CAJAL: A general framework for the combined morphometric, transcriptomic, and physiological analysis of cells using metric geometry

Kiya W. Govek[1,2], Jake Crawford[3], Artur B. Saturnino[4], Kristi Zoga[1], Michael P. Hart[1], and Pablo G. Camara[1,2,5,#]

[1] Department of Genetics, [2] Institute for Biomedical Informatics, [3] Department of Systems Pharmacology and Translational Therapeutics, and [5] Center for Artificial Intelligence and Data Science for Integrated Diagnosis, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104.

[4] Department of Mathematics, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104

[#] Correspondence to: pcamara@pennmedicine.upenn.edu

Abstract

Cell morphology reflects the progression of high-level cellular processes, such as neuronal diversification, cell migration, or immune cell activation, and is one of the most described cellular phenotypes in biology. Multimodal techniques like Patch-seq enable simultaneously profiling the morphological, transcriptomic, and physiological characteristics of individual cells. However, computational methods that can summarize the great diversity of complex cell shapes found in tissues and infer associations with other single-cell data remain scarce. Here we report a computational framework, named *CAJAL*, for the morphometric and multi-modal analysis of cells. *CAJAL* uses concepts from metric geometry to accurately build, visualize, and integrate cell morphology summary spaces and establish associations with molecular and physiological data of individual cells. We demonstrate the utility of *CAJAL* by applying it to published Patch-seq, patch-clamp, serial electron, and two-photon microscopy data, and show that it represents a substantial improvement in functionality, scope, and accuracy with respect to current methods for cell morphometry.

Introduction

Since the advent of staining techniques in the 19th century, cell morphology has become one of the most described phenotypes in biology. The idea that the morphology of a cell is related to its function has been central to major discoveries, including the neuron doctrine[1], the molecular basis of sickle cell disease[2], and the pathways for cell migration and chemotropic sensing[3]. In the digital era, the study of cell morphology continues to have a prominent role. Cell shape can be indicative of disease and is a key diagnostic tool for some pathologies[4]. Image-based screens on cultured cells have uncovered the mechanism of action of multiple drugs[5, 6]. In the nervous system, thousands of neurons have been morphologically characterized using whole-cell patch-clamp[7], and the recent incorporation of high-throughput single-cell RNA-seq onto this technique, known as Patch-seq[8-12], has led to deeper characterizations including morphological, transcriptomic, and electrophysiological information of the same cells[13]. The combination of these techniques with Cre-dependent sparse labeling, high-resolution microscopy, and slice alternation methods is producing accurate morphological reconstructions of individual cells across macroscopic volumes[14, 15]. The potential of this new array of techniques is immense, not only for cell taxonomic purposes[16-21], but also for uncovering the molecular pathways that are associated with, and may ultimately drive, morphological cellular processes. However, to ensure progress in these directions, the implementation of high-resolution cell morphology profiling techniques needs to be accompanied by the development of computational methods that can take full advantage of them.

Algorithms for cell morphometry seek to determine similarities among the morphology of individual cells in digitally reconstructed microscopy images. They extract a set of shape descriptors that summarize the morphology of each cell, and then quantify differences between descriptors. Simple geometric descriptors consist of general morphological readouts like the area, perimeter, and lengths of major and minor axes of the cell[22, 23]. Although they can be

applied to most cell types and are invariant under rigid transformations, they cannot accurately discriminate complex cell morphologies like those of neurons and glia. Thus, cell-type-specific descriptors include more complex morphological characteristics, such as neuronal branching patterns (e.g. Sholl analysis[24], L-measure[25], and SNT[26]), which can be used to summarize complex morphologies. However, these descriptors need to be tailored to the specific cell type of interest and cannot be broadly applied. In addition, they are arbitrary with respect to the features that are used, and the weight assigned to them. To overcome some of these limitations, other methods generate a similarity score based on tree alignment (e.g. NBLAST[27] and BlastNeuron[28]) or decomposition in Fourier, Zernike, or spherical harmonic moments[29-32]. These methods require building combinations of descriptors that are invariant under rigid transformations or carefully pre-aligning the cells using Procrustes analysis, and therefore fail to accurately quantify morphological differences between highly dissimilar cells. Additionally, the similarity scores produced by these and other methods, like those based on persistent homology[33, 34], do not directly reflect the biophysical processes involved in cell morphological changes and do not lead to an actual mathematical distance function. These limitations have precluded the development of algebraic and statistical approaches for integrating and batch-correcting cell morphology spaces, constructing consensus cell morphologies, or inferring cell state trajectories associated with morphological processes.

Here, we build upon recent developments in applied geometry and shape registration[35-37] to establish a computational framework for summarizing complex and heterogeneous 2D and 3D cell morphologies across the broad range of cells found in tissues. This framework enables the characterization of morphological cellular processes from a biophysical perspective and produces an actual mathematical distance upon which rigorous algebraic and statistical analytic approaches can be built. The resulting framework has the generality and stability of simple geometric shape descriptors, the discriminative power of cell-type specific descriptors, and the unbiasedness and hierarchical structure of moments-based descriptors. Using this approach,

we address several outstanding methodological gaps in relating cell morphology to mechanism and function, including the combined analysis of morphological, molecular, and physiological information of individual cells, the integration of morphological information across technologies, and the analyses of morphological covariates. We expect this framework, and the accompanying software *CAJAL*, to greatly enhance morphometric analyses by not only increasing their accuracy and versatility, but also by enabling currently unavailable analyses such as the integration of cell morphology data across experiments.

Results

A general framework for the quantitative analysis of cell morphology data

In its simplest formulation, the study of cell morphology involves the quantitative comparison of cell shapes irrespective of distance-preserving transformations (*isometries*), such as rotations and translations. From a mathematical standpoint, this is a problem of metric geometry. The Gromov-Hausdorf (GH) distance[38, 39] measures how far two compact metric spaces are from being isometric. In physical terms, it determines the minimum amount of deformation required to convert the shape of an object into that of another. The use of the GH distance to describe cell shapes is therefore broadly applicable to any cell type, as it does not rely on geometric features that are particular to the cell type or require pre-aligning the cells to a reference shape. Because of these reasons, we sought to develop a general framework for cell morphometry by building upon these concepts in metric geometry.

Since computing the GH distance is intractable even for relatively small datasets, we based our approach upon a computationally efficient approximation, referred to as the Gromov-Wasserstein (GW) distance[35-37]. The GW distance preserves most of the mathematical properties of the GH distance and leads to an actual distance function[36]. Although its running

time grows cubically with the number of points, its efficiency can be further improved by means of nearly linear-time approximations that build upon optimal transport regularization[40, 41] and nesting strategies[42].

The starting point to our analytic framework is the 2D or 3D segmentation masks of individual cells, which are discretized by evenly sampling points from their outline (Fig. 1a). For each cell, we compute the pairwise distance matrix (d_i) between its sampled points. Then, for each pair of cells, i and j, we compute the GW distance between the matrices d_i and d_j using optimal transport (Fig. 1b). The result is a pairwise GW distance that quantifies the morphological differences between each pair of cells.

Different metrics for measuring distances between sampled points lead to different properties of the GW distance that may be advantageous in specific applications (Fig. 1a). For example, using Euclidean distance results in a GW matrix that accounts for the positioning of cell appendages, which can be particularly useful in the study of neuronal projections. On the other hand, using geodesic distance results in a GW matrix that is invariant under bending deformations of the cell, and it is therefore particularly sensitive to topological features such as the branching structure of cell appendages.

In all cases, the resulting GW distance can be thought of as a distance in a latent space of cell morphologies (Fig. 1c). In this latent space, each cell is represented by a point, and distances between cells indicate the amount of physical deformation needed to change the morphology of one cell into that of another. By formulating the problem in this way, we can use statistical and machine learning approaches to define cell populations based on their morphology; dimensionally reduce and visualize cell morphology spaces; integrate cell morphology spaces across tissues, technologies, and with other single-cell data modalities (for example, single-cell RNA-seq or ATAC-seq data); or infer trajectories associated with continuous morphological processes. We have implemented these analyses in an open-source

Python library, called *CAJAL*, which can be used with arbitrarily complex and heterogeneous cell populations (Fig. 1d).

GW cell morphology spaces accurately summarize complex cell shapes

To assess the ability of GW cell morphology spaces to summarize complex cell shapes, we applied *CAJAL* to the 3D basal and apical dendrite reconstructions of 506 neurons from the mouse visual cortex profiled with patch-clamp[20]. The resulting space of cell morphologies recapitulated the neuronal morphological types of the visual cortex (Fig. 2a). Cells with a similar morphology appeared in proximity in the UMAP representation of this space. Molecularly defined neuronal types were also localized in the representation (Fig. 2b), consistent with the presence of morphological characteristics that are unique to each molecular subtype. Excitatory and inhibitory neurons clustered separately, and individual neurons were organized in the cell morphology space according to their cortical layer and Cre driver line (Fig. 2b). Clustering the morphology space using Louvain community detection[43] partitioned it into 9 morphological populations. Using the metric structure of the cell morphology space, we then computed the medoid and average cell morphology for each cluster (Fig. 2c). These summaries accurately represented the main morphological characteristics of each cell population and were consistent with the diversity of neuronal morphologies found in the mouse visual cortex[20].

To quantitatively evaluate the ability of the GW distance to accurately summarize complex neuron morphologies in comparison to state-of-the-art methods for neuron morphometry, we considered two Patch-seq datasets of the visual[19] and motor cortex[21] in addition to the patch-clamp dataset of the visual cortex. For each dataset, we assessed the ability of *CAJAL* and 5 other methods (Sholl analysis[24], L-measure[25], SNT[26], NBLAST[27], and TMD[33]) to identify morphological differences between molecularly-defined neurons. In the case of the patch-clamp dataset, we considered neurons labeled with different Cre driver lines, for a

total of 31 lines, with the understanding that each line preferentially labels distinct neuronal types. In the case of the two Patch-seq datasets, we considered the known classification of motor and visual cortex neurons into 9 and 6 transcriptionally-defined subtypes (t-types), respectively, based on their gene expression profile[19, 21]. We used three different metrics of performance to evaluate the ability of each method to predict the molecular type of each individual neuron based on its morphology: the multi-class Matthews Correlation Coefficient of a k nearest neighbor classifier, the Calinski-Harabasz clustering score of neurons from the same molecular type in the cell morphology space, and the benchmarking score introduced in a previous study of cell shape analysis methods[30]. In this comparative study, we found that the GW distance outperformed existing methods for neuron morphometry (Fig. 2d). In addition, some of the best performing methods, such as TMD, produced errors and were not able to summarize the morphology of 26 neurons for which the model assumptions were violated. A similar evaluation of the ability of each algorithm to identify 47 t-types of inhibitory neurons and 26 t-types of excitatory neurons in the motor cortex[21] led to consistent results (Supplementary Fig. 1). As expected, the accuracy and running time of *CAJAL* in these analyses increased with the number of points that are sampled from the outline of the cell (Supplementary Fig. 2). However, for these datasets the accuracy saturated at approximately 100 points, indicating no major advantage in sampling a larger number of points. Taken together, these results demonstrate the utility and versatility of the GW distance to perform unbiased studies of complex cell morphologies.

