Data-driven fine-grained region discovery in the mouse brain with transformers

Alex J. Lee[1,2], Shenqin Yao[3], Nicholas Lusk[3], Lydia Ng[3], Michael Kunst[3], Hongkui Zeng[3], Bosiljka Tasic[3], Reza Abbasi-Asl[1,2,*]

[1]University of California, San Francisco
[2]Weill Institute for Neurosciences
[3]Allen Institute for Brain Science

*Corresponding author: Reza Abbasi-Asl (Reza.AbbasiAsl@ucsf.edu)

Abstract

Technologies such as spatial transcriptomics offer unique opportunities to define the spatial organization of the mouse brain. We developed an unsupervised training scheme and novel transformer-based deep learning architecture to detect spatial domains across the whole mouse brain using spatial transcriptomics data. Our model learns local representations of molecular and cellular statistical patterns which can be clustered to identify spatial domains within the brain from coarse to fine-grained. Discovered domains are spatially regular, even with several hundreds of spatial clusters. They are also consistent with existing anatomical ontologies such as the Allen Mouse Brain Common Coordinate Framework version 3 (CCFv3) and can be visually interpreted at the cell type or transcript level. We demonstrate our method can be used to identify previously uncatalogued subregions, such as in the midbrain, where we uncover gradients of inhibitory neuron complexity and abundance. Notably, these subregions cannot be discovered using other methods. We apply our method to a separate multi-animal whole-brain spatial transcriptomic dataset and show that our method can also robustly integrate spatial domains across animals.

Introduction

Hierarchical spatial organization is ubiquitous in tissue and organ biology. Systematic, high-dimensional phenotypic measurements of this organization, generated through experimental tools such as spatial transcriptomics, multiplex immunofluorescence, and electron microscopy, are also becoming increasingly available as large, open datasets[1–3]. However, transforming this abundance of data into information and hypotheses can be difficult, even for fields with a wealth of prior knowledge, such as neuroanatomy.

The creation of open data resources such as the Allen Brain Cell Mouse Whole Brain (ABC-MWB) Atlas[1], a multi-million cell spatial and single-cell RNA sequencing (scRNA-seq) atlas, provides a new opportunity to investigate whether computational tools can help biologists understand spatial cellular and molecular organization. However, the size of these datasets presents computational challenges for existing methods, many of which do not extend readily to a multi-section or multi-animal scenario. One reason for this is because existing methods for spatial niche or spatial domain detection often operate on the entire dataset at once, for example an entire cell by gene matrix for a tissue section. This poses significant obstacles for deep-learning based methods requiring GPUs, as most systems do not have the

GPU memory required to load multiple sections of data or compute intermediary representations such as pairwise distance matrices[4]. A significant class of methods relies on Gaussian processes, which feature a costly cubic computational scaling in the number of observations[5]. Our method, CellTransformer, addresses this concern by focusing on a spatially constrained representation of cells in space. We show this representation is sufficient to recover important anatomically relevant structures in the data while remaining computationally efficient.

CellTransformer is a transformer[6], an established architectural tool in deep representation learning, which we use to compute a latent variable representation of cellular neighborhoods and their molecular features. In order to train this network without predefined region labels, we introduce a novel self-supervised training objective reminiscent of graph masked autoencoding[7,8]. During training time, the network is optimized to predict masked data for cells in the same neighborhood using the neighborhood representation, which is a compressed vector representation of the molecular and cellular characteristics of that cellular neighborhood. We interpret this objective function as quantifying the extent of spatial correlation between a cell and its cellular neighborhood.

We evaluate CellTransformer on the ABC-MWB dataset, demonstrating its effectiveness in producing completely data-driven spatial domains of the mouse brain by comparing to the Allen Mouse Brain Common Coordinate Framework version 3 (CCFv3[9]), a consensus hand-drawn 3D reference space compiled from a large multimodal data corpus. At coarse scale (meaning a small number of spatial clusters) CellTransformer regions are similar to CCFv3 in terms of cell type composition; at fine scale, we uncover previously uncharacterized subregions with unique cell type and gene expression patterns. CellTransformer is scalable to spatial transcriptomics datasets with millions of cells and also allows for integration of regions across animals. Importantly, we show that other methods are unable to resolve fine grained and spatially coherent spatial domains that CellTransformer can, demonstrating the practical utility of our method for accurate and scalable whole-organ spatial domain detection.

Results

Data-driven discovery of fine-grained spatial domains in the mouse brain spatial transcriptomics data with CellTransformer

CellTransformer is a graph transformer[10] neural network that models spatial cellular neighborhoods. We define a cellular neighborhood as any cells within a user-specified distance cutoff in microns away from a reference or center cell. Our model takes the gene expression (here, MERFISH probe counts) and imputed cell type identity from the ABC-MWB scRNA-seq taxonomy for each cell and outputs a latent variable representation for each of these neighborhoods across the brain. One of the principal operations in a transformer is the self-attention operation, which computes a feature update based on pairwise interactions between elements in a sequence, which are referred to as tokens (here, cells). Accordingly, one interpretation of our model is of learning an arbitrary and dynamic pairwise interaction graph among cells.

Restricting this graph to a small subgraph of the whole-tissue-section graph has benefits for both computational resource usage and biological interpretability. We interpret the size of the neighborhood as a constraint on the physical distance at which statistical correlations between the observed cells and their gene expression profiles can be directly captured. A secondary aspect is that truncating neighborhoods using a fixed spatial threshold allows the network to account for the differential density of cells in space. In this way, our framework incorporates a notion of both cytoarchitecture (relative density and proximity) and molecular variation (cell type and transcript-level variation) in the data.

