Exploring the space of soma and nucleus features often co-varies with variation in connectivity patterns of individual cells, suggesting that further sub-divisions are necessary and justified by these data. Furthermore, the precise sub-class definitions of individual excitatory cells is subject to reasonable debate. In particular, classifying excitatory cells across the layer ⅔ and layer 4, and layer 5 and layer 6 borders can vary from rater to rater on a 10-20 microns length scale. Nonetheless, the variation across the major classes and sub-classes shown here demonstrates that the somatic structure of individual neurons is modulated by cell type specialization.

The classifications we have produced here will be immensely useful for researchers mining these data for biologically interesting patterns of connectivity. Because the features used in this study do not require immense proofreading, the predictions are accurate across most of the dataset. This means the cell type connectivity of individual neurons or axonal fragments can be directly queried for targets which contain a single soma in the dataset. In the present release of the dataset, this means over 50% of a typical cell's outputs can be classified.[38] Another use case for these predictions is to aggregate common inputs to a subclass of cells to find axon fragments with strong specificity for that subclass. Furthermore, these features can be extracted

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

from datasets where full reconstruction is not possible, as we demonstrated with Dataset 2. These use cases demonstrate how an accurate data set wide cell classification transforms such datasets into a research resource for exploring the cell type specificity of connectivity. This success reinforces a more general phenomenon, that not only are different cell types expressing different proteins, but that these changes leave a unique imprint on cells that can be measured in highly localized structures with electron microscopy.

Methods

Dataset 1

Dataset 1 consists of a 1.4mm x 800um x 800um volumetric serial section EM dataset from mouse visual cortex of a male P87 mouse. This covers all layers of cortex and spanning primary visual cortex and two higher visual areas. The dataset has been described in detail elsewhere.[35] Briefly, two photon imaging was performed on the mouse, which was subsequently prepared for electron microscopy. The specimen was then sectioned and imaged using transmission electron microscopy.[36] The images were then stitched, aligned,[7] and processed through a deep learning segmentation algorithm,[7] followed by manual proofreading (manuscript in preparation). For dataset 1, all cells within a 100um region were labeled by human experts as a nonneuron, neuron or error with class and subclass labels where applicable. This is described in detail elsewhere (manuscript in preparation). Cells where the nucleus was surrounded by error ID 0 were excluded to remove edge volume edge effects. Due to high levels of proofreading in the column, there were very few errors thus the training set was augmented with manually labeled errors from the entire dataset.

Dataset 2

The second dataset covers a millimeter square cross-sectional area, and 40 microns of depth within primary visual cortex of a P49 male mouse. This dataset is described in detail elsewhere (manuscript in preparation). The largest available segmentation spans Layer 2/3 of cortex through to Layer 6. After applying the nuclear detection model (manuscript in preparation) and filtering out all nuclear objects below 25 um^3 and cells that were cut off by the volume border, 82,148 cells were used for the analysis. Class type of each cell was labeled manually and used as ground truth. Due to thinness of the volume, much of the distal cell morphologies were cut off and thus subclass type labeling was not possible. Nuclear and somatic meshes were cleaned using the same heuristic procedure outlined above.

Generating Nucleus Features

We analyzed nuclei using the results of a deep neural network segmentation (manuscript in preparation), extracted the mesh using marching cubes and obtained the largest component of the detected mesh.

We further filtered the detections from this algorithm in that any detected objects less than 25 um^3 were filtered out as errors. Nuclear features were then extracted on the remaining meshes (n=129,769). These features included, nucleus volume, nucleus area, the area to volume ratio,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nucleus surface area within an infolding, the fraction of the total surface area within an infolding, and cortical depth (measured as the distance from the pial surface). Nucleus fold features were extracted by creating a shrink wrapped[37] mesh for each nucleus mesh. We then calculated the distance of each vertex on the nucleus mesh from the shrink-wrapped mesh and any vertex further than 150 nm was considered within an infolding.

Generating Somatic Features

For each nucleus detection the somatic compartment was identified as the ID in the segmentation which surrounded >80% of the nucleus. Somatic segmentations went through a heuristic cleaning procedure to remove missing slices of data and incorrectly merged fragments. Since each soma was matched to its corresponding nucleus, 15 microns surrounding the nucleus center of mass was cut out from the dense segmentation and converted into a binary mask. Binary dilation by 5 voxels in 3d was performed, followed by filling of all holes, and then binary erosion of 3 voxels. The resulting binary mask was meshed using marching cubes and connected component analysis was run on the result. The largest connected component mesh was retained, and any disconnected components were dropped. Somatic features were extracted for all 129,769 nuclear detections within the dataset. These somatic features included soma area, soma volume, the area to volume ratio, the number of synapses on the somatic cutout, and the soma synapse density. Using both the somatic and nucleus meshes, we calculated the ratio between the nucleus volume and soma volume and the offset between the two, measured as the euclidean distance between nuclear center of mass and soma center of mass.