GW cell morphology spaces summarize cell shapes across heterogeneous cell populations

We next evaluated the ability of the GW distance to summarize cell morphologies across very heterogenous cell populations. For that purpose, we used *CAJAL* to study the morphologies of 70,510 cells from a cubic millimeter volume of the mouse visual cortex profiled

by the Machine Intelligence from Cortical Networks (MICrONS) program using two-photon microscopy, microtomography, and serial electron microscopy[15]. This dataset includes not only neurons, but also multiple types of glia and immune cells. The UMAP representation of the cell morphology space produced by *CAJAL* recapitulated in an unsupervised manner the broad spectrum of cell types that are present in the tissue, including several populations of neurons, astrocytes, microglia, and immune cells (Fig. 3a). These populations were consistent with the manual annotations of 185 individual cells provided by the MICrONS program (Fig. 3b). Neurons from different cortical layers were associated with distinct regions of the cell morphology space, indicating the presence of morphological differences between neurons from different layers (Fig. 3c). In addition, our analysis uncovered morphological differences between astrocytes located in different cortical layers (Figs. 3d, e). Specifically, layer 1 astrocytes were substantially smaller, and layer 2/3 astrocytes were elongated perpendicularly to the pial surface, in comparison to astrocytes residing in other cortical layers (Figs. 3d, e). These morphological differences were consistent with recent observations in Glast-EMTB-GFP transgenic mice[44].

To quantitatively evaluate the ability of GW distance to summarize cell morphologies across different cell types in comparison to existing general approaches for cell morphometry, we considered the 3D morphological reconstructions of 512 T cells from the mouse popliteal lymph node, submandibular salivary gland, and skin, profiled with intra-vital two-photon microscopy[29], in addition to the neuronal patch-clamp dataset of the mouse visual cortex[20]. We evaluated the ability of *CAJAL* and 4 other general approaches for cell shape analysis (CellProfiler[45], SPHARM[29, 32], Zernike moments[30, 31], and the PCA-based approach of Celltool[30] and VAMPIRE[46]) to predict the anatomical location of each T cell and the Cre driver line of each neuron based on their morphology. In this study, *CAJAL* was again the most capable algorithm at separating the morphologies of T cells from different tissues according to most metrics (Fig.

3f). In addition, all methods for general cell shape analysis except for *CAJAL* performed poorly in the analysis of complex neuronal morphologies, showing that no other method was able to perform well in both datasets (Fig. 3f). Altogether, these results demonstrate that the GW distance overcomes the limitations of current methods to summarize the broad range of complex cell shapes present in mammalian tissues.

Multimodal analyses of GW cell morphology spaces enable uncovering genetic determinants of cell morphology

The combined analysis of morphological and genomic data from individual cells has the potential to unravel the genetic and molecular pathways that are associated with the progression of high-level cellular processes such as cell differentiation and plasticity. Since changes in cell morphology are continuous, establishing associations between cell morphology and other data is best accomplished by methods of analysis that are purpose-built for continuous processes. We extended our previous work on clustering-independent analyses of omics data[47] to implement a statistical approach for identifying molecular and physiological features that are associated with changes in cell morphology. We use the Laplacian score for feature selection[48] to test the association between the values of each feature and the structure of the morphology space, while accounting for user-specified covariates such as the age of the individual (Fig. 4a). To illustrate this approach, we used it to identify genes that contribute to the morphological plasticity of neurons in the *C. elegans*. For that purpose, we considered the 3D morphological reconstructions of the male *C. elegans* GABAergic DVB interneuron in 12 gene mutants, 5 double mutants, and controls across days 1 to 5 of adulthood (Supplementary Table 1), including 7 gene mutants and 1 double mutant from a previous study[49]. The DVB neuron develops post-embryonically in the dorsorectal ganglion and undergoes post-developmental neurite outgrowth in males, altering its morphology and synaptic connectivity, and contributing to changes in the spicule protraction step of male mating behavior[50]. We applied our approach to

identify loss of function mutations that are associated with changes in the dynamic morphology of the DVB neuron, taking the age of the worm as a covariate to reliably compare morphological changes across timepoints in adulthood. This analysis identified mutations in the genes *unc-97, lat-2*, *nlg-1, unc-49, nrx-1*, and *unc-25* as significantly affecting the morphology of the DVB neuron (Fig. 4b-d; Laplacian score permutation test, FDR < 0.05). By repeating this analysis for worms of each age separately, we identified the age at which each of these mutations starts significantly affecting the morphology of the DVB neuron (Fig 4e). To interpret these morphological differences in terms of neuronal characteristics, we used the same approach to evaluate the association of 33 morphological features with the structure of the cell morphology space (Supplementary Table 2). Within these significantly associated features, we found that mutations in *nlg-1* and *unc-25* caused an increase in neurite length and number of branches compared to control worms (Supplementary Fig. 3), while inactivating mutations in *unc-97* and *nrx-1* stunted neurite growth (Supplementary Fig. 3). Altogether, these results are consistent previous findings[49] and extend them by uncovering new genetic determinants of neuronal plasticity in *C. elegans* and quantitative differences in the age of onset of the morphological alterations induced by different genes.

An integrative analysis of molecular, physiological, and morphological data from single cells identifies continuous morpho-transcriptomic trajectories

The incorporation of single-cell RNA-seq onto whole-cell patch-clamp, known as Patch-seq, has enabled concurrent high-throughput measurements of the transcriptome, physiology, and morphology of individual cells[13]. The integrative analysis of these multi-modal data has the potential to uncover the transcriptomic and physiological programs associated with morphological changes of cells.

We used *CAJAL* to analyze the basal and apical dendrites of 370 inhibitory and 274 excitatory motor cortex neurons profiled with Patch-seq[21]. Consistent with our previous results,

the GW cell morphology space captured morphological differences between the dendrites of neurons from different neuronal t-types and cortical layers (Fig. 5a, b). By representing the pairwise distance between each pair of cells in the transcriptomic, electrophysiological, and morphological latent spaces as a point in a 2D simplex, we found a large degree of variability in the morphology of the dendrites of extratelecephalic-projecting (ET) neurons that was not paralleled by their gene expression profile (Fig. 5c). In contrast, the dendrites of *Lamp5*$^{+}$ and bipolar (*Vip*$^{+}$) GABAergic neurons presented limited morphological variability in comparison to their gene expression profile (Fig. 5c).

We characterized the gene expression and electrophysiological programs associated with morphological differences between neurons by using the Laplacian score approach described above. We performed this analysis separately for inhibitory and excitatory neurons to identify 173 and 556 genes, and 14 and 22 electrophysiological features, respectively, that were significantly associated with the morphological diversity of their dendrites (Fig. 5d and Supplementary Tables 3 and 4; Laplacian score permutation test, FDR < 0.1). Among the 7 genes that were significant for both excitatory and inhibitory neurons, there were several that have been previously reported to be involved in dendrite morphogenesis, such as *Dscam*, which plays a central role in dendritic self-avoidance[51], and *Pcdh7*, which regulates dendritic spine morphology and synaptic function[52]. Consistent with these results, a gene ontology enrichment analysis for biological processes identified neuron projection morphogenesis among the top ontologies associated with the morphological diversity of excitatory neurons (GO enrichment adjusted *p*-value = 0.009), and cellular response to chemical stimulus, neuron differentiation, and taxis among the top ontologies associated with the morphological diversity of inhibitory neurons (GO enrichment adjusted *p*-values = 0.006, 0.008, and 0.01, respectively).

We next investigated if some of these gene expression programs form part of continuous morpho-transcriptomic cellular processes. We computed the RNA velocity field to predict the future gene expression state of each cell based on the observed ratio between un-spliced and

spliced transcripts[53, 54]. The time scale of these predictions is determined by the mRNA degradation rate and is of the order of hours. We reasoned that by projecting the RNA velocity field onto the GW cell morphology space and looking for transcriptomic trajectories that also appear as trajectories in this space, we could identify continuous cellular processes that involve consistent changes in gene expression and cell morphology. This approach revealed several morpho-transcriptomic trajectories involving chandelier, basket, and *Lamp5*$^{+}$ neurons (Fig. 5e). Cells along these trajectories showed increased complexity in their apical and basal dendrites in parallel to changes in their gene expression profile, in agreement with the presence of molecular programs associated with the plasticity of these neuronal types. To characterize these molecular programs, we focused our analysis on 78 genes that were associated with the RNA velocity field of inhibitory neurons and computed the Laplacian score of each of these genes in the GW cell morphology space. This analysis revealed that 32 of the 78 genes were also significantly associated with the structure of the cell morphology space (Supplementary Table 5; Laplacian score permutation test, FDR < 0.05). The list of significant genes included multiple genes coding for secreted factors, such as *Spon1*, *Fgf13*, *Rspo2*, and *Reln* (Supplementary Fig. 4), and was enriched for genes involved in memory and cognition (GO enrichment adjusted *p*-value = 0.007).

Taken together, these results demonstrate the utility of *CAJAL* to identify and characterize molecular and electrophysiological programs associated with cell morphological changes based on single-cell Patch-seq data.

GW cell morphology spaces enable the integration of cell morphology data across technologies

Advances in cell morphology profiling techniques have led to an explosion of high-resolution cell morphology data over the past decade[7]. The ability to perform integrated analyses of such data regardless of the experimental approach and technology that was used to

generate them would be a powerful tool for imputing missing data and refining taxonomic classifications of cells. For example, by integrating patch-clamp and Patch-seq data the transcriptome of cells profiled with patch-clamp could be predicted based on their morphology.

We used *CAJAL* to build a combined morphology space of the basal and apical dendrites of visual cortex neurons profiled with patch-clamp[20] and visual and motor cortex neurons profiled with Patch-seq[19, 21]. The combined dataset consisted of 1,662 neurons, of which 1,156 had associated single-cell RNA-seq data. Inhibitory and excitatory neurons from different datasets clustered together in separated regions of the combined morphology space (Fig 6a), indicating that the structure of this space is mostly driven by biological differences rather than by technical differences. To evaluate the consistency of the combined cell morphology space, we considered the t-type[18] of the cells profiled with Patch-seq, and quantitatively assessed the distance in the combined cell morphology space between cells of the same t-type (for the Patch-seq data) or cells labelled with the corresponding Cre driver line (for the patch-clamp data). Cells of same t-type but from different Patch-seq datasets, as well as cells from the matching Cre driver line in the patch-clamp dataset, were closer to each other in the combined morphology space than cells from different t-types or Cre driver lines (Fig. 6b and Supplementary Fig. 5; Wilcoxon rank-sum test p-value $< 10^{-100}$), demonstrating the utility of the GW distance for integrating cell morphology data across experiments and technologies.