We designed a self-supervised training scheme to induce our graph transformer to learn biologically relevant latent features in the data without spatial region labels. In this scheme, the model learns to compute latent variables of cellular neighborhoods such that they are useful to predict characteristics (observed MERFISH probe counts) of masked cells in the neighborhood. We train the model to use only information about the cell's neighbors to predict the gene expression profile of the reference cell (**Figure 1a**). Specifically, we first compute latent features for every cell in the neighborhood without the reference cell. The features for these cells interact and iteratively update based on interactions between cells, which are parameterized via the self-attention operator. We then aggregate these features using attention pooling (see **Methods**), compressing features for cells across the network into a single, fixed, vector latent variable representation. This vector is then concatenated to a token which represents the cell type of the reference cell and is used to compute a final prediction of the observed gene expression counts of the reference cell (**Figure 1b**). At test time, we extract the neighborhood representation for each cell and use *k*-means to identify discrete spatial domains. Decoupling the feature learning and spatial domain clustering steps into components that can be computed in a batched fashion allows us to scale to very large datasets. We also optionally introduce a smoothing step prior to *k*-means, which we applied (see **Supplementary Note 1**) to spatially smooth the embeddings, which we applied at 40 micron full-width at half-maxima (sigma of 12 micron). See **Supplementary Note 1** for a discussion on the effects of smoothing on detected domains.

The ABC-MWB spatial transcriptomics dataset is divided into two parts. Tissue from five animals was collected, with one animal processed by the Allen Institute for Brain Science with a 500 gene MERFISH panel and the others generated by the Zhuang lab (Zhang et al. (2023)[11]) with a 1149 gene panel. We initially trained CellTransformer on the Allen Institute for Brain Science mouse in the ABC-MWB dataset using cell type labels at the subclass level of the ABC-MWB cell type ontology, subsequently extracting embeddings for each neighborhood in the dataset and clustering using *k*-means.

We found a high spatial consistency of domains across tissue sections and visual correspondence with CCFv3 (**Figure 2a**) using the preliminary registration available in the October 2023 data release. We provide representative images of spatial clusters across the brain at different *k* in **Supplementary Figures 1-3**. We observed that increasing the number of spatial domains *k* yields subregions of higher-level regions in a neuroanatomically consistent manner. For example, at *k*=25, we observed high level structures such as the striatum and gross cortical layers (layer 1, layers 2 and 3, and layers 4 and 5 are grouped together) in anterior cingulate (AC), motor cortex (MO) and somatosensory (SS), shown in different shades of green in the CCFv3 reference (**Figure 2a**). At *k*=300 we observed spatial clusters for subregions of the striatum, such as nucleus accumbens and olfactory tubercle, as well as finer layers across cortex, where CellTransformer is able to identify per-region cortical layers across cortex, such as in MO and SS, as well as in AC and insula (**Figure 2b**). At *k*=500, we observed subregions in areas such as pallidum, where substructures such as the bed nucleus (BST) become visible. We also observed several subregions tiling the dorsal-ventral axis of caudoputamen (**Figure 2b)**, and observed anterior-posterior divisions within domains, for example in layers 1-3 of anterior cingulate, which are grouped into an anterior segment (sections 50 and 49) and a posterior segment (sections 48 and 47). We also observed fine-grained domains in the subcortex, such as the Islands of Calleja (**Supplementary Figure 4a**) and subregions of the caudoputamen (**Supplementary Figure 4b**).

To quantify the observation of high-level similarity with CCF, we selected three levels of granularity in CCF annotations: division (25 regions), structure (354 regions), and substructure (670 regions). We also computed corresponding regions using our method and CellCharter[12,] another deep learning model, with the same number of spatial domains for comparison. We could not run GPU-accelerated methods including STACI[13] and spaGCN[4] on our system, which used a NVIDIA A6000 GPU with 48GB RAM, on the entire dataset. We could not run scENVI[14] due to an out-of-(system)-memory error on our machine and on the largest nodes of the UCSF Wynton HPC cluster we could access. For

each spatial domain (from CCF or from one of the methods), we computed a cell type composition vector using the ABC-MWB taxonomy and used the similarity between these vectors as an index of similarity to CCF. The ABC-MWB taxonomy is hierarchical, and we chose the subclass level (338 types) for computational expediency. We calculated the similarity between CellTransformer and CellCharter region sets with CCF by correlating the cell type composition vectors of a given pipeline's region with its best match in CCF, showingCellTransformer domains are more similar than CellCharter (**Figure 2c**). Particularly at higher numbers of regions, both CellTransformer and CellCharter domains are very similar to at least one CCF region. However, CellTransformer performs better particularly at large numbers of regions. We also computed the average correlation when the comparison is mutually exclusive (multiple spatial domains cannot match with the same CCF region, **Figure 2d**), showing that similarity of spatial domains to CCF are not merely driven by redundant matches of similar regions. Again, CellTransformer performs relatively better at high numbers of regions. Finally we developed an index of spatial consistency of domains from CCF, CellTransformer, and CellCharter (**Figure 2e**) using a nearest-neighbor approach (see **Methods**), demonstrating CellTransformer produces much more spatially discrete domains than CellCharter. Visualization of spatial clusters from CellCharter (**Supplementary Figure 5-6**) shows that even at *k*=670, regions are much less spatially discrete in the midbrain. Particularly in midbrain and hindbrain, there are few discrete regions, demonstrating the practical utility of CellTransformer in uncovering structure in areas that have been difficult to anatomically characterize.

The hierarchical grouping of domains at different choices of *k* prompted us to develop a strategy to preliminarily determine an optimal number of spatial domains based on two metrics. We computed 20 clustering instances with different random initializations for a large range of choices of *k* and analyzed them using these metrics (**Supplementary Figure 7**). The first, inertia, is an overall goodness-of-fit measurement that quantifies the sum of squares errors for the points in latent space to their nearest centroid. We also used the Amari-type distance[15], which measures the stability of clusters at a given *k,* to quantify the similarity of clusters across random initializations. Based on this analysis we chose *k*=1300 which was the point of second derivative crossing for the product of the two metrics.