Filtering procedure: After identifying the segment IDs within a 15um bounding box around each nucleus, if over 20% of these IDs corresponded to error ID 0, they were filtered out. The majority of these error cases were cells close to the volume border or areas in the volume with higher segmentation errors such as those near blood vessels. Somatic meshes that were merged with other somas in the dataset were also filtered out. These cells were statistical outliers that clustered together in the low embedding feature space based on soma and nucleus features (Supplemental Fig. 1b) and thus were removed based on their location within that space. Finally, cells that were predicted as errors based on the soma and nucleus feature object model were also removed from analysis. This resulted in a final set of 82,454 cells, neuronal and nonneuronal.

Generating PSS Features

Around each synapse, we extracted a 3500nm region to obtain the synapse region mesh. This mesh was then segmented using the CGAL surface segmentation algorithm[38] which splits regions based on differences in thickness. We adapted our previously developed method[24] to identify the PSS region by using a local skeleton calculated from the synapse region mesh, rather than a precomputed whole cell mesh. This allowed us to adapt this method for cells in the dataset without the need for proofreading.

Given a cell for which all PSS have been extracted within a 60 micron radius from the nucleus center, the objective was to build a descriptor that encapsulates the various properties of the PSS. In particular, we aim to capture two of these properties: the type of shape of the PSS and the distance of the PSS from the soma. For the shape, a dictionary of all shape types is built

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

using the dictionary dataset from.[24] These shapes were rotationally normalized and used to train a pointnet autoencoder[39] to learn a latent representation of size 1024. The high dimensional latent space spanning all these shapes is a continuous space (Extended Data Fig 3) which was used to generate a Bag of Words model[40] for the shapes. For this, we performed K-means clustering with K=30 to estimate cluster centers. The top row of Fig. 4c shows the shape in the dictionary that is closest to each of these cluster centers. For distance binning, we split the 60 micron radius around the nucleus center into four 15 micron radial bins (Figure 4b). All PSS were then binned according to their shape and distance properties.

Hierarchical Framework

We defined an object as the segmentation associated with a predicted nucleus (manuscript in preparation) from which nucleus, soma, and PSS features could be extracted. A hierarchical framework was designed to predict the cell type of any such object (Fig. 5c). From Dataset 1, there were 84,539 objects. The first level in the hierarchy predicts whether the object was a neuron, nonneuron, or an error. All objects predicted as errors were excluded from all subsequent analyses except for metamodel prediction evaluation. Nonneuronal cells were then classified as one of the following: astrocyte, microglia, oligodendrocyte, oligodendrocyte precursor cells (OPC), or pericyte. For neurons, cells were predicted as either excitatory or inhibitory followed by a separate subclass classifier for each class type. Excitatory subclasses: Layer 2/3 pyramidal, Layer 4 pyramidal, Layer 5 IT, Layer 5 ET, Layer 5 near projecting (NP) pyramidal, Layer 6 IT, Layer 6 CT pyramidal. Inhibitory subclasses: Basket cells, Bipolar cells, Martinotti cells, and Neurogliaform cells.

Soma and nucleus features were extracted from the 3D mesh of all objects and PSS features were extracted from all neurons within the column. For each level of the hierarchy, multiple classifiers were trained using either nucleus and depth features only, nucleus depth and soma features, or nucleus soma depth and PSS features. Within each level, performance of the various classifiers was evaluated based on accuracy and F1 score (a measure for precision and recall). The metamodel was defined as the sequential combination of the best performing classifiers at each level.

Individual Classifiers

For each classifier, model type was chosen using a randomized grid search for the following models: Support Vector Machine SVM with a linear kernel, SVM with a radial basis function kernel, Nearest Neighbors, Random Forest Classifier, Decision Tree and Neural Network. For each type, 50 models were trained with varying parameters and the top performing model was chosen. This model was then further optimized using 5-fold cross validation with 20% held out for validation. Training, and test examples were held consistent across models for direct performance comparison within each level.