We then used the same approach to refine the annotation of visual cortical neurons profiled with serial electron microscopy by the MICrONS program[15]. We considered 883 full neuron reconstructions from the two Patch-seq datasets and created a combined morphology space of these cells along with a subset of 1,000 evenly sampled and 139 manually annotated neurons from the MICrONS dataset. As with the integration of patch-clamp and Patch-seq data, the manually annotated neuronal types from the MICrONS dataset were closer to Patch-seq cells of the matching t-type in the consolidated cell morphology space than to non-matching t-types (Fig 6c and Supplementary Fig. 6; Wilcoxon rank-sum test p-value $< 10^{-100}$). For example,

the only chandelier cell annotated in our MICrONS dataset was closer to *Pvalb Vipr2* t-type cells from the Patch-seq data than to cells from other t-types (Supplementary Fig. 6; Wilcoxon rank-sum test *p*-value = 0.035).

Using this combined cell morphology space, we refined some of the manual annotations of the MICrONS data with more precise transcriptomic definitions. For example, of the three Martinotti cells annotated in the MICrONS dataset, one cell presenting a distinct morphology with a densely arborized axon was closer in the morphology space to Patch-seq cells of the *Sst Chrna2* t-type (Fig. 6d; Wilcoxon rank-sum test *p*-value = 0.045), while the other two Martinotti cells were closer to *Sst Calb2* t-type cells (Fig. 6d; Wilcoxon rank-sum *p*-value = 10^{-3}). This is consistent with previous results showing that expression of *Chrna2* is characteristic of layer 5 Martinotti cells that project into layer 1[55], and we confirmed that the soma of the predicted *Chrna2* Martinotti cell was indeed located in layer 5 while its long axon ended in layer 1 (Fig. 6d). Similarly, among the manually annotated basket cells in the MICrONS dataset, one had a more condensed morphology than the others (Fig. 6e). This smaller basket cell was close in the cell morphology space to *Vip Chat Htr1f* and *Vip Col15a1 Pde1a* t-type Patch-seq cells (Fig. 6e; Wilcoxon rank-sum *p*-value = 0.02), while larger basket cells were closer to *Pvalb Sema3e Kank4* and *Pvalb Gpr149 Islr* t-type Patch-seq cells (Fig. 6e; Wilcoxon rank-sum *p*-value = 10^{-14}). These results were again in agreement with the molecular characterization of small and large basket cells in the somatosensory cortex[56].

Taken together, these results demonstrate the utility of GW cell morphology spaces to perform integrative analyses of cell morphological data across technologies and represent a conceptual basis for the development of algorithms for cell morphology data integration and batch-correction.

Discussion

Shape registration has experienced several breakthroughs over the past 15 years with the formalization of new paradigms that allow for more flexibility in the quantification of morphology[57]. Here, we built upon one of these constructions, the GW distance, to develop a general computational framework and software for the combined morphometric, transcriptomic, and physiological analysis of individual cells. The proposed framework does not rely on predefined morphological features, is insensitive to rigid transformations, and can be efficiently used with arbitrarily complex and heterogeneous cell morphologies. Using this approach, we have been able to accurately build, analyze, and visualize cell morphology summary spaces, where each cell is represented by a point and distances between cells indicate the amount of physical deformation needed to change the morphology of one cell into that of another. We have integrated morphological data across experiments and technologies; identified morphological, molecular, and physiological features that define cell populations; and established associations between morphological, molecular, and electrophysiological cellular processes. Our quantitative comparative studies using published Patch-seq, patch-clamp, electron, and two-photon microscopy datasets demonstrate that the proposed approach represents a boost in accuracy, functionality, and scope with respect to current methods for the analysis of cell morphology data.

There are several limitations of the proposed approach, the most important one perhaps being its running time. In our studies, the application of *CAJAL* to 644 digital neuron reconstructions on an 8-core desktop computer took 70 minutes. We expect that the implementation of recent strategies for reducing the computing time of the GW distance[40, 42] might improve the scalability of *CAJAL* to larger datasets without having to reduce the number of sampled points per cell. This will become particularly important as the throughput of technologies for high-resolution cell morphology profiling continues to grow. In addition, the

interpretability of cell morphology spaces in terms of specific morphological features is limited in some situations. To address this aspect, in our studies we evaluated interpretable morphological descriptors produced by methods like SNT[26], L-measure[25], or CellProfiler[45] on the cell morphology space produced by *CAJAL* to combine the interpretability of these descriptors with the unbiasedness and accuracy of GW cell morphology spaces.

Overall, the application of metric geometry to cell morphometry serves as a pillar for the development of currently missing computational methods for integrating and batch-correcting cell morphology data, as well as for modelling continuous morphological cellular processes. We expect that the development of these methods during the next few years will significantly impact our understanding of the relation between the morphological, molecular, and physiological diversification of cells.

Methods

Computation of GW cell morphology spaces

We build upon the application of metric geometry to the problem of finding a correspondence between two point-clouds such that the size of non-isometric local transformations is minimized[35-37]. *CAJAL* takes as input the digitally reconstructed cells. For each cell i, it samples n points regularly from the outline and computes their pairwise distance matrix, d_i. It then computes the GW distance between every pair of distance matrices

$$GW(d_i, d_j) = \frac{1}{2} \min_{\mathrm{T_{ij}} \in \mathrm{C}} \sum_{\alpha,\beta,\delta,\gamma} \left| (d_i)_{\alpha\beta} - (d_j)_{\gamma\delta} \right|^2 (T_{ij})_{\alpha\gamma} (T_{ij})_{\beta\delta}$$

where the matrix T_{ij} specifies a weighted pointwise matching between the points of cells i and j, and C represents the space of all possible weighted assignments[36]. By construction, *CAJAL* does not require pre-aligning cell outlines, since the input to GW is the pairwise distance matrix within each cell, d_i, which is invariant under rigid transformations. Depending on the application, we consider two choices for the distances d_i: Euclidean and geodesic distance.

The output is a metric space for cell morphologies which can then be clustered and visualized using standard procedures, such as Louvain community detection[43] and UMAP[58]. For each population of cells, $\mathcal{X}$, we compute its average morphology as the distance matrix

$$(\hat{d}_{\mathcal{X}})_{\alpha\beta} = \frac{1}{|\mathcal{X}|} \sum_{i,\gamma} (T_{i\,\mathrm{med}(\mathcal{X})})_{\alpha\gamma} (d_i)_{\gamma\beta}$$

where $\mathrm{med}(\mathcal{X})$ denotes the medoid of $\mathcal{X}$ with respect to the GW distance matrix. The morphology can then be visualized by computing the shortest-path tree or multidimensional scaling (MDS) of $\hat{d}_{\mathcal{X}}$. In addition, to facilitate the interpretation of morphology spaces, we find it is useful to plot the values of standard morphological descriptors like cell height, width, diameter, neuronal depth, or fractal dimension[25, 26, 45] in the UMAP representation of the cell morphology space.

Evaluation of features on the cell morphology space

To evaluate features, such as gene expression or electrophysiological properties, on GW cell morphology spaces, we build upon a spectral approach for clustering-independent analyses of multimodal data[47, 48]. We first construct a radius neighbor graph of the GW distance with radius ε. Each feature g is represented by a vector f_g of length the number of cells. The Laplacian score of g on the cell morphology space is then given by[48]

$$C_g = \frac{\sum_{ij} \left((f_g)_i - (f_g)_j \right)^2 A_{ij}}{\mathrm{Var}(f_g)}$$

where A is the adjacency matrix of the radius neighbor graph, and $\mathrm{Var}(f_g)$ the estimated variance of f_g. Features with a low C_g score are associated with morphologically similar cells. The significance of the score can be statistically assessed for each feature by means of a one-tailed permutation test and adjusted for multiple hypothesis testing using Benjamini-Hochberg procedure. To assess the significance of a feature g in the presence of a set of covariates h_m, we perform a permutation test where the entries of f_g and f_{h_m} are simultaneously permuted and the scores C_g and C_{h_m} are computed at each permutation. We denote these values collectively as C_g^{null} and $C_{h_m}^{\mathrm{null}}$. We then solve the regression problem

$$C_g^{\mathrm{null}} \sim \beta_0 + \sum_m \beta_m C_{h_m}^{\mathrm{null}}$$

and consider the distribution of residuals as the null distribution for the adjusted score $\tilde{C}_g = C_g - \beta_0 - \sum_m \beta_m C_{h_m}$.

Processing of Patch-seq and patch-clamp morphological reconstructions

We downloaded the morphological reconstructions of neurons from several repositories. For the Gouwens *et al.* [20] Patch-clamp dataset, we downloaded 509 reconstructions in SWC format from the Allen Cell Types database, using Cell Feature Search and selecting for “Full” or

“Dendrite Only” reconstruction types. Three of the SWC files were unsorted and were left out of further processing, for a total of 506 neurons. For the Gouwens *et al.*[19] Patch-seq dataset, we downloaded 574 reconstructions from the Brain Image Library (BIL) repository. We removed 62 neurons that did not have assigned transcriptomic types, for a total of 512 neurons. Lastly, for the Scala *et al.* [21] Patch-seq dataset, we downloaded 645 reconstructions from the inhibitory and excitatory sets from the BIL repository, skipping the one inhibitory neuron that had no dendrites, for a total of 644 neurons.

The SWC format represents each neuron as a tree of vertices, such that an edge can be drawn between a vertex and its parent, forming the skeleton of the neuron. From this format, we sampled 100 points radially around the soma at a given step size. We used a binary search to identify the step size which returns the required amount of evenly spaced points. To calculate the pairwise geodesic distance between these points, we constructed a weighted graph with weights given by the distance to the latest sampled point. We then used the Floyd-Warshall algorithm implemented in Networkx[59] to compute all pairwise shortest path distances in this graph. Alternatively, we computed the pairwise Euclidean distance between the 3D coordinates of these points.

We the computed the GW distance between each pair of cells as described above (subheading “Computation of GW cell morphology spaces”) using the ot.gromov.gromov_wasserstein function of the “POT: Python Optimal Transport” Python library[60]. We then used this precomputed distance to build a 2D visualization of the morphology space using the https://github.com/tkonopka/umap package in R. We computed the Louvain clusters of a KNN graph of the GW distance using the multilevel.community function of the igraph R package[61].

Average shape of neurons

To compute the average shape of a cluster of cells in the GW cell morphology space, we first found the medoid cell as the cell with the minimum sum of distances to all other cells in the cluster. To compute a morphological distance between cells, the GW algorithm identifies an optimal matching T_{ij} between the points we have sampled (subheading "Computation of GW cell morphology spaces"). We used this matching to align the other cells in the cluster to the medoid, by reordering the pairwise geodesic distance matrix of their sampled points to match the distance matrix of the medoid cell. We rescaled the geodesic distance matrix of each cell into an unweighted graph distance by dividing out the minimum distance between any two points, so that the rescaled distances were integers. We set a threshold on these distances at 2, such that the distance was 0 from the point to itself, 1 to an adjacent point in the tree of the neuron trace, and 2 to any farther point. We averaged all of these distance matrices together over the cells in the cluster and built a $k = 3$ nearest neighbor graph, essentially connecting each sampled point to the three other points it was most often adjacent in the neurons of that cluster. We took the shortest path tree in this graph as the average shape for that cluster using Dijkstra's algorithm. We color each point in this average shape by a confidence value based on its minimum original unweighted graph distance, summed over the cluster, to any other point.