Mapping of spatial domains in the hippocampal formation

We first qualitatively determined the correspondence of CellTransformer spatial domains with known anatomical structure in the hippocampal formation, notably the subiculum (SUB) and prosubiculum (PS), which were investigated in Ding et al. (2020)[16]. We focused on this area because it is well characterized with respect to both connectivity[17] and transcriptomic composition[18,19]. Specifically, recent works have noted the extensive laminar organization (superficial layers to deeper layers), and dorsal-ventral organization of subiculum. This organization has been attributed to distinct and correlated patterns of gene expression and connectivity. We qualitatively compare spatial domains discovered by CellTransformer with *k*=1300 to the anatomical borders identified in Ding et al. (**Figure 3a**), where we found CellTransformer was robustly able to recover this organization.

The subiculum is known to feature a three layer organization of a molecular layer, a pyramidal cell layer, and polymorphic cell layer. **Figure 3b** shows discovered spatial domains at *k*=1300, with a subset corresponding to SUB and PS in **Figure 3c**. Both SUB and PS are clearly identified in section 32, although the polymorphic layer of PS stretches into CA1v. This is consistent with results identified in Cembrowski et al. (2018)[18], which identified a spatially discrete boundary between the polymorphic layer in PS and along the edge of CA1. We also observed a dorsal-medial polymorphic domain (dark red, PSmd-po) and a medial-ventral domain (brown, PSmv-po) component as well. We also observed the pyramidal (blue, HApy) and polymorphic layers (orange, HApo) of the hippocampal-amygdaloid transition area. In section 30, we observed the previously mentioned three layer structure of PS, and in section 28 the pyramidal and polymorphic layers of SUB.

By visualizing gene expression in the MERFISH panel within these cells (**Figure 3e**) we identify several genes which vary along either the superficial-deep or core-periphery directions. We observe that similarly to Cembrowski et al. (2018)[18], many genes expressed in SUB and PS traverse their long axis. We show the probe counts for differentially expressed genes among these areas in **Figure 3e**. Using hierarchical clustering on the MERFISH cell by gene matrix for cells in these regions, we see that the regions cluster according to spatial proximity, for example components of HA together and dorsal or ventral components of PS in the same group in the hierarchy.

The identification of spatial domains which subdivided specific layers of PS and SUB similarly to the results in Ding et al., and featured similar types of gene expression gradients as existing literature, suggests that our pipeline was successful in learning neuroanatomically useful information. Encouraged by this result, we continued our analysis in the superior colliculus.

CellTransformer allows for quantification of laminar and columnar organization in superior colliculus

Recent studies using systematic mapping of cortico-tectal fibers in superior colliculus (SC) have identified distinct laminar and columnar structure[20], suggestive of the complex role SC plays in integration of sensory information and coordination of signals. Therefore, SC presented an excellent opportunity to identify transcriptomic and cellular correlates of connectomic variation. We observed a strong correspondence of three of our spatial clusters (k=1300) with three layers of superior colliculus, particularly the zonal, superficial gray, and optic layers (**Figure 4a**) across a set of tissue sections spanning ~600 µm from anterior to posterior (**Supplementary Figure 8a**). We present a map of CCF regions registered to the section coordinates for reference. Notably, CellCharter was unable to identify these structures. At 1300 Gaussians the CellCharter clustering was heterogeneous and difficult to interpret, so we presented 670 Gaussians instead, demonstrating CellTransformer's practical utility for fine-grained domain discovery (**Supplementary Figure 6**).

By visualizing the cell type composition within the top-ten most abundant types for these three spatial domains (**Supplementary Figure 8b**), we were able to identify cell types that were highly selective for our data-driven SC layers, such as types 0873 SCsg Gabrr2 Gaba_2, 0861 SCs Pax7 Nfia Gaba_3, and 0788 SCop Sln Glut_1. We emphasize that the cell types, which have already been annotated as being associated with the zonal, superficial gray, or optic layers, are identified automatically by CellTransformer. By counting the number of unique supertype level cell types in these regions (**Supplementary Figure 8c**), we observed a gradient of inhibitory neuron complexity, where our optic layer had the most inhibitory neuron types, and the zonal layer had the least. We also found major differences in excitatory-inhibitory composition; for example, the superficial optic layer had the fewest types of inhibitory neurons, but inhibitory neurons were the largest proportion of cells in this region.

Encouraged by these findings, we also identified subregions of the intermediate gray and intermediate white areas of the motor-related areas in SC (**Figure 4a-b**), where we identify regions which are not annotated in CCF. We identified selective cell types, including 0849 SCm-PAG Cdh23 Gaba_2 (enriched in the medial intermediate white layer, shown in dark green), 0769 SCig SCiw Pitx2 Glut_3 (enriched in lateral intermediate white, shown in light blue), and 0764 SCig-an-PTT Foxb1 Glut_1 (enriched in medial intermediate gray, shown in dark blue) by visualizing the top ten most abundant types at the supertype cell type level (**Figure 4c**). We note that particularly for the subregions of superior colliculus motor-related areas, differences between spatial domains may be attributed to differences in the composition of rare cell types (**Fig 4c**) instead of majority or common cell types. There were also large differences in the proportion of glial cell types such as 1184 MOL NN-4.

We observed complex cell type abundance gradients by visualizing the supertype cell type composition and relative proportions of different neurotransmitter types (**Figure 4d-e**). Overall there is a gradient in the distribution of neurons and glia, where non-neuronal cell types are more enriched in lateral areas (such as the light and dark blue areas, which represent lateral areas of intermediate gray and intermediate white). GABA-ergic neurons are more abundant in the medial areas (brown and purple-gray). The medial subdomain of intermediate white (shown in brown) has the fewest excitatory neurons, followed by the medial subdomain of intermediate gray (shown in purple-gray).