The top level of the hierarchy (the object model), distinguished neurons from non-neurons as well as erroneous detections. Here, the nucleus features alone were sufficient to separate neuronal from non-neuronal nuclei as well as erroneous detections with an F1 score of 98.7%. Adding somatic features to this model did not significantly change performance. For the non-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

neuronal class model,, somatic features did notably improve the accuracy of distinguishing between non-neuronal types from 84% to 95%. The neuron class model simply distinguished excitatory from inhibitory neurons. Here, again the soma features improved the classification accuracy over the nucleus only features from 88% to 98%. Consistent with the qualitative impression given by the low dimensional projections, the principle confusion for the soma-nucleus models exists between the inhibitory and excitatory cells, particularly in Layer 5 (Extended Data Fig. 4). Adding in the PSS features raised performance to 99.6%. The excitatory subclass model distinguished between 7 different laminar specific pyramidal types. Here, the soma and nucleus features performed the best, with 90% accuracy. The inhibitory subclass model distinguished between the 4 sub-classes of inhibitory cells found in the columnar region. Performance improves significantly as you add in the somatic and PSS features, topping out with an F1 score of 93.1%.

Metamodel

We tested the overall performance for predicting the leaves of the tree using different combinations of the individual classifiers. We found that the best performing meta-model used the best performing classifier at each step (bolded Fig. 5d), with the exception of one, the object class step. This was because the classifier trained on nucleus features alone had a larger number of confusions between neurons and non-neurons than the classifier trained with soma and nucleus features, leading to lower overall performance (accuracy & F1) in the meta-model.

Contributions

Analyzed Data: LE, SS, FC, CSM
Developed Nucleus Model: SM, GM, LE, FC
Cell Typing: AB, NDC, CSM
Paper Writing: LE, SS, FC
Dataset + segmentation generation: collaboration Proofreading: collaboration
Analysis infrastructure: collaboration

References

1. Oberlaender, M. *et al.* Cell type-specific three-dimensional structure of thalamocortical circuits in a column of rat vibrissal cortex. *Cereb. Cortex N. Y. N 1991* **22**, 2375–2391 (2012).
2. Kanari, L. *et al.* Objective Morphological Classification of Neocortical Pyramidal Cells. *Cereb. Cortex* **29**, 1719–1735 (2019).
3. Delgado-Escueta, A. V. The Fine Structure of the Nervous System: Neurons and Their Supporting Cells. Alan Peters , Sanford L. Palay , Henry de F. Webster. *Q. Rev. Biol.* **67**, 80–80 (1992).
4. Wang, Y. *et al.* Anatomical, physiological and molecular properties of Martinotti cells in the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

somatosensory cortex of the juvenile rat. *J. Physiol.* **561**, 65–90 (2004).

5. Martin, K. A. & Whitteridge, D. Form, function and intracortical projections of spiny neurones in the striate visual cortex of the cat. *J. Physiol.* **353**, 463–504 (1984).
6. Gilbert, C. D. & Wiesel, T. N. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature* **280**, 120–125 (1979).
7. Markram, H. *et al.* Reconstruction and Simulation of Neocortical Microcircuitry. *Cell* **163**, 456–492 (2015).
8. Gouwens, N. W. *et al.* Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat. Neurosci.* **22**, 1182–1195 (2019).
9. Scala, F. *et al.* Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature* **598**, 144–150 (2021).
10. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
11. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
12. Jones, E. G. Varieties and distribution of non-pyramidal cells in the somatic sensory cortex of the squirrel monkey. *J. Comp. Neurol.* **160**, 205–267 (1975).
13. Ascoli, G. A. *et al.* Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nat. Rev. Neurosci.* **9**, 557–568 (2008).
14. Consortium, Mic. *et al.* Functional connectomics spanning multiple areas of mouse visual cortex. 2021.07.28.454025 Preprint at https://doi.org/10.1101/2021.07.28.454025 (2021).
15. Macrina, T. *et al.* Petascale neural circuit reconstruction: automated methods. 2021.08.04.455162 Preprint at https://doi.org/10.1101/2021.08.04.455162 (2021).
16. Shapson-Coe, A. *et al.* A connectomic study of a petascale fragment of human cerebral cortex. 2021.05.29.446289 Preprint at https://doi.org/10.1101/2021.05.29.446289 (2021).
17. Motta, A. *et al.* Dense connectomic reconstruction in layer 4 of the somatosensory cortex.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Science **366**, eaay3134 (2019).