Comparison of *CAJAL* with current methods for neuronal morphometry

We compared our approach to five other morphological methods for neuron analysis by applying them to the dendrite reconstructions of the neuronal datasets listed above (subheading "Processing of Patch-seq and patch-clamp morphological reconstructions"). These methods have stricter assumptions on the input, forcing us to remove disconnected dendrites from the reconstructions. We applied NBLAST[27], as implemented in the nat.nblast R package (https://github.com/natverse/nat.nblast). We calculated a pairwise distance between all neurons using the nblast_allbyall function with the mean normalization method. We ran the Topological

Morphology Descriptor (TMD) method of Kanari *et al.*[33] using the TMD Python package (https://github.com/BlueBrain/TMD). We followed their distances example (https://github.com/BlueBrain/TMD/blob/master/examples/distances_example.py) to compute the persistence image difference between every pair of neurons. We skipped 26 neurons across the two Patch-seq datasets for which get_ph_neuron or get_persistence_image_data errored due to a lack of bifurcating branches. We used the Measure Multiple Files batch script of the ImageJ SNT plugin[26] to compute morphological features of neurons, including the Sholl features[24]. We also computed morphological features using L-Measure[25], selecting all of their provided functions.

We used three different metrics to assess the ability of these algorithms to identify morphological differences between Cre lines or transcriptomic types. We computed the Calinski-Harabasz clustering score using cluster.stats from the fpc R package (https://cran.r-project.org/package=fpc). We also implemented the median-based group discrimination statistic used by Pincus *et al.*[30] to compare methods for cell-shape analysis. Lastly, we used a 7-fold $k = 10$ nearest neighbor classifier from the scikit-learn Python library to predict the Cre driver line or t-type of each cell based on morphological distance and used the Matthew's correlation coefficient to evaluate the accuracy of the predictions.

Morphological analysis of the MICrONS dataset

We downloaded the 113,182 static cell segmentation meshes from MICrONS using the trimesh_io module from the package MeshParty (https://meshparty.readthedocs.io/) at the lowest resolution (resolution 3). We then downloaded higher resolution meshes for cells that had less than 10,000 vertices at this lowest resolution. Cells with less than 1,001 vertices at the lowest resolution were re-downloaded at the highest resolution (resolution 0). Cells with 1,001 to 3,000 vertices at the lowest resolution were re-downloaded at resolution 1, and cells with 3,001 to 10,000 vertices were re-downloaded at resolution 2.

Along with other metadata available through the CAVEclient (https://github.com/seung-lab/CAVEclient), such as the 3D coordinates of neuron soma, we collected the cell IDs for each manually annotated cell type provided by the MICrONS program in their website. We used the layer 2/3, layer 4, and layer 5 manually annotated cells to estimate cortical layer boundaries in the y values of the 3D soma coordinates. We placed these cutoffs at layer 1 < 104,191 < layer 2/3 < 133,616 < layer 4 < 179,168 < layer 5 < 213,824 < layer 6.

We sampled 50 vertices from the triangular mesh of each cell, using the linspace function of the NumPy package[62] to evenly select vertices, since vertices were roughly ordered by proximity, and this gives an approximation of even sampling over the 3D space. We skipped the very large blood vessel mesh and 240 meshes with less than 50 vertices, for a total of 112,941 meshes. We used the heat method[63], implemented in the MeshHeatMethodDistanceSolver function of the Python library potpourri3D (https://github.com/nmwsharp/potpourri3d), to compute geodesic distances between the sampled points on the mesh. We parallelized the computation of the pairwise GW distance between the 112,941 meshes on 128 cores, but otherwise used the same process as with the Patch-seq and patch-clamp datasets (subheading “Processing of Patch-seq and patch-clamp morphological reconstructions”). Due to the large size of the resulting GW pairwise distance matrix, we used the Python libraries leidenalg[64] and umap-learn[65] to cluster the cells and compute 2D UMAP visualizations, respectively. We labelled the clusters based on the manually annotated cells provided by the MICrONS program.

We found that some morphological clusters mostly consisted of artifacts or neuron-glia doublets and removed those. In addition, another morphological cluster contained both neuron-neuron doublets and individual neurons with complex morphologies, so we devised an approach to remove meshes containing multiple somas from that cluster. We determined the number of somas in each mesh from that cluster by using MeshParty to skeletonize the meshes and convert them into graph representations where nodes have a radius value, and nodes within

soma regions fall in a specific range of radii. For us, this range was 4,000 to 30,000. We used HDBSCAN[66] to cluster these nodes in the 3D space and counted each cluster with at least three nodes as a soma. Meshes with more than one soma were removed from the cluster. Lastly, we noticed that many meshes with very high y coordinates appeared stretched, so we removed meshes with a y soma coordinate greater than 240,000. After removing all these artifacts, we recomputed a UMAP visualization of the remaining 70,510 cells in the cell morphology space using umap-learn.

For each astrocyte, we measured the bounding box by placing lower and upper bounds on the 1% and 99% quantiles of the mesh vertices along each of the first three principal components. We took the arccosine of the first principal component along the y axis to be the orientation angle of the astrocyte and measured its deviation from perpendicular.

Morphological analysis of T cells

We retrieved 512 3D meshes of T cells from Medyukhina *et al.*[29]. We evenly sampled 200 points from the list of vertices in each mesh, which approximates an even sampling in 3D space since the vertices are roughly ordered in a spiral down the cell. We computed the GW distance of the pairwise Euclidean distances between these points as described above (subheadings “Computation of GW cell morphology spaces” and “Processing of Patch-seq and patch-clamp morphological reconstructions”).

Comparison of CAJAL with general methods for cell morphometry

We applied the Celltool method of Pincus *et al.*[30] using their Python package (https://github.com/zpincus/celltool). Since this method only works with 2D cell segmentations, we sampled the 2D boundary of the projection of each cell along the first two axes to the same number of sampled points used for *CAJAL*. We aligned these contours using a maximum of 20 iterations, allowing for reflections, and saved the non-normalized PCA values from the shape model. We used CellProfiler 4.0.3[22] on binary 2D projection images to compute both general

shape features and Zernike moments using MeasureObjectSizeShape. We ran SPHARM[29] using their Python package (https://github.com/applied-systems-biology/Dynamic_SPHARM) on all of the mesh vertices for each cell. For neurons, we used the marching_cubes function of the scikit-learn Python library to define 3D mesh vertices. We used the spectrum.return_feature_vector function of SPHARM to extract the amplitude of harmonic components from the spectra produced by compute_spharm. We compared these methods to *CAJAL* using the same metrics described above (subheading "Comparison of *CAJAL* with current methods for neuronal morphometry").

Evaluation of the accuracy and runtime of *CAJAL* as a function of the number of sampled points

We sampled 25, 50, 75, 100, and 200 points from each cell from the patch-clamp dataset of Gouwens *et al.*[20] and applied *CAJAL* as described above to compute the GW distance between cells. We used the Calinski-Harabasz score, the median-based statistic of Pincus *et al.*[30], and the Matthews coefficient of a $k = 10$ nearest neighbor classifier (subheading "Comparison of *CAJAL* with current methods for neuronal morphometry") to assess how the number of sampled points affects the ability of *CAJAL* to capture morphological differences between cells from different Cre driver lines. Runtimes were determined based on 12 threads of a desktop computer with an 8-core Intel Xeon E5-1660 3.20 GHz CPU.

Morphological analysis of the DVB neuron

We considered the neurite reconstructions of the DVB neuron from 799 adult male *C. elegans* aged 1 to 5 days from control strains or strains containing mutations in the genes *nrx-1, mir-1, unc-49, nlg-1, unc-25, unc-97, lim-6, lat-2, ptp-3, sup-17,* or *pkd-2* (Supplementary Table 2). Reconstructions were created from confocal images of the DVB neuron using SNT[26] in Fiji[67] as previously described[49]. We computed the GW distance between these morphological reconstructions as described above (subheadings "Computation of GW cell morphology spaces"

and “Processing of Patch-seq and patch-clamp morphological reconstructions”), based on the geodesic pairwise distance of 100 points sampled from each neuron. We then introduced an indicator function for each mutated gene, which took values 0 or 1 on each cell depending on whether the worm had a wild-type or a mutated version of the gene, respectively. To determine which of the 11 mutated genes were associated with changes in morphology, we computed the Laplacian score of each indicator function on the GW cell morphology space as described above (subheading “Evaluation of features on the cell morphology space”). To compute the score we used the R package RayleighSelection[47] with 1,000 permutations, ε equal to the median GW distance and the age of the worm in days as a covariate. In the same way, we used RayleighSelection to determine which of 33 morphological features computed with SNT were significantly localized in the cell morphology space. In addition, we performed the same analysis using only neurons from a single day, for each day, to determine the age at which the effect of significant mutations on the morphology of the DVB neuron starts to emerge.

Identification of genes and electrophysiological features associated with the morphology of neuronal dendrites

We used the same process described above (subheading “Processing of Patch-seq and patch-clamp morphological reconstructions”) to compute the GW distance between the morphological reconstructions of the dendrites of 644 neurons profiled by Scala *et al.*[21]. We sampled 100 points from each dendrite and used geodesic distance to measure the distance between points. To determine which genes are associated with morphological variability we computed the Laplacian score of each gene on the GW morphology space using RayleighSelection, as described above (subheading “Evaluation of features on the cell morphology space”). Gene expression values were normalized as $\log(1 + 5000 \cdot$size-normalized expression), we used 1,000 permutations, and ε was given by the median GW distance. We only tested genes expressed in at least 5% and less than 90% of cells. We identified gene

ontology enrichments using the R package gProfileR[68], where we performed an ordered query of the significant genes based on their Laplacian score and restricted the search to biological process (BP) gene ontologies. We used the same procedure based on the Laplacian score to determine which electrophysiological features were associated with changes in the morphology of the dendrites.

Computation of RNA velocity trajectories

We clipped 3' Illumina adapters and aligned FASTQ files to the GRCm38 mouse reference genome using the STAR aligner[69]. We used the command "run_smartseq" from the velocyto command line tool[53] to create a Loom file of spliced and unspliced reads. We then used the scvelo Python package[54] to compute RNA velocity trajectories. We tested scvelo in dynamical or stochastic mode with 0, 10, or 20 minimum counts; 500, 1000, or 2000 top variable genes; 10, 20, or 30 principal components; and 10 or 30 neighbors. We kept the velocity trajectories with the highest average confidence per arrow, defined by agreement with neighboring arrows. These trajectories were produced using stochastic mode with 0 minimum counts, 500 top variable genes, 10 principal components, and 30 neighbors. We computed the pseudotime using the velocity graph. We took all 78 genes which passed the basic default filters in rank_velocity_genes() to be velocity-related genes and used the Laplacian score to assess their morphological association.