The division of subdomains of SC into lateral and medial components was previously identified from experiments in Benavidez et al. (2021)[20], (see **Figure 4f-g** for reproduction from Benavidez et al., granted with permission). Briefly, connectivity was measured via both anterograde and retrograde tracing, revealing specific and radially organized patterns of fiber termination in SC from sensorimotor and visual areas. The strong correspondence of CellTransformer discovering wholly molecular domains which are correlated with connectivity is a significant step forward in describing the potential covariance between molecular properties and other properties of subcortical brain areas.

A medial-lateral gradient of inhibitory neurons in the midbrain reticular nucleus

Next, we investigated the midbrain reticular nucleus (MRN), a subcortical structure with few anatomical annotations in CCF. MRN is highly enriched for interneurons in a dense array of connections, CellTransformer identifies four subregions of the MRN, which are not included in the existing CCF annotation (**Figure 5a**). Plotting cell type proportions across the MRN, we identified cell types which are enriched for a given region, although all regions were predominantly glial (e.g. 1184 MOL NN_4 supertype is abundant in all regions, **Figure 5b**). Visualizing gene expression patterns across these subregions (**Figure 5c**), we identified genes which are similarly subregion selective (**Figure 5d**) and form dorsal-ventral expression gradients. We observed that the number of types of excitatory neurons varied along an inferior-superior axis, with the yellow (inferior) region having the fewest types and the purple (superior) having the most (**Figure 5e-f**). Interestingly, we also found that the two inferior regions (yellow and gray) featured Glut-Sero and Dopa neurons which were absent in the superior ones.

CellTransformer enables scaling up to multi-animal, million cell datasets

The availability of multi-animal data in ABC-MWB from the Zhuang lab led us to investigate the possibility of performing automated spatial-domain integration across animals. To investigate this, we trained CellTransformer on the Zhang et al. (2023)[1] MERFISH data. Note that we initialized new projection layers and output layers in our network to enable the modeling of a larger MERFISH gene panel in the Zhuang lab data. We computed embeddings for each neighborhood as in the previous analysis and performed *k*-means similarly, concatenating representations for all mice. This provided an opportunity to examine whether CellTransformer could adapt to a multi-animal case in addition to finding spatial domains across tissue sections of the same animal.

We observed a strong concordance between spatial domains across all mice, which we demonstrated by plotting sequential tissue sections across mice (**Figure 6a**). There is also clear anatomical correspondence between regions across coronal and sagittal mice, which can be observed by comparing animals; for example, in section 45 of mouse 2 and sections 006 and 007 of mouse 3, where

hippocampal formation is identified similarly across mice, and sections 019 (mouse 3) and section 002 (mouse4) with almost perfect correspondence of regions.

One concern with cross-animal integration is whether per-animal spatial clusters emerge. Although it is ambiguous whether a spatial cluster appears only in one animal due to sectioning or batch-related effects, we wanted to understand the relative abundance of per-animal spatial clusters. To quantify the emergence of these spatial clusters, we computed several clustering solutions with differing k and quantified how many spatial domains are not found in all four mice (**Figure 6b**). We found that at higher values of *k* the number of per-mouse spatial domains increases significantly. There is a marked increase in overall correlation relative to the Allen Institute for Brain Science dataset. Although the Allen Institute dataset is smaller (3,737,551 cells) it is all in one animal, whereas the total number of cells in the Zhuang lab datasets (5,822,737 cells) is larger but spread across animals.W e also noted that in general per-mouse domains appeared for animal two. For example, for one region set at *k*=630, 78% of per-mouse domains were found in mouse 2 (18.5% for mouse 1 and 3.7% for mouse 3; note that mouse 4, which has the fewest number of cells, did not develop per-mouse clusters at the tested parameter range).

To investigate whether animal-level differences in the embeddings might be related to overall appearance of spatially noncontiguous domains, we used two regression analyses to quantify the degree of positional and animal-specific information linearly contained in the embeddings. We regressed CCF-registered (x, y, z) coordinate position on the embeddings and separately used a logistic regression to classify animal identities. We then plotted the median absolute prediction error from the (x, y, z) coordinate regression against the predicted probabilities from the logistic regression (**Figure 6c**). Overall neighborhood level prediction of donor identity was very accurate (>90% for all animals). Median absolute prediction error was also relatively accurate, although there were differences in decoding accuracy in both metrics associated with different amounts of data points. Animal 4 has the smallest number of measured cells compared to the others (162,578 cells vs the next largest, animal 2, with 1,227,408 cells).

We then investigated whether pooling across animals affected the correspondence of CellTransformer domains with CCF ones in terms of cell type composition. We compared cell type composition estimates for two sets of spatial clusters: one where neighborhood embeddings are clustered per-animal and one where neighborhood embeddings are clustered across animals by concatenating representations. We found that pooling across clusters did not improve overall correlation with CCF compared to per-animal clustering (**Figure 6d**) although overall similarity to CCF internally to the datasets were similar to when CellTransformer was compared to CCF for the Allen Institute for Brain Science mouse. We also found that the high spatial smoothness of domains detected in the Allen Institute data is also observed in the Zhuang lab datasets (**Supplementary Figure 9**). We visualized the differences between clustering solutions with only one animal or all animals, finding they were visually similar in both cases (**Supplementary Figure 10**).

Discussion

In this study, we present a transformer-based pipeline to combine scRNA-seq and MERFISH atlases and perform whole-brain, data-driven parcellation. This representation space learned by our model can be clustered to identify progressively finer-scale regions using local cellular and molecular information alone, without predefined spatial labels. We show these regions can be interpreted at the gene or cell level. Our pipeline allows extraction of a very high number (several hundred or more) domains, which retain high correspondence to existing brain region ontology (CCF). These domains are also highly spatially consistent both within and across tissue sections. This capability is intrinsic to our model, and is learned despite any conditional modeling for donor or section-level covariates, indicating the robustness of learned features.