18. Scorcioni, R., Polavaram, S. & Ascoli, G. A. L-Measure: a web-accessible tool for the analysis, comparison and search of digital reconstructions of neuronal morphologies. *Nat. Protoc.* **3**, 866–876 (2008).
19. Wittmann, M. *et al.* Synaptic Activity Induces Dramatic Changes in the Geometry of the Cell Nucleus: Interplay between Nuclear Structure, Histone H3 Phosphorylation, and Nuclear Calcium Signaling. *J. Neurosci.* **29**, 14687–14700 (2009).
20. Peters, A. & Kara, D. A. The neuronal composition of area 17 of rat visual cortex. I. The pyramidal cells. *J. Comp. Neurol.* **234**, 218–241 (1985).
21. Villa, K. L. & Nedivi, E. Excitatory and Inhibitory Synaptic Placement and Functional Implications. in (2016). doi:10.1007/978-4-431-56050-0_18.
22. Davis, T. L. & Sterling, P. Microcircuitry of cat visual cortex: classification of neurons in layer IV of area 17, and identification of the patterns of lateral geniculate input. *J. Comp. Neurol.* **188**, 599–627 (1979).
23. Hwang, Y.-S. *et al.* 3D Ultrastructure of Synaptic Inputs to Distinct GABAergic Neurons in the Mouse Primary Visual Cortex. *Cereb. Cortex* **31**, 2610–2624 (2021).
24. Campagnola, L. *et al.* Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* **375**, eabj5861 (2022).
25. Hage, T. A. *et al.* Synaptic connectivity to L2/3 of primary visual cortex measured by two-photon optogenetic stimulation. *eLife* **11**, e71103 (2022).
26. Georgiou, C. *et al.* A subpopulation of cortical VIP-expressing interneurons with highly dynamic spines. *Commun. Biol.* **5**, 1–15 (2022).
27. Turner, N. L. *et al.* Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity. *Cell* **185**, 1082-1100.e24 (2022).
28. Jorgens, D. M. *et al.* Deep nuclear invaginations are linked to cytoskeletal filaments – integrated bioimaging of epithelial cells in 3D culture. *J. Cell Sci.* **130**, 177–189 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

29. Kim, E. J., Juavinett, A. L., Kyubwa, E. M., Jacobs, M. W. & Callaway, E. M. Three Types of Cortical Layer 5 Neurons That Differ in Brain-wide Connectivity and Function. *Neuron* **88**, 1253–1267 (2015).

30. Pchitskaya, E. & Bezprozvanny, I. Dendritic Spines Shape Analysis—Classification or Clusterization? Perspective. *Front. Synaptic Neurosci.* **12**, (2020).

31. Seshamani, S. *et al. Automated Neuron Shape Analysis from Electron Microscopy*. http://arxiv.org/abs/2006.00100 (2020) doi:10.48550/arXiv.2006.00100.

32. Sholl, D. A. Dendritic organization in the neurons of the visual and motor cortices of the cat. *J. Anat.* **87**, 387-406.1 (1953).

33. Shimojima, M. *et al.* Emerin plays a crucial role in nuclear invagination and in the nuclear calcium transient. *Sci. Rep.* **7**, 44312 (2017).

34. Schoen, I., Aires, L., Ries, J. & Vogel, V. Nanoscale invaginations of the nuclear envelope: Shedding new light on wormholes with elusive function. *Nucleus* **8**, 506–514 (2017).

35. Kim, J.-Y. & Paredes, M. F. Implications of Extended Inhibitory Neuron Development. *Int. J. Mol. Sci.* **22**, 5113 (2021).

36. Sorensen, S. A. *et al.* Correlated Gene Expression and Target Specificity Demonstrate Excitatory Projection Neuron Diversity. *Cereb. Cortex* **25**, 433–449 (2015).

37. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

38. Harnett, M. T., Makara, J. K., Spruston, N., Kath, W. L. & Magee, J. C. Synaptic amplification by dendritic spines enhances input cooperativity. *Nature* **491**, 599–602 (2012).

39. Alvarez, V. A. & Sabatini, B. L. Anatomical and Physiological Plasticity of Dendritic Spines. *Annu. Rev. Neurosci.* **30**, 79–97 (2007).

40. Cornejo, V. H., Ofer, N. & Yuste, R. Voltage compartmentalization in dendritic spines in vivo. *Science* **375**, 82–86 (2022).

41. Knott, G. W., Holtmaat, A., Wilbrecht, L., Welker, E. & Svoboda, K. Spine growth precedes

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

synapse formation in the adult neocortex in vivo. *Nat. Neurosci.* **9**, 1117–1124 (2006).