Consistency between transcriptomic, electrophysiological, and morphological spaces

We defined the transcriptomic distance (d_T) between two cells as the Spearman correlation distance between their log-normalized gene expression profile, and their electrophysiological distance (d_E) as the Euclidean distance between their electrophysiological feature vectors. We compared these distances and the GW morphological distance (d_M) between all pairs of cells in the Scala *et al.*[21] dataset by representing them on a 2-simplex. For that purpose, we standardized the logarithm of pairwise distances independently for each data

modality. We then took the axes of the 2-simplex to be the given by the difference between each pair of distances ($d_M - d_T$, $d_T - d_E$, $d_E - d_M$), so that the sum of the coordinates equals 0 for each pair of cells. We plotted cell pairs in the middle 98% of each axis.

Integrative analysis of Patch-seq and patch-clamp data

We combined the patch-clamp and two Patch-seq datasets into one cell morphology space by computing the GW distance between the morphological reconstructions of the dendrites of all 1,662 neurons from the 3 datasets. We sampled 100 points from each dendrite and used geodesic distance to measure distances between points.

To evaluate the integration of the Patch-seq datasets, we utilized the classification of neurons into the t-types of Tasic *et al.*[18]. This classification is provided by Gouwens *et al.*[19] as their transcriptomic alias, and we computed the classification for the dataset of Scala *et al.*[21] using their ttype-assignment Jupyter notebook. We tested the overlap between neurons of the same t-type but from different datasets in the cell morphology space by performing a Wilcoxon rank-sum test, comparing the distribution of GW distances within the same t-type with the distribution of GW distances between t-types.

To evaluate the integration between the two Patch-seq datasets and the patch-clamp dataset, we matched the neuronal t-types in the Patch-seq datasets with the Cre driver lines in the patch-clamp dataset. We used only the first marker in the t-types and considered markers that existed in at least five cells of two of the three datasets. This left Sst, Pvalb, and Vip as major markers between the t-types and Cre lines, and Lamp5 and Sncg as markers between t-types only. We again used the Wilcoxon rank-sum test to compare the distributions of GW distances within and between these five major transcriptomic types.

Integrative analysis of Patch-seq and MICrONS neuronal data

We calculated a combined GW morphological space for the two Patch-seq datasets and 1,000 neurons evenly sampled from the MICrONS dataset, in addition to 140 manually

annotated neurons by the MICrONS program. We sampled 50 points from the full neuronal reconstructions from the Patch-seq datasets. In the case of the Scala *et al.*[21] dataset, this restricted our analysis to 370 neurons with full reconstructions. Since the SWC format used in the Patch-seq datasets contains a trace reconstruction, and the triangular cell segmentation meshes used in the MICrONS dataset contain cell surface reconstructions, we computed the GW distance based on the pairwise Euclidean distances between 50 points sampled from each neuron, instead of geodesic distance.

To evaluate the integration, we matched some of the manually annotated cells from the MICrONS dataset with t-types from the Patch-seq datasets. Following the results of Tasic *et al.*[18], we assigned the *Sst Calb2/Chrna2* t-types (*Sst Calb2 Pdlim5, Sst Calb2 Necab1, Sst Chrna2 Ptgdr, Sst Chrna2 Glra3*) to Martinotti cells, and the *Pvalb Vipr2* t-type to chandelier cells. Some other *Pvalb* t-types were assigned to basket cells, such as *Pvalb Sema3e Kank4* and *Pvalb Gpr149 Islr,* whereas CCK or small basket cells were associated with *Vip* t-types such as *Vip Chat Htr1f* and *Vip Col15a1* [56]. Since cells of the *Vip* subclass have bipolar morphologies[18], we assigned all other *Vip* subtypes to bipolar cells. We then evaluated the consistency of the cell morphology space with these assignments by using a Wilcoxon rank-sum test to compare the distribution of GW distances between matching types across datasets with the distribution GW distances between non-matching types across datasets.

Code availability

The source code of *CAJAL* is available at https://github.com/CamaraLab/CAJAL.

Data availability

All the datasets used in this study are publicly available. The morphological reconstructions of the DVB neuron have been deposited in the neuromorpho.org database (Hart archive). The patch-clamp data of Gouwens *et al.*[20] can be accessed at the Allen Brain Atlas data portal (http://celltypes.brain-map.org/data). The Patch-seq datasets of Gouwens *et al.*[19] and Scala *et*

al.[21] can be accessed at the Brain Image Library (BIL) using the URLs https://download.brainimagelibrary.org/biccn/zeng/pseq/morph/200526/ and https://download.brainimagelibrary.org/biccn/zeng/tolias/pseq/morph/, respectively. The two-photon microscopy data of Medyukhina *et al.*[29] can be accessed at https://asbdata.hki-jena.de/publidata/MedyukhinaEtAL_SPHARM/. The MICrONS program dataset can be accessed using the MICrONS Explorer (https://www.microns-explorer.org/cortical-mm3#segmentation-meshes).

Acknowledgements

The authors are grateful to Dr. Zhaolan Zhou for his constructive comments on the manuscript, Matthew Jozwik for assisting with the conversion of files for the DVB analysis, and Jiazhen Rong and Alice Wang for collaboration on related aspects.

Author contributions

K.G. implemented *CAJAL* and performed all the computational analyses of this work. J.C. performed preliminary computational analyses for this work. A.B.S. implemented the covariate analysis of the Laplacian score. K.Z. and M.P.H. generated the morphological reconstructions of the DVB neuron and assisted with their analysis. P.G.C. and K.G. jointly wrote the manuscript. P.G.C. supervised the work.

Declaration of competing interests

The authors declare no competing interests.

Figures

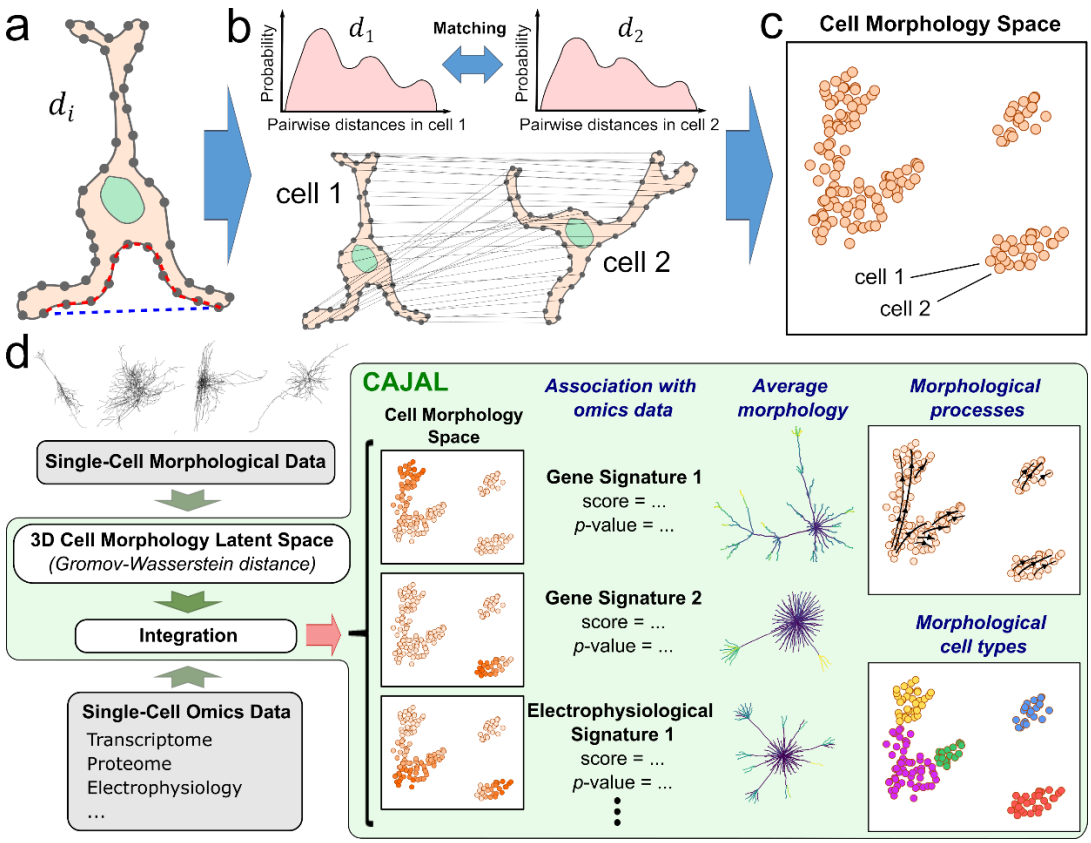

Figure 1. A general framework for the quantitative analysis and integration of cell morphology data based on metric geometry. a) *CAJAL* takes as input the 2D or 3D cell segmentation masks or traces of a set of cells. For each cell, a set of points is evenly sampled from the outline and their pairwise distance matrix d_i is computed. The Euclidean and geodesic distances between 2 sampled points is indicated with a blue and red dashed line, respectively. Different metrics for computing d_i capture different aspects of cell morphology. **b)** An optimal matching between the discretized morphologies of each pair of cells is established by computing the GW distance between their corresponding d_i matrices. Computationally efficient approximations to the GW distance use optimal transport theory to establish a map between the distributions of sampled points pairwise distances in each cell. The value of the cost function at the minimum measures the amount of deformation that is needed to convert the shape of one cell in that of another. **c)** The GW distance matrix can be thought of as a distance in a latent

space of cell morphologies. This enables the application of statistical and machine learning methods for the analysis and integration of point clouds. **d)** Overview of the open-source software *CAJAL*. The software takes single-cell morphological data in the form of segmentation masks or neuronal traces as input and enables its integration with other single-cell data modalities, the visualization and clustering of cell morphology spaces, the identification of molecular and electrophysiological features associated with changes in cell morphology, the computation of average or representative cell shapes, and the visualization of trajectories in the cell morphology space.