The block structure in our quantitative comparison of CellTransformer's spatial domains to CCF domains illustrates the potential biological relevance of our method, directly demonstrating Cell Transformer's ability to detect brain areas concordant with existing observations as well as detection of novel subregions. We interpret the length and width of the blocks as a measurement of structural complexity in a group of domains from either CCF or CellTransformer. For example, CellTransformer regions similar to isocortex regions are in compact blocks (e.g., upper left hand of **Figure 2f**). Subcortical areas such as the midbrain comprise more regions in CCF than our method at a low *k* (*k*=670). However, our ability to search for more fine-grained regions at higher values of *k,* where we still observe high spatial regularity, provides an excellent opportunity to understand the structural organization of these complex areas. We performed proof-of-concept analyses identifying spatially varying gene expression and cell type abundance and complexity patterns in SC and MRN, with some regions not previously delineated in CCF.

Finally, we demonstrate that our method can be effective in a multi-animal setting in uncovering spatially smooth regions with high similarity to CCF. We found that "batch-effect"-signals such as donor identity and (x, y, z) coordinates are linearly extractable from the embeddings and vary among animals, suggesting distribution shifts among animals but not a specific cause. We also demonstrate that pooling across animals dramatically did not dramatically increase correlation of CellTransformer regions to CCF based on region composition.

We highlight several advantages of our architecture and approach. Although the use of graph-structured architecture or self-attention to model cells in a neighborhood graph is not novel[21–23], we contribute a novel self-supervised training objective that can be interpreted by the extent of spatial correlation between a cell and its cellular neighborhood. CellTransformer is also relatively computationally efficient. We incorporate a packed sequence format (see **Methods**) to reduce memory requirements, with total GPU memory usage <28 GB during training on the ABC-MWB dataset (although we also apply gradient accumulation to decrease batch size). In comparison, we could not run methods such as SpaGCN and STACI due to GPU memory requirements. The intersection of graph neural networks, transformers, and representation learning research is a rich and rapidly moving research area. Potential methodological improvements include incorporation of physical distances and coordinates (for example, via Laplacian positional encoding[24]), Perceiver-based accommodation of variable numbers of input genes[25,26], or rotationally-equivariant architectures[27]. Addendums for interpretability might include biasing of attention operations via prior knowledge about cell-cell interactions or molecular similarity.

There are also significant opportunities to extend CellTransformer's local representation framework to include other data modalities. Both cell-level (e.g. neurophysiology[28,29]) and pixel-level data (e.g. mesoscale axonal connectivity[13], and magnetic resonance imaging[30]) can be tokenized and included in our framework.

Caveats include the necessity of a user-specified spatial radius (for neighborhood computation) and choice of *k* for spatial cluster detection. The stability of detected spatial domains at a given radius or *k* poses an interesting future angle from which to study anatomical hierarchies in the brain, although we qualitatively found that a smaller radius detracted from visual quality of spatial clusters (**Supplementary Figure 11**). Users must also have access to GPUs (to allow for timely model fitting), which reduces overall accessibility. We include a small discussion of observed failure modes (see **Supplementary Note 3**), often manifesting as left-right asymmetry and high-frequency, non-uniform spatial clusters. Most often these can be isolated to a specific tissue section.

CellTransformer advances the state of the art for automated domain detection in the mouse brain, which could potentially inform a diverse set of fields. As spatially resolved transcriptomic and multi-omic studies of the brain become more prevalent[3,31,32], tools such as CellTransformer provide avenues to scalably integrate these data into iteratively refined anatomical maps of the brain and other complex organs.

Methods

Allen Brain Cell Mouse Whole Brain (ABC-MWB) dataset processing

We downloaded the log-transformed MERFISH probe counts and metadata from the Allen Institute public beta (https://alleninstitute.github.io/abc_atlas_access/intro.html) access for ABC-MWB for both the Allen Institute for Brain Science and Zhuang lab mice. We use the ontology from the "20231215" data release. We transformed the (x, y) coordinates of each cell into microns (instead of mm as provided). Otherwise the dataset was used as-is for neural network training. All sections in this dataset are coronal sections.

Zhuang lab (Zhang et al.) dataset processing

For the Zhuang lab dataset mice we downloaded the "20230830" release of data from the Allen Institute ABC-MWB public data. Otherwise, the data were used as-is for neural network training. This dataset consists of four mice worth of data. Two animals (animals 1 and 2) were sectioned coronally and two animals (animals 3 and 4) were sectioned sagittally. Animals 1 and 2 consist of 2,846,909 cells and 1,227,409 cells, respectively. Animals 3 and 4 consist of 1,585,844 cells and 162,579 cells, respectively.

Cellular neighborhood construction

We consider cells in the same neighborhood as a reference cell if the distance between them is within a user-defined threshold. For all datasets we used a radius of 85 micron.

CellTransformer architecture

We construct a CellTransformer to generate a latent representation from a cellular neighborhood where this representation is composed of both molecular and cell type information. We represent cells as nodes in an undirected graph, $\mathcal{G} = \mathcal{V}, \mathcal{E}$ where $\mathcal{V}$ indexes the nodes in the graph (cells) and we add an edge (v_i, v_j) to the edge set $\mathcal{E}$ if $d_{i,j} < r$, with r a user-specified distance in microns. We assume also that for each node we have access to $\mathbf{x}_i \in \mathbb{R}^g$, a g-dimensional vector of MERFISH probe counts. We also assume we are given class labels $\mathbf{c} = \{c_i \in \{1, ..., C\}\}_{i=1}^N$ for each of the N cells. The user must also specify an embedding dimension and number of transformer encoder and decoder layers; in all experiments in this paper we use an embedding dimension of 384, 4 encoder layers, and 4 decoder layers. The meaning of these specific architectural details will be explained in the next paragraph.