42. Jiang, X. *et al.* Principles of connectivity among morphologically defined cell types in adult neocortex. *Science* **350**, aac9462 (2015).

43. Hioki, H. *et al.* Cell Type-Specific Inhibitory Inputs to Dendritic and Somatic Compartments of Parvalbumin-Expressing Neocortical Interneuron. *J. Neurosci.* **33**, 544–555 (2013).

44. Consortium, Mic. *et al.* Functional connectomics spanning multiple areas of mouse visual cortex. 2021.07.28.454025 Preprint at https://doi.org/10.1101/2021.07.28.454025 (2021).

45. Yin, W. *et al.* A petascale automated imaging pipeline for mapping neuronal circuits with high-throughput transmission electron microscopy. *Nat. Commun.* **11**, 4949 (2020).

46. Kobbelt, L. P., Vorsatz, J. & Labsik, U. and. A Shrink Wrapping Approach to Remeshing Polygonal Surfaces. *Comput. Graph. Forum* **18**, 119–130 (1999).

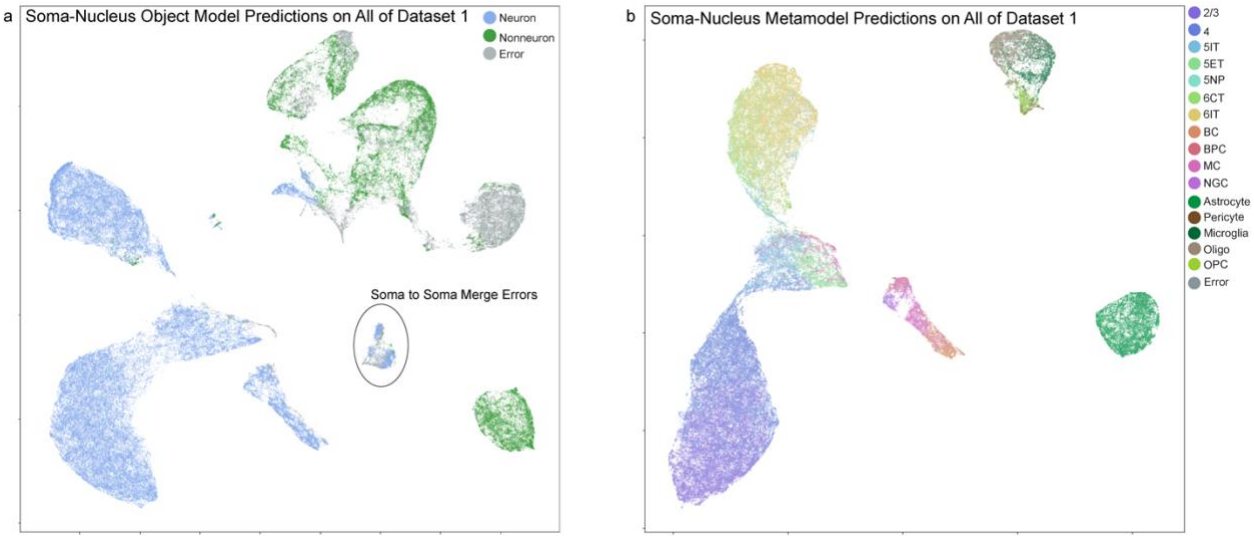
47. Ilker O. Yaz and Sebastien Loriot. CGAL 5.4.1 - Triangulated Surface Mesh Segmentation: User Manual. https://doc.cgal.org/latest/Surface_mesh_segmentation/index.html.

48. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Preprint at https://doi.org/10.48550/arXiv.1612.00593 (2017).

49. Fei-Fei, L. & Perona, P. A Bayesian hierarchical model for learning natural scene categories. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* vol. 2 524–531 vol. 2 (2005).