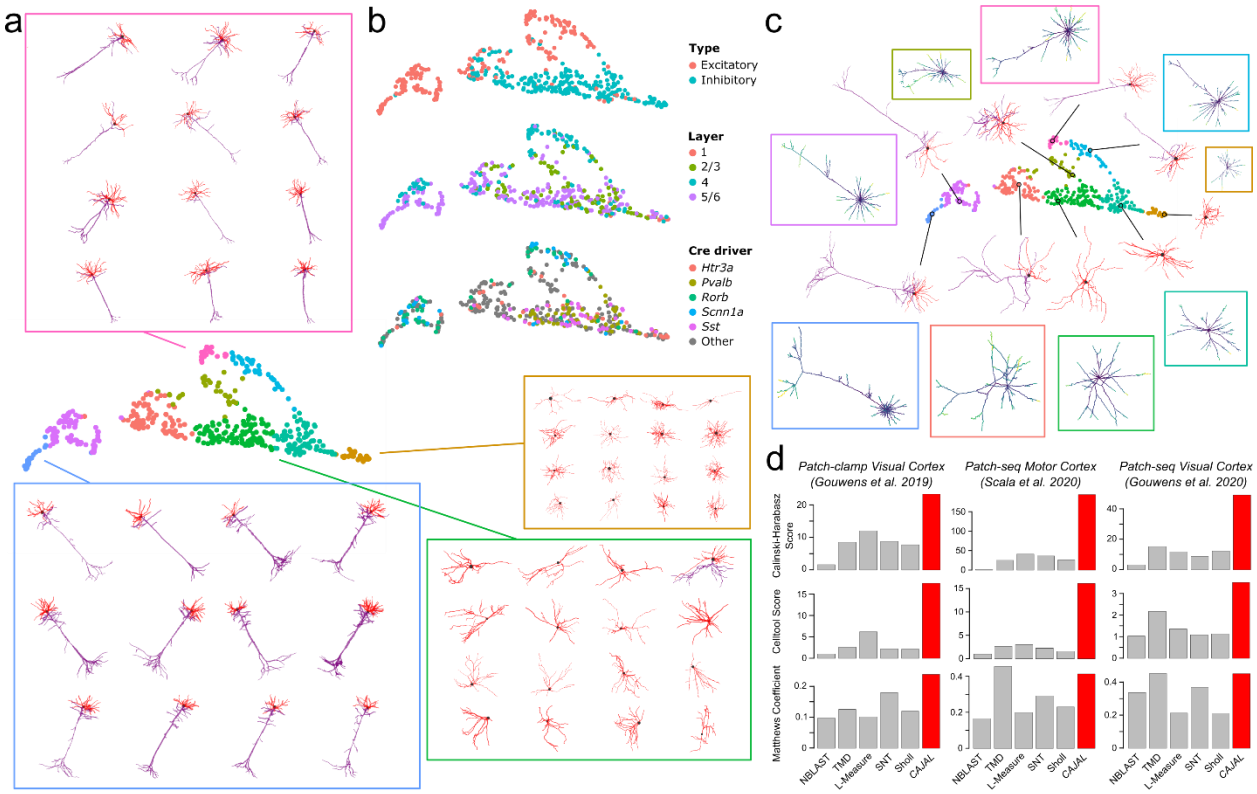

Figure 2. Cell morphology spaces accurately summarize the complexity of cell shapes. a) UMAP representation of the cell morphology space of 506 neurons from the murine visual cortex profiled with whole-cell patch-clamp. The representation is colored by the morphological cell populations that resulted from clustering the cell morphology space using Louvain community detection. The morphologies of individual neurons sampled from 4 of the populations are shown for reference. Apical and basal dendrites are indicated in purple and red, respectively. **b)** The UMAP representation is colored by the neuronal type (top), cortical layer (middle), and Cre driver line (bottom). The GW cell morphology space captures morphological differences between neurons of different molecular type and anatomic location. **c)** The metric structure of the GW morphology space enables performing algebraic operations such as averaging shapes. The figure shows the medoid (indicated with a circle) and average morphologies (in boxes) computed for each of the morphological populations in (a). **d)** The ability of *CAJAL* to identify morphological differences between molecularly defined neurons is

assessed in 3 datasets in comparison to 5 currently available methods. In this study, *CAJAL* offered substantially better or similar results compared to the best performing method in each dataset according to the three metrics that were used for the evaluation.

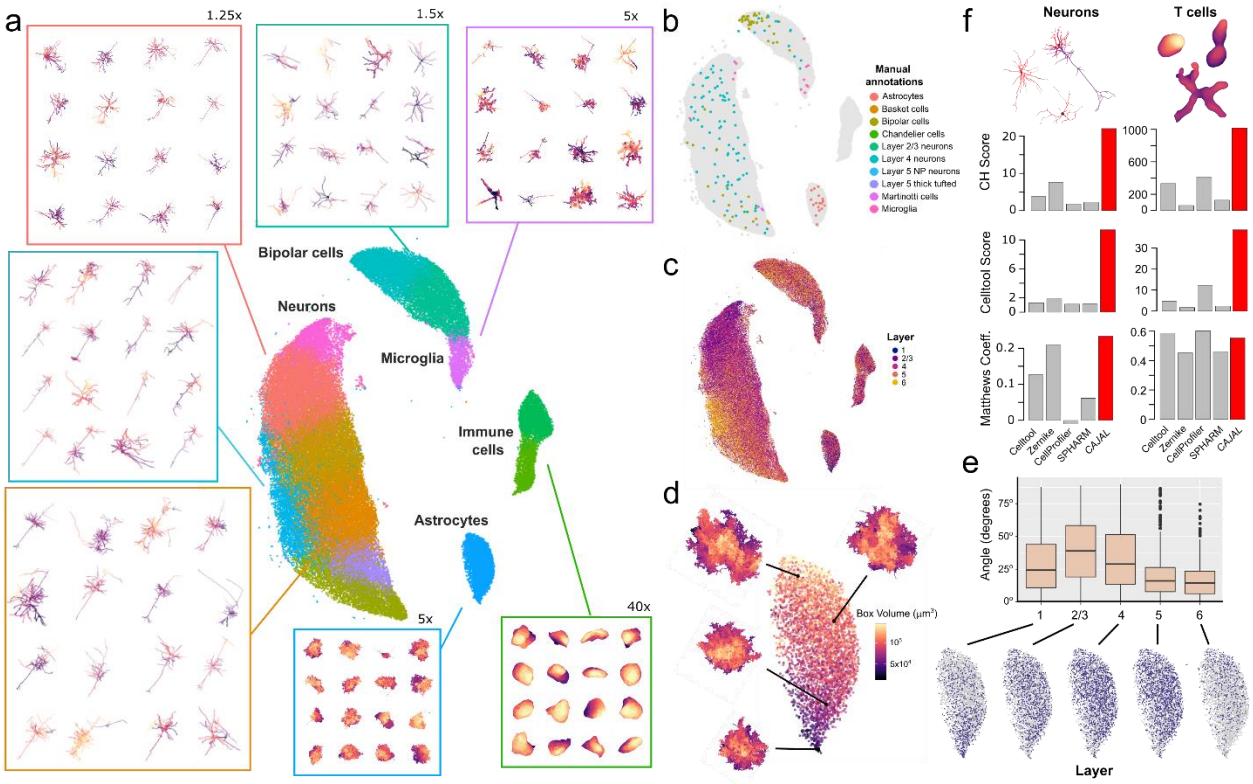

Figure 3. GW cell morphology spaces summarize cell shapes across heterogeneous cell types. a) UMAP representation of the cell morphology space of 70,510 cells from a cubic millimeter volume of the mouse visual cortex profiled by the MICrONS program using a combination of two-photon microscopy, microtomography, and serial electron microscopy [15]. The representation is colored by the cell populations identified by clustering of the cell morphology space. The morphology of randomly sampled cells from several populations is also show for reference. The magnification is indicated in cases where the morphology of the cells has been zoomed in to facilitate visualization. The cell morphology space correctly recapitulates the diversity of cell types present in the visual cortex and morphological diversity within each cell type. **b)** The position of 185 cells that were manually annotated by the MICRoNS program is indicated in the UMAP representation, showing consistency with the structure of the cell morphology space. **c)** The UMAP representation of the cell morphology space is colored by the cortical layer of each cell. The morphology space recapitulates morphological differences

between neurons and astrocytes from different cortical layers. **d)** The part of the UMAP representation corresponding to astrocytes is colored by the volume of the minimum-size box containing the astrocyte. Astrocytes in the lowe part of the UMAP have smaller dimensions than at the top. For reference, the morphology of 4 astrocytes from different parts of the UMAP is also shown. **e)** Boxplot summarizing the distribution of the angle of the major axis of astrocytes from different layers with respect to the pial surface. Astrocytes from layer 2/3 are elongated perpendicularly with respect to the pial surface. For reference, the part of the UMAP representation corresponding to astrocytes colored by the cortical layer each also shown. **f)** The ability of *CAJAL* to identify morphological differences between neurons of different molecular type and T cells from different anatomical locations is evaluated in comparison to four general algorithms for cell morphometry and according to three different metrics of performance. Contrary to state-of-the-art methods for cell morphometry, the geometric approach implemented in *CAJAL* offered accurate results for both T cells and neurons.

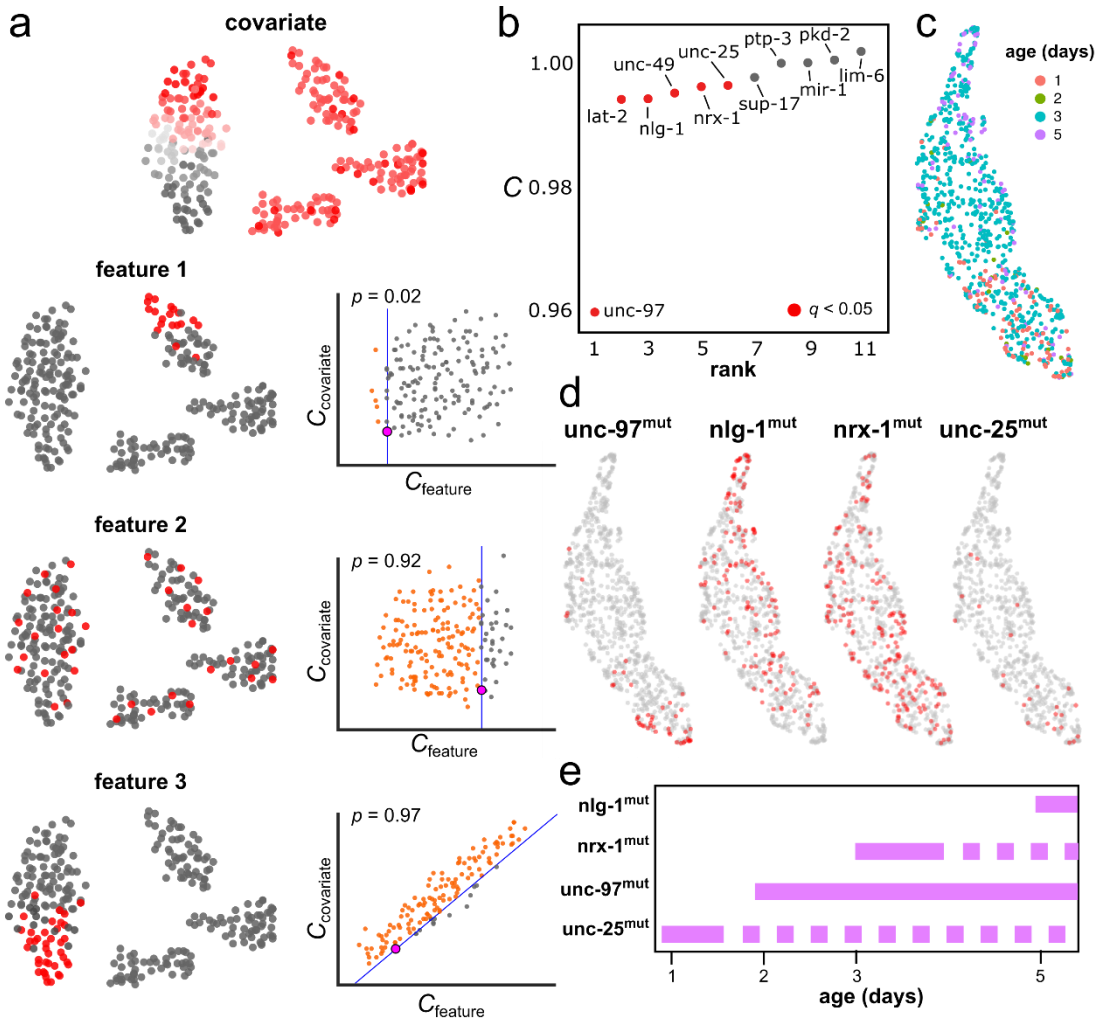

Figure 4. Identification of mutations that have an impact in the morphology of an individual neuron. a) Schematic of approach for identifying features (gene expression, mutations, protein expression, etc.) associated with cell morphological changes based on multi-modal data. Features and covariates can take binary (e.g. mutations) or continuous (e.g. gene expression) values. For each feature, the degree of consistency between the feature values and the structure of the cell morphology space is quantified using the Laplacian score (*C*) of the feature in this space. Features with a low score are associated with local regions of the cell morphology space. The statistical significance of each feature in relation to the covariates is evaluated by means of a permutation test, where cell labels are reshuffled. In the figure, examples of features that are significantly localized in the cell morphology space (feature 1, a