To generate a neighborhood embedding, we first identify a particular cell which we call a reference cell. Its first degree neighbors are extracted from $\mathcal{G}$. We first apply a shallow encoder (two layer perceptron with GELU nonlinearity) function $f_\theta : \mathbb{R}^g \rightarrow \mathbb{R}^{192}$ which maps the gene expression into embedding space. We likewise construct and apply the function $g_\theta : \mathbb{R}^C \rightarrow \mathbb{R}^{192}$ to map one-hot encoded cell type labels to embedding space, here a simple lookup into a learnable embedding table. These representations are concatenated into a single 384-dimensional representation. We apply these transformations to each cell in the neighborhood, not including the reference cell. At this point, all operations have been per-cell without interactions. In addition to these cell tokens, we also instantiate for each neighborhood a register token which we use to accumulate global information across the neighborhood. We refer to this token as a <cls>-like token in keeping with previous literature.

We then apply a transformer encoder to the cells, only allowing cells within the same neighborhood and their <cls>-like token. We use 8 attention heads with GELU activations and layer norm prior to attention and perceptron residual addition. We also use a bias for all layers. Following the transformer encoder, we use attention pooling to aggregate the cell and <cls> representations for each

neighborhood into a single token with embedding dimension 384. We refer to these as the neighborhood representations.

We then instantiate a new token from each reference cell that is a learned embedding for each cell type (separately from the encoder cell type embedding). These are concatenated to the neighborhood representations. We then apply a transformer decoder to the tokens. This decoder also has embedding dimension 384 (although this is not a requirement) and has 8 attention heads and bias terms. Only the reference cell token and its corresponding neighborhood representation are allowed to attend to each other.

During training, we extract only the reference cell tokens. We then use separate linear projections to output the mean, dispersion, scale, and zero inflation logit parameters for zero-inflated negative binomial regression. We optimize the model by minimizing the log likelihood of a negative binomial distribution using an 80% train split size (random splitting across the entire dataset) and the ADAM optimizer over 40 epochs. We perform a linear warmup for 500 steps to a peak learning rate of 0.001 and use an inverse-square root learning rate scheduler to decay the learning rate continuously. We use a weight decay value of 0.00005 which we do not warm up. We trained all versions of CellTransformer on a system with 2 NVIDIA A6000 GPUs with effective batch size 256 using gradient accumulation.

Spatial domain detection

Once trained, we apply CellTransformer to a given dataset and instead of extracting reference cell tokens we extract the neighborhood representation. We then cluster this representation using *k*-means. We use the cuml library (`cuml.KMeans`) to perform this operation on GPU, with arguments `n_init=3`, `oversampling_factor=3`, and `max_iter=1000`.

Optional smoothing of embeddings

We observe spatial domains are spatially smooth. However in the case that there is a high-frequency signal that the end-user would like to filter, we optionally introduce a step prior to *k*-means where we smooth the embeddings using a Gaussian filter. For all comparisons except those in **Supplementary Figure 12**, smoothing was performed with a Gaussian filter with 40 micron full-width at half maxima (sigma of 12.01 microns).

Model fitting on Zhuang lab datasets

We perform training from scratch without transfer. We trained for 40 epochs with the same settings as for the Allen Institute for Brain Science mouse.

Computation of stability metrics

We follow Wu et. al[15] in using an Amari-type distance to compare clustering solutions. Briefly, we compute several replicates (20 in this work) of *k*-means at a given choice of *k* with different random seed as $\mathbf{D}_{k,i}$, with i indexing the different centroids for a given solution. We then measure the stability of a given choice of *k* by comparing the similarity of all pairs of $\mathbf{D}_k$. Define $\mathbf{C}$ the cross-correlation matrix between pairs $\mathbf{D}$ and $\mathbf{D}'$. Then we define:

$$\mathrm{diss}(\mathbf{D}, \mathbf{D}') = \frac{1}{2K}\left(2K - \sum_{i=1}^{K} \max_{1\leq k\leq K} \mathbf{C}_{kj} - \sum_{j=1}^{K} \max_{1\leq k\leq K} \mathbf{C}_{kj}\right).$$

to identify the choice of *k* which is most stable.

Regional matching with CCF computation

To quantify overall similarity of regions extracted using CellTransformer with CCF, we first extract cell type composition vectors for each region at a given level of the hierarchy. For all comparisons in **Figure 2**, we use the subclass level (338 cell types), resulting in *k*-region by 338 matrices. We then compute for a region derived from either CellCharter or CellTransformer the best match (maximum value of Pearson correlation, non-exclusively) with CCF (**Figure 2c**) or an exclusive match (using the Hungarian matching algorithm, also using Pearson correlation as the metric, **Figure 2d**).

CellCharter

To run CellCharter we first generated scVI embeddings using the default settings for depth and width of the network and with the tissue section labels as conditional batch variables. We trained for 50 epochs with early stopping, using the `early_stopping=True` setting. We then aggregated across 3 (default settings), 6, 9 layers using the `cellcharter.gr.aggregate_neighbors` function. We then applied CellCharter's Gaussian mixture model implementation at various choices of the number of Gaussians. We could not run the mixture model with our hardware (A6000, 48GB GPU memory) for more than 9 layers, which was also the number which produced the highest correspondence with CCF and is reported in **Figure 2e**.

Nearest-neighbor smoothness computation

To quantify smoothness of the spatial domains, we use a nearest-neighbor approach. We extract approximate spatial neighbors for each cell using `cuml.NearestNeighbor` with 100 neighbors, restricting neighbors to be within the same tissue section. For a given domain set, either from CCF, CellTransformer, or CellCharter, we extract the spatial domain label of the given cell and count the proportion of times that cell is observed in the 100 neighbors. These proportions are averaged across all cells and tissues.