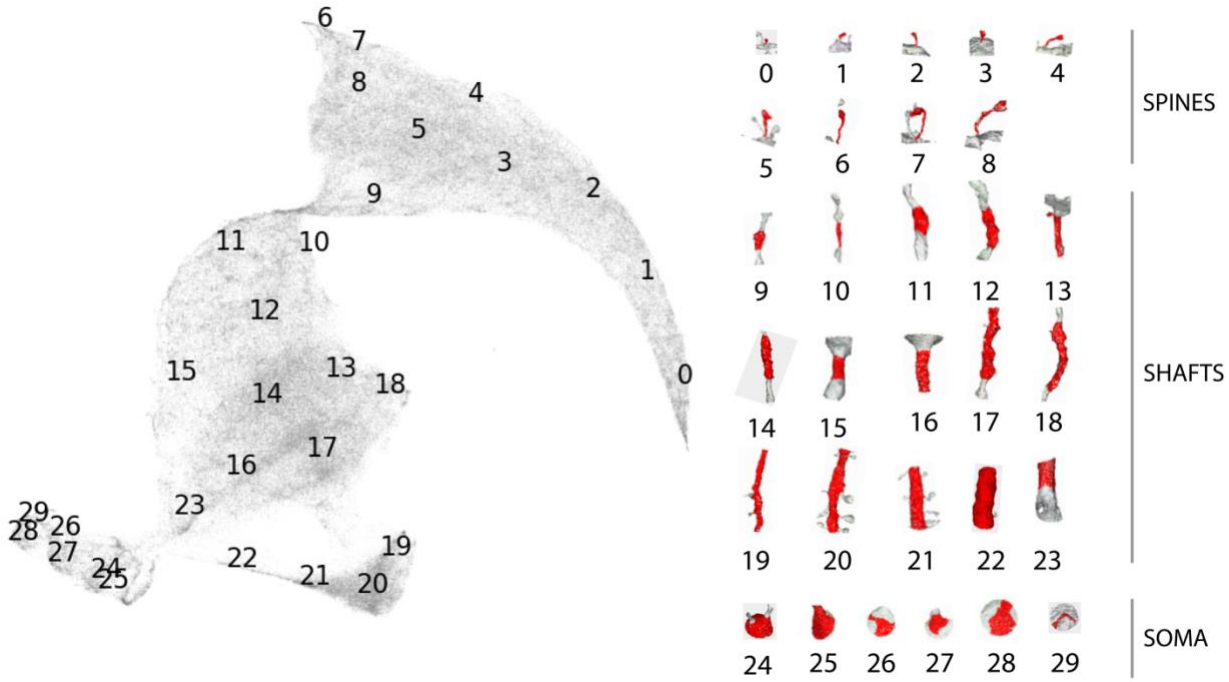
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Extended Data



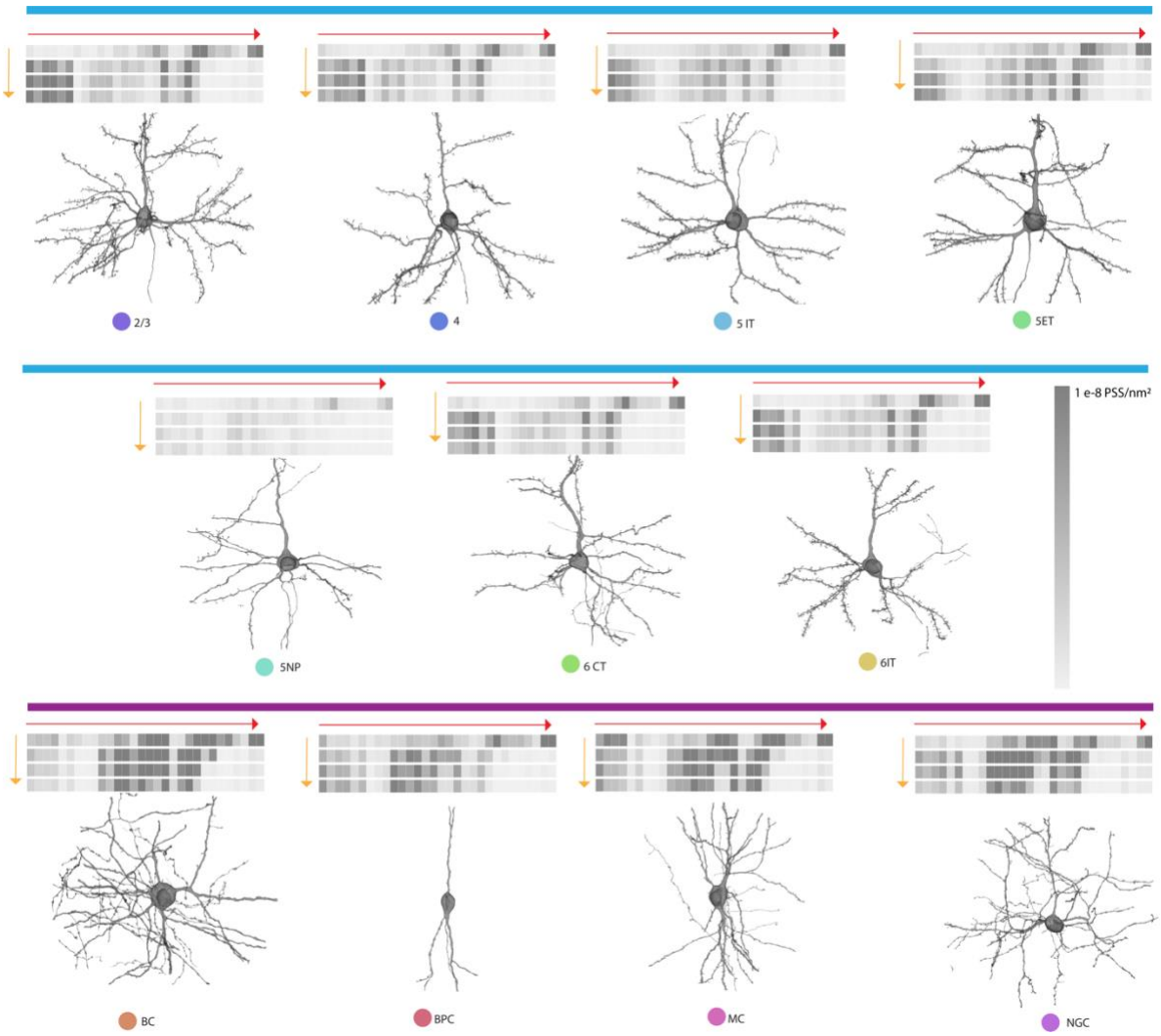
Extended Data Fig. 1) 2D UMAP embedding of soma and nucleus features across the dataset 1. a) UMAP embedding of all the cells in the dataset, including errors (n=129,769). Each point represents an individual cell, color represents the object class prediction from the soma and nucleus based model (Fig. 5), blue represents neuronal predictions, green nonneuronal and gray errors. Note that soma to soma segmentation merge errors are quantitatively distinguishable from other neurons and errors and thus filtered out from further analyses based on these metrics. b) UMAP embedding inferred from soma nucleus features from neuronal and nonneuronal predictions only (n=82,454 cells). Each point represents an individual cell and color represents subclass predictions based on a hierarchical model where each classifier is trained on soma and nucleus features only.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