small number of random configurations have a smaller value of $C_{feature}$, independently of the value of $C_{covariate}$), not significantly localized in the cell morphology space (feature 2, a large number of random configurations have a smaller value of $C_{feature}$), and substantially localized in the morphology space but in association with the covariate (feature 3, a small number of random configurations have smaller value of $C_{feature}$, but they are not independent on the value of $C_{covariate}$), are presented. **b)** Mutations that have an impact on the morphology of the DVB interneuron in *C. elegans*. Null alleles are ranked according to their Laplacian score (*C*) in the cell morphology space of the DVB interneuron. Samples were imaged across 4 time points and the age of the worm used as a covariate. Genes that significantly impact the morphology of the DVB interneuron according to this approach are indicated in red (FDR < 0.05). **c, d)** UMAP visualization of the cell morphology space of the DVB interneuron colored by the the age of each worm (c) and the mutation status of *unc-97, nlg-1*, *nrx-1*, and *unc-25* (red: mutated; gray: wild-type). **e)** Restricting the analysis to worms of the same age allows us to identify the age of onset of the morphological effects induced by each significant mutation (FDR < 0.05). This analysis shows that *unc-97* and *unc-25* mutations have an earlier onset in morphology than *nlg-1* and *nrx-1* mutations. Dashed lines indicate time points with limited data availability.

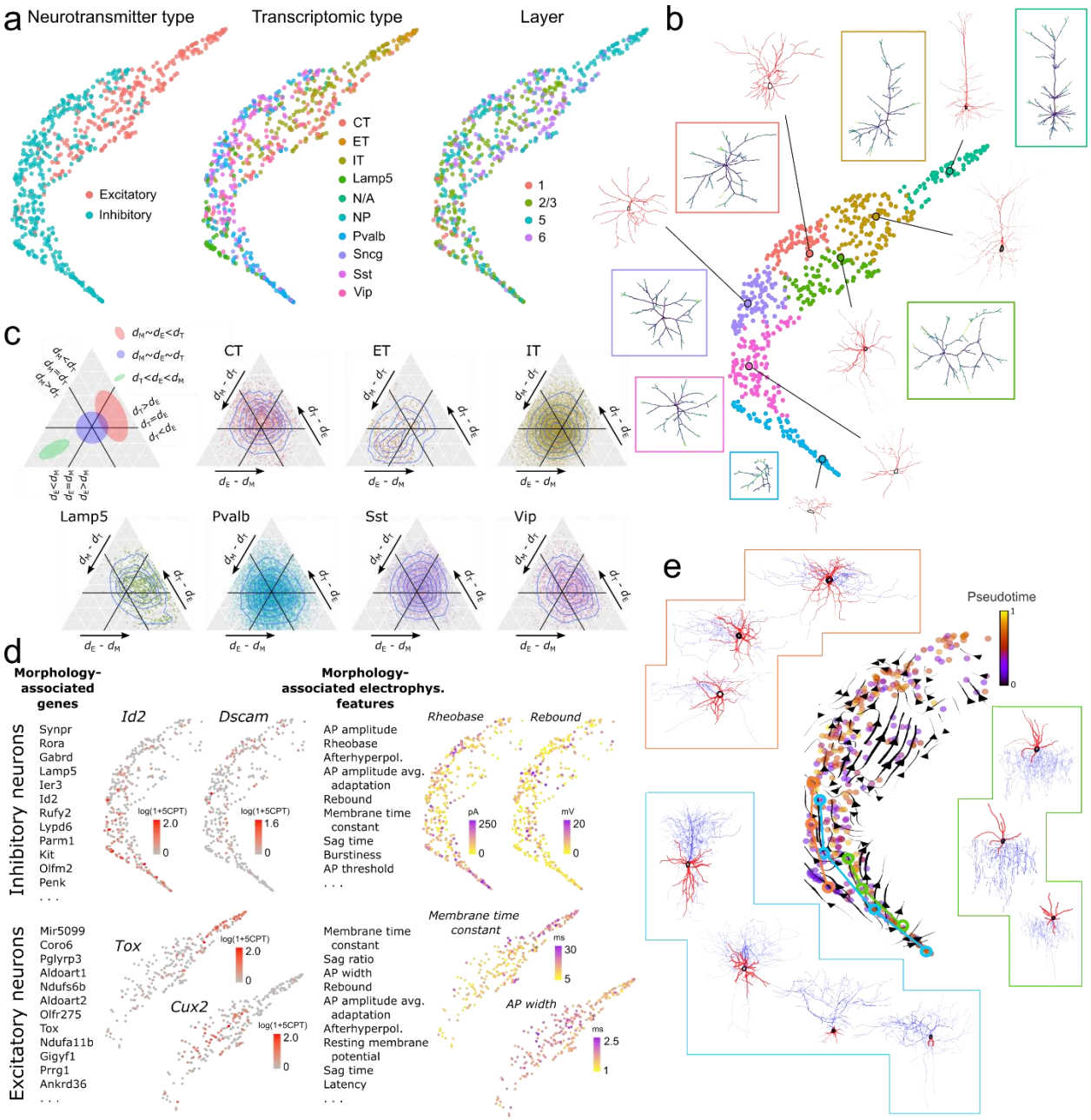

Figure 5. Integrative analysis of molecular, physiological, and morphological data of mouse motor cortex neurons. a) UMAP representation of the GW cell morphology space of the dendrites of 370 inhibitory neurons and 274 excitatory neurons from the mouse motor cortex profiled with Patch-seq by Scala et al. [21]. The representation is colored by the neurotransmitter type (excitatory/inhibitory), the transcriptomic type, and the cortical layer of the cells, showing a large degree of localization of molecular and physiological features on the morphological space.

b) UMAP representation colored by the morphological cell populations defined by Louvain clustering. The medoid and average cell morphology (in boxes) are shown for each cell population. **c)** Ternary plots showing the discrepancy between pairwise distances between cells in the morphology (M), transcriptomic (T), and electrophysiology (E) latent spaces for each transcriptomically-defined population. The dendrites of ET excitatory neurons present a large degree of variability in their morphology which is not paralleled by consistent changes in their gene expression profile, whereas the dendrites of *Lamp5*$^{+}$ GABAergic neurons present limited morphological variability in comparison to their gene expression profile. CT: corticothalamic neurons, ET: extratelecephalic neurons, IT: intratelenchephalic neurons. **d)** Top genes and electrophysiological features that are significantly associated with the morphological diversity of excitatory and inhibitory neurons according to their Laplacian score in the cell morphology space (FDR < 0.1). The part of the UMAP corresponding to excitatory or inhibitory neurons is colored by the expression level and values of some of the significant genes and electrophysiological features, respectively. CPT: counts per thousand. **e)** Morpho-transcriptomic trajectories computed by projecting the RNA velocity field in the cell morphology space. The morphology of several chandelier, basket, and *Lamp5*$^{+}$ neurons along the trajectories is shown for reference.

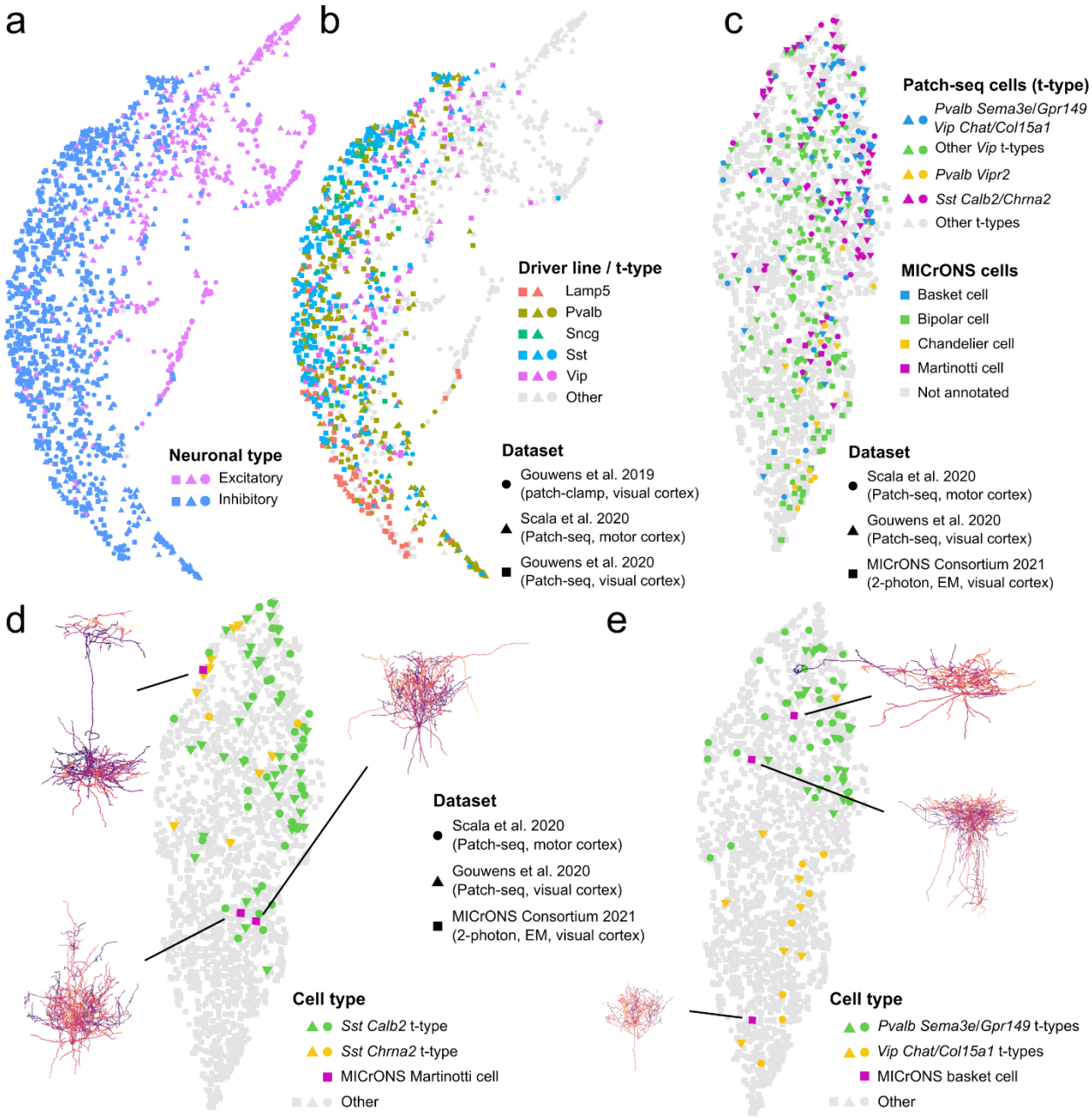

Figure 6. Integration of cell morphology data across experiments and technologies. a) UMAP representation of the combined cell morphology space of the basal and apical dendrites of visual cortex neurons profiled with patch-clamp[20], and visual cortex and motor cortex neurons profiled with Patch-seq[19, 21]. The combined dataset consists of 1,662 neurons. The representation is colored by the neuronal type. Inhibitory and excitatory neurons from different datasets cluster together in separated regions of the combined morphology space. **b)** The same