Linear probing experiments

We extract neighborhood representations for each of the cells in the Zhuang lab datasets. First, we regress these embeddings on the (x, y, z) coordinates. We then computed the absolute prediction error in terms of the coordinates and then reported the average. We also fit a multi-class logistic regression using the mouse donor identity. For the logistic regression we use `cuml.LogisticRegression` with default settings in cuml. For the linear regression we apply simple least squares using PyTorch via QR decomposition.

Quantification of spatial contribution to gene expression

We interpret the accuracy of the gene expression predictions for a given cell as an index of correlation of an instance of a particular cell type with its surrounding neighbors. To do this, we compute a simple baseline model which predicts average gene expression (computed across the entire Allen Institute for Brain Science mouse dataset) for each cell. We compute the average Pearson correlation for each instance of a given cell type and average across instances to obtain an average Pearson correlation. We then compute a Pearson correlation between each cell's observed gene expression and the CellTransformer predictions, averaging similarly across instances of a given cell type. The difference between the baseline and model predictions is displayed, per cell type, and grouped across neurotransmitter types in **Supplementary Figure 13**.

Zhuang lab dataset per-animal CCF comparison

We contrast two methods of extracting spatial domains from the Zhuang lab dataset, which consists of four mice. We first fix *k,* the number of desired spatial domains. Then we fit one *k*-means model on all of the neighborhood embeddings for all of four mice together. We also fit a *k*-means model to each of the

mice separately. We then compute the similarity of these region sets using the same method used to quantify differences between CellCharter and CellTransformer by comparing the regional cell type composition vectors.

Other software

Principal software used in this work includes PyTorch[33], numpy[34], scikit-learn[35], scipy[36], scanpy[37], cuml[38], matplotlib[39], and seaborn[40].

Code and data availability

Code will be publicly available at https://github.com/abbasilab/celltransformer. All data used for this publication is available either at the Allen Institute ABC-MWB data portal (https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas) or the CZI cellxgene portal[41].

Acknowledgments

The authors would like to acknowledge Forrest Collman for help with Neuroglancer data visualization of spatial domains. We acknowledge Patrick Xian for helpful discussions. AJL acknowledges Katharine Z. Yu, Chang Kim, Parker Grosjean, Lee Rao, and Tom Nowakowski for feedback on neuroscientific and machine learning contexts of paper. RA, BT, AJL, and HZ would like to acknowledge support from the Weill Neurohub through the Weill Neurohub's Next Great Ideas Award. RA would like to acknowledge support from the National Institute of Mental Health of the National Institutes of Health under award number RF1MH128672 and Sandler Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. AJL acknowledges support from the UCSF Discovery Fellowship and the ARCS Fellowship. Generation of Allen Institute for Brain Science data in ABC-MWB and subsequent registration to CCFv3 was funded through 1U24MH130918-01.

References

1. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. 2023.03.06.531121 Preprint at https://doi.org/10.1101/2023.03.06.531121 (2023).
2. Consortium, T. 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 (2023).
3. Chen, A. *et al.* Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex. *Cell* **186**, 3726-3743.e24 (2023).
4. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
5. Townes, F. W. & Engelhardt, B. E. Nonnegative spatial factorization applied to spatial genomics. *Nat.*

Methods **20**, 229–238 (2023).

6. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
7. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. Preprint at https://doi.org/10.48550/arXiv.2111.06377 (2021).
8. Hou, Z. *et al.* GraphMAE: Self-Supervised Masked Graph Autoencoders. Preprint at https://doi.org/10.48550/arXiv.2205.10803 (2022).
9. Wang, Q. *et al.* The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181**, 936-953.e20 (2020).
10. Veličković, P. *et al.* Graph Attention Networks. Preprint at https://doi.org/10.48550/arXiv.1710.10903 (2018).
11. Zhang, M. *et al.* A molecularly defined and spatially resolved cell atlas of the whole mouse brain. 2023.03.06.531348 Preprint at https://doi.org/10.1101/2023.03.06.531348 (2023).
12. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nat. Genet.* **56**, 74–84 (2024).
13. Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
14. Haviv, D. *et al.* The covariance environment defines cellular niches for spatial inference. *Nat. Biotechnol.* 1–12 (2024) doi:10.1038/s41587-024-02193-4.
15. Wu, S. *et al.* Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci.* **113**, 4290–4295 (2016).
16. Ding, S.-L. *et al.* Distinct Transcriptomic Cell Types and Neural Circuits of the Subiculum and Prosubiculum along the Dorsal-Ventral Axis. *Cell Rep.* **31**, 107648 (2020).
17. Bienkowski, M. S. *et al.* Integration of gene expression and brain-wide connectivity reveals the multiscale organization of mouse hippocampal networks. *Nat. Neurosci.* **21**, 1628–1643 (2018).
18. Cembrowski, M. S. *et al.* The subiculum is a patchwork of discrete subregions. *eLife* **7**, e37701 (2018).
19. Bienkowski, M. S. *et al.* Homologous laminar organization of the mouse and human subiculum. *Sci. Rep.* **11**, 3729 (2021).

20. Benavidez, N. L. *et al.* Organization of the inputs and outputs of the mouse superior colliculus. *Nat. Commun.* **12**, 4004 (2021).
21. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
22. Brbić, M. *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nat. Methods* **19**, 1411–1418 (2022).
23. Fischer, D. S., Schaar, A. C. & Theis, F. J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol.* **41**, 332–336 (2023).
24. Maskey, S. *et al.* Generalized Laplacian Positional Encoding for Graph Representation Learning. *arXiv.org* https://arxiv.org/abs/2210.15956v2 (2022).
25. Connell, W., Khan, U. & Keiser, M. J. A single-cell gene expression language model. *arXiv.org* https://arxiv.org/abs/2210.14330v1 (2022).
26. Jaegle, A. *et al.* Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv.org* https://arxiv.org/abs/2107.14795v3 (2021).
27. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *arXiv.org* https://arxiv.org/abs/2006.10503v3 (2020).
28. Laboratory, I. B. *et al.* A Brain-Wide Map of Neural Activity during Complex Behaviour. 2023.07.04.547681 Preprint at https://doi.org/10.1101/2023.07.04.547681 (2023).
29. de Vries, S. E., Siegle, J. H. & Koch, C. Sharing neurophysiology data from the Allen Brain Observatory. *eLife* **12**, e85550 (2023).
30. Hike, D. *et al.* High-resolution awake mouse fMRI at 14 Tesla. *eLife* **13**, (2024).
31. Chen, X. *et al.* High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. *Cell* **179**, 772-786.e19 (2019).
32. Chen, X. *et al.* Whole-cortex in situ sequencing reveals input-dependent area identity. *Nature* 1–10 (2024) doi:10.1038/s41586-024-07221-6.
33. Paszke, A. *et al.* Automatic differentiation in PyTorch. (2017).
34. Array programming with NumPy | Nature. https://www.nature.com/articles/s41586-020-2649-2.
35. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830

(2011).

36. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).
37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
38. Raschka, S., Patterson, J. & Nolet, C. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *ArXiv Prepr. ArXiv200204803* (2020).
39. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
40. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
41. Program, C. S.-C. B. *et al.* CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. 2023.10.30.563174 Preprint at https://doi.org/10.1101/2023.10.30.563174 (2023).
42. Ollivier, M. *et al.* Crym-positive striatal astrocytes gate perseverative behaviour. *Nature* **627**, 358–366 (2024).

Supplementary Notes

Supplementary Note 1: effect of smoothing and analysis on striatal glial populations

When scaling up the number of regions past 500, we observed that almost all spatial clusters were spatially smooth except for a recurring pattern in the striatum. We plot (**Supplementary Figure 12a-b**) six sequential sections where we identified an irregular (which we define here as a broadly non-convex shape that does not form relatively singular connected component) pattern of cells in the striatum and only the striatum (note the spatial uniformity of areas surrounding striatum in cortex and endopiriform area, nucleus accumbens etc.). We identified cells in these areas and found they were mostly non-neuronal, with astrocytic types (such as 1163 **Astro-TE NN_3, Supplementary Figure** 12c) forming a large proportion of cells.

We sought to understand whether these spatial clusters might be biologically relevant or somehow related to noise. We observed a recent paper, Ollivier et al. (2024)[42], which identified a novel population of *Crym+* astrocytes in a similar spatial distribution as observed in our regions, specifically in a dorsoventral and lateromedial distribution (see **Supplementary Figure 12e** for reproduction from Ollivier et al., granted with permission). As *Crym* was included in the MERFISH panel, we quantified *Crym* expression in astrocytes within these areas, finding that all but two of these spatially irregular domains

had very high levels of *Crym* expression. Notably, the two groups with lower expression, spatial clusters 457 and 758, were the most dorsolateral, and are distributed where *Crym+* astrocytes were not observed in Ollivier et al. We reasoned that these spatial clusters, therefore, may have biological relevance.

However, to simplify downstream analyses, we applied a simple smoothing operation (see **Methods**), which removed this spatial cluster in successive clustering operations. We use a very small smoothing window (12 micron sigma, or 40 micron full-width at half-maxima), and reasoned that because otherwise regions were highly spatially smooth that any downstream metrics were minimally affected.

Supplementary Note 2: interpreting the CellTransformer objective as a measure of spatial dependence

One interpretation of the CellTransformer architecture is learning two representations of cellular gene expression. The first (the learned embedding for each cell type) is unconditional on the spatial neighborhood information. The second is one that is conditional on the spatial neighborhood information learned in the encoder portion of the network and parameterized as a residual update in the transformer decoder, suiting the established mathematical structure of the transformer block. We compare the increase in gene expression prediction accuracy from using the conditional spatial neighborhood representation to the unconditional one and interpret this value as an index of spatial dependence. A trivial or poorly fit optimization solution would produce a small value of this index, and so we hypothesized CellTransformer would achieve reasonable improvements in expression prediction, but also cell type-to-type differences.

Specifically, we use the change in Pearson correlation from using our model as a predictor and a trivial predictor of the per-cell type average gene expression. We observe increases in the predictive accuracy (mean 0.10 +/- 0.0701 in correlation). Moreover, there are few cell types for which there is a decrease in predictive accuracy, indicating that our model has nontrivially learned the objective (**Supplementary Figure 13a**). This accuracy has only a mild correlation with cellular density (**Supplementary Figure 13b**). Note that we compute accuracies at the subclass level (338 types). Increase in accuracy does correlate strongly with log-number of observations per cell type, (Pearson correlation of 0.71, **Supplementary Figure 13c**), possibly indicating inefficiencies in pretraining.

We note another way to interpret the model is to consider it a contextualized embedding for the reference cell based on spatial information, however we do not explore the implications of this formulation in this manuscript..

Supplementary Note 3: visual failure mode characterization in Allen Institute for Brain Science Mouse

We characterize two major failure modes in the Allen Institute for Brain Science mouse data.

On sections 54 and 51 there is consistently a tear artifact across the section (**Supplementary Figure 14a**, brown cluster in striatum) which does not appear in sequential sections, although some structures are found (for example, yellow sections of cortex at *k*=100) despite presence of this artifact. At higher *k* the tearing artifacts appear to subdivide into different groups.

On section 39 (**Supplementary Figure 14b**) there is a similar tearing artifact around the edges of the section, where tissue loss has occurred. Also, in section 40 we repeatedly observed across choices of *k* appearance of asymmetric thalamic and cortical regions (although not all are asymmetric) and we note that in other sections there can be mild symmetry breaking. Tissue sectioning variability across the

midline (left and right sides appear mildly offset in anterior-posterior axis when viewing the CCF region registration) is difficult to disambiguate from errors in the algorithm.

Figures