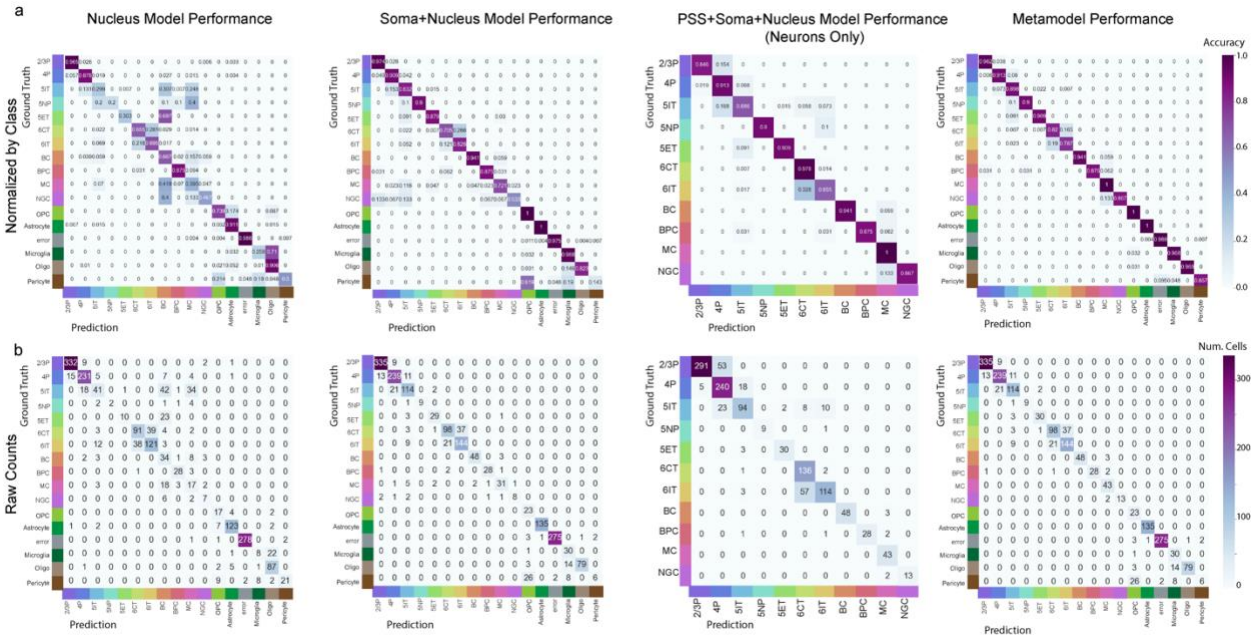
Extended Data Fig. 2) 2D UMAP embedding of all shapes in the PSS Dictionary. The numbers indicate the bin centers mapped in this 2D space and the corresponding PSS meshes on the right show the shape associated with each bin center. Bins 1-8 range in spine shapes, Bins 9-23 are shaft shapes and Bins 24-29 are soma shapes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