UMAP representation is colored by the Cre driver line (for cells from the patch-clamp dataset) or the t-type (for cells from the Patch-seq datasets). Cells of same t-type but from different Patch-seq dataset, and cells from the corresponding Cre driver line in the patch-clamp dataset, localize in the same regions of the combined morphology space. **c)** UMAP representation of the combined cell morphology space of 883 full neuron reconstructions from the motor and visual cortices profiled with Patch-seq[19, 21] and 1,139 neurons from the mouse visual cortex with a combination of two-photon microscopy, microtomography, and serial electron microscopy [15], 139 of which have been manually annotated by the MICrONS program. The manually annotated neuronal types from the MICrONS dataset localize in the same regions of the morphology space than Patch-seq cells from the corresponding t-type. **d)** Refined annotation of 3 Martinotti cells that were manually annotated by the MICrONS program. One of the Martinotti cells presents a distinct morphology with a densely arborized axon and is close in the cell morphology space to Patch-seq cells of the *Sst Chrna2* t-type, while the other two Martinotti cells are closer to *Sst Calb2* t-type cells. **e)** Refined annotation of 3 basket cells that were manually annotated by the MICrONS program. One basket cell has a more condensed morphology than the others is close in the morphology space to Path-seq cells of the *Vip Chat Htr1f* and *Vip Col15a1 Pde1a* t-types, while the other two larger basket cells are close to *Pvalb Sema3e Kank4* and *Pvalb Gpr149 Islr* t-type cells.

References

1. Ramón y Cajal, S. Studies on vertebrate neurogenesis. *Guth, trans., Thomas, Springfield, IL* (1960).
2. Pauling, L., Itano, H.A., Singer, S.J. & Wells, I.C. Sickle cell anemia, a molecular disease. *Science* **110**, 543-548 (1949).
3. Wessells, N. et al. Microfilaments in cellular and developmental processes. *Science* **171**, 135-143 (1971).
4. Ford, J. Red blood cell morphology. *Int J Lab Hematol* **35**, 351-357 (2013).
5. Boutros, M., Heigwer, F. & Laufer, C. Microscopy-Based High-Content Screening. *Cell* **163**, 1314-1325 (2015).
6. Mattiazzi Usaj, M. et al. High-Content Screening for Quantitative Cell Biology. *Trends Cell Biol* **26**, 598-611 (2016).
7. Ascoli, G.A., Donohue, D.E. & Halavi, M. NeuroMorpho.Org: a central resource for neuronal morphologies. *J Neurosci* **27**, 9247-9251 (2007).
8. Bardy, C. et al. Predicting the functional states of human iPSC-derived neurons with single-cell RNA-seq and electrophysiology. *Mol Psychiatry* **21**, 1573-1588 (2016).
9. Cadwell, C.R. et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat Biotechnol* **34**, 199-203 (2016).
10. Chen, X. et al. Coupled electrophysiological recording and single cell transcriptome analyses revealed molecular mechanisms underlying neuronal maturation. *Protein Cell* **7**, 175-186 (2016).
11. Foldy, C. et al. Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E5222-5231 (2016).
12. Fuzik, J. et al. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat Biotechnol* **34**, 175-183 (2016).
13. Lipovsek, M. et al. Patch-seq: Past, Present, and Future. *J Neurosci* **41**, 937-946 (2021).
14. Winnubst, J. et al. Reconstruction of 1,000 Projection Neurons Reveals New Cell Types and Organization of Long-Range Connectivity in the Mouse Brain. *Cell* **179**, 268-281 e213 (2019).
15. Consortium, M. et al. Functional connectomics spanning multiple areas of mouse visual cortex. *bioRxiv*, 2021.2007.2028.454025 (2021).
16. Jiang, X. et al. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science* **350**, aac9462 (2015).
17. Markram, H. et al. Reconstruction and Simulation of Neocortical Microcircuitry. *Cell* **163**, 456-492 (2015).
18. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78 (2018).
19. Gouwens, N.W. et al. Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells. *Cell* **183**, 935-953 e919 (2020).
20. Gouwens, N.W. et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat Neurosci* **22**, 1182-1195 (2019).
21. Scala, F. et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature* **598**, 144-150 (2021).
22. Carpenter, A.E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* **7**, R100 (2006).

23. Legland, D., Arganda-Carreras, I. & Andrey, P. MorphoLibJ: integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics* **32**, 3532-3534 (2016).
24. Sholl, D.A. Dendritic organization in the neurons of the visual and motor cortices of the cat. *J Anat* **87**, 387-406 (1953).
25. 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).
26. Arshadi, C., Gunther, U., Eddison, M., Harrington, K.I.S. & Ferreira, T.A. SNT: a unifying toolbox for quantification of neuronal anatomy. *Nat Methods* **18**, 374-377 (2021).
27. Costa, M., Manton, J.D., Ostrovsky, A.D., Prohaska, S. & Jefferis, G.S. NBLAST: Rapid, Sensitive Comparison of Neuronal Structure and Construction of Neuron Family Databases. *Neuron* **91**, 293-311 (2016).
28. Wan, Y. et al. BlastNeuron for Automated Comparison, Retrieval and Clustering of 3D Neuron Morphologies. *Neuroinformatics* **13**, 487-499 (2015).
29. Medyukhina, A. et al. Dynamic spherical harmonics approach for shape classification of migrating cells. *Sci Rep* **10**, 6072 (2020).
30. Pincus, Z. & Theriot, J.A. Comparison of quantitative methods for cell-shape analysis. *J Microsc* **227**, 140-156 (2007).
31. Khotanzad, A. & Hong, Y.H. Invariant image recognition by Zernike moments. *IEEE Transactions on pattern analysis and machine intelligence* **12**, 489-497 (1990).
32. Brechbühler, C., Gerig, G. & Kübler, O. Parametrization of closed surfaces for 3-D shape description. *Computer vision and image understanding* **61**, 154-170 (1995).
33. Kanari, L. et al. A Topological Representation of Branching Neuronal Morphologies. *Neuroinformatics* **16**, 3-13 (2018).
34. Li, Y., Wang, D., Ascoli, G.A., Mitra, P. & Wang, Y. Metrics for comparing neuronal tree shapes based on persistent homology. *PLoS One* **12**, e0182184 (2017).
35. Mémoli, F. On the use of Gromov-Hausdorff distances for shape comparison. (2007).
36. Mémoli, F. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* **11**, 417-487 (2011).
37. Mémoli, F. & Sapiro, G. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics* **5**, 313-347 (2005).
38. Gromov, M. Groups of polynomial growth and expanding maps (with an appendix by Jacques Tits). *Publications Mathématiques de l'IHÉS* **53**, 53-78 (1981).
39. Edwards, D.A. in Studies in topology 121-133 (Elsevier, 1975).
40. Scetbon, M., Cuturi, M. & Peyré, G. in International Conference on Machine Learning 9344-9354 (PMLR, 2021).
41. Solomon, J., Peyré, G., Kim, V.G. & Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)* **35**, 1-13 (2016).
42. Chowdhury, S., Miller, D. & Needham, T. in Joint European Conference on Machine Learning and Knowledge Discovery in Databases 811-827 (Springer, 2021).
43. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

44. Lanjakornsiripan, D. et al. Layer-specific morphological and molecular differences in neocortical astrocytes and their dependence on neuronal layers. *Nature communications* **9**, 1623 (2018).
45. McQuin, C. et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol* **16**, e2005970 (2018).
46. Wu, P.H. et al. Evolution of cellular morpho-phenotypes in cancer metastasis. *Sci Rep* **5**, 18437 (2015).
47. Govek, K.W., Yamajala, V.S. & Camara, P.G. Clustering-independent analysis of genomic data using spectral simplicial theory. *PLoS computational biology* **15**, e1007509 (2019).
48. He, X., Cai, D. & Niyogi, P. in Advances in neural information processing systems 507-514 (2006).
49. Hart, M.P. & Hobert, O. Neurexin controls plasticity of a mature, sexually dimorphic neuron. *Nature* **553**, 165-170 (2018).
50. LeBoeuf, B. & Garcia, L.R. Caenorhabditis elegans Male Copulation Circuitry Incorporates Sex-Shared Defecation Components To Promote Intromission and Sperm Transfer. *G3 (Bethesda)* **7**, 647-662 (2017).
51. Fuerst, P.G., Koizumi, A., Masland, R.H. & Burgess, R.W. Neurite arborization and mosaic spacing in the mouse retina require DSCAM. *Nature* **451**, 470-474 (2008).
52. Wang, Y. et al. PCDH7 interacts with GluN1 and regulates dendritic spine morphology and synaptic function. *Sci Rep* **10**, 10951 (2020).
53. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
54. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* (2020).
55. Hilscher, M.M., Leao, R.N., Edwards, S.J., Leao, K.E. & Kullander, K. Chrna2-Martinotti Cells Synchronize Layer 5 Type A Pyramidal Cells via Rebound Excitation. *PLoS Biol* **15**, e2001392 (2017).
56. Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C.Z. & Markram, H. Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cereb Cortex* **12**, 395-410 (2002).
57. Biasotti, S., Cerri, A., Bronstein, A. & Bronstein, M. Recent Trends, Applications, and Perspectives in 3D Shape Similarity Assessment. *Computer Graphics Forum* **35**, 87-119 (2016).
58. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018).
59. Hagberg, A., Swart, P. & S Chult, D. (Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008).
60. Flamary, R. et al. Pot: Python optimal transport. *Journal of Machine Learning Research* **22**, 1-8 (2021).
61. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1-9 (2006).
62. Harris, C.R. et al. Array programming with NumPy. *Nature* **585**, 357-362 (2020).
63. Crane, K., Weischedel, C. & Wardetzky, M. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)* **32**, 1-11 (2013).

64. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
65. Leland, M., John, H., Nathaniel, S. & Lukas, G. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018).
66. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* **2**, 205 (2017).
67. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).
68. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research* **35**, W193-200 (2007).
69. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).