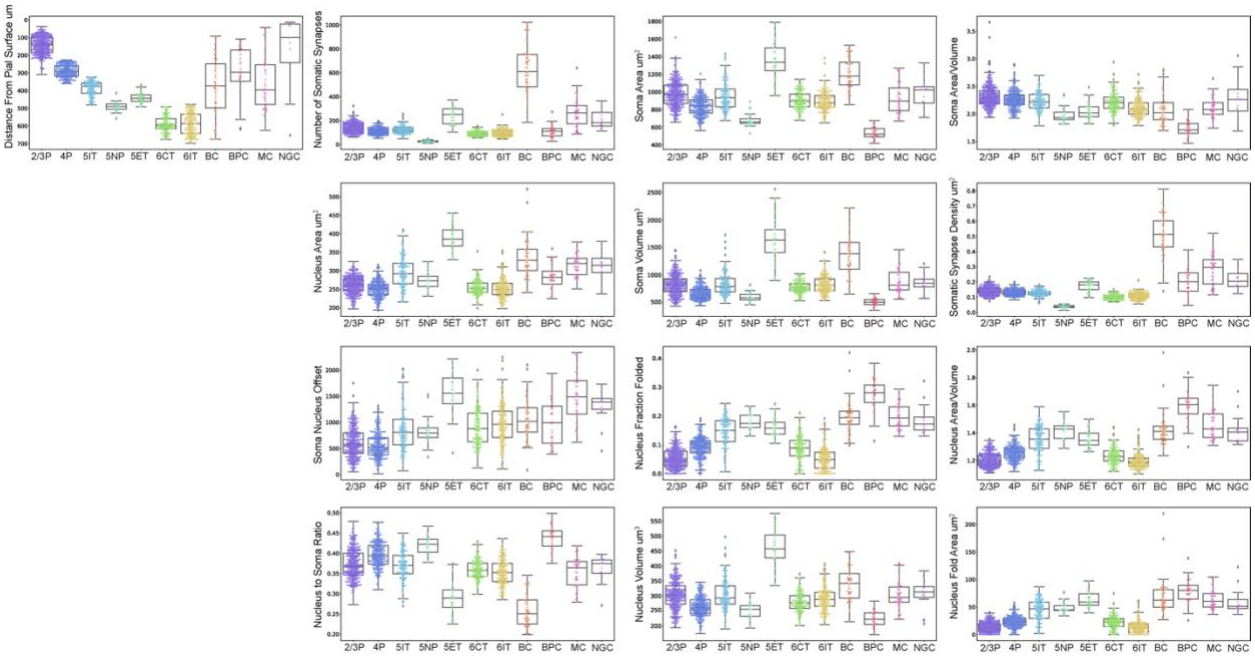


Extended Data Fig. 3) Average PSS histogram visualization for all cells within each subclass category in column. Note that these are different from the histograms in Figure 4 since those show the histograms for specific cells. Below each histogram is a mesh extracted within a 60 micron radius of the center of the nucleus.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.499976; this version posted July 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Extended Data. Fig. 4) Performance of hierarchical classifiers as constrained by various feature sets. a) Confusion matrices for model performance across the human labeled column for classifiers trained on nucleus only features, soma and nucleus features, pss soma and nucleus features and the metamodel (as described in Fig. 5). Rows represent ground truth labels and columns represent predictions. Performance noted as the normalized accuracy across different subclasses. b) The same as a, however performance is noted as the number of cells correct irrespective of subclass size.



Extended Data Fig. 5) Neuronal subclass distribution of individual soma and nucleus features. Distribution and variation of individual features plotted by neuronal subclass labels from the human labeled column dataset. Mean and variance each subclass represented by the boxplots while individual cells are noted in the overlaid swarm plots. Color denotes human assigned subclass labels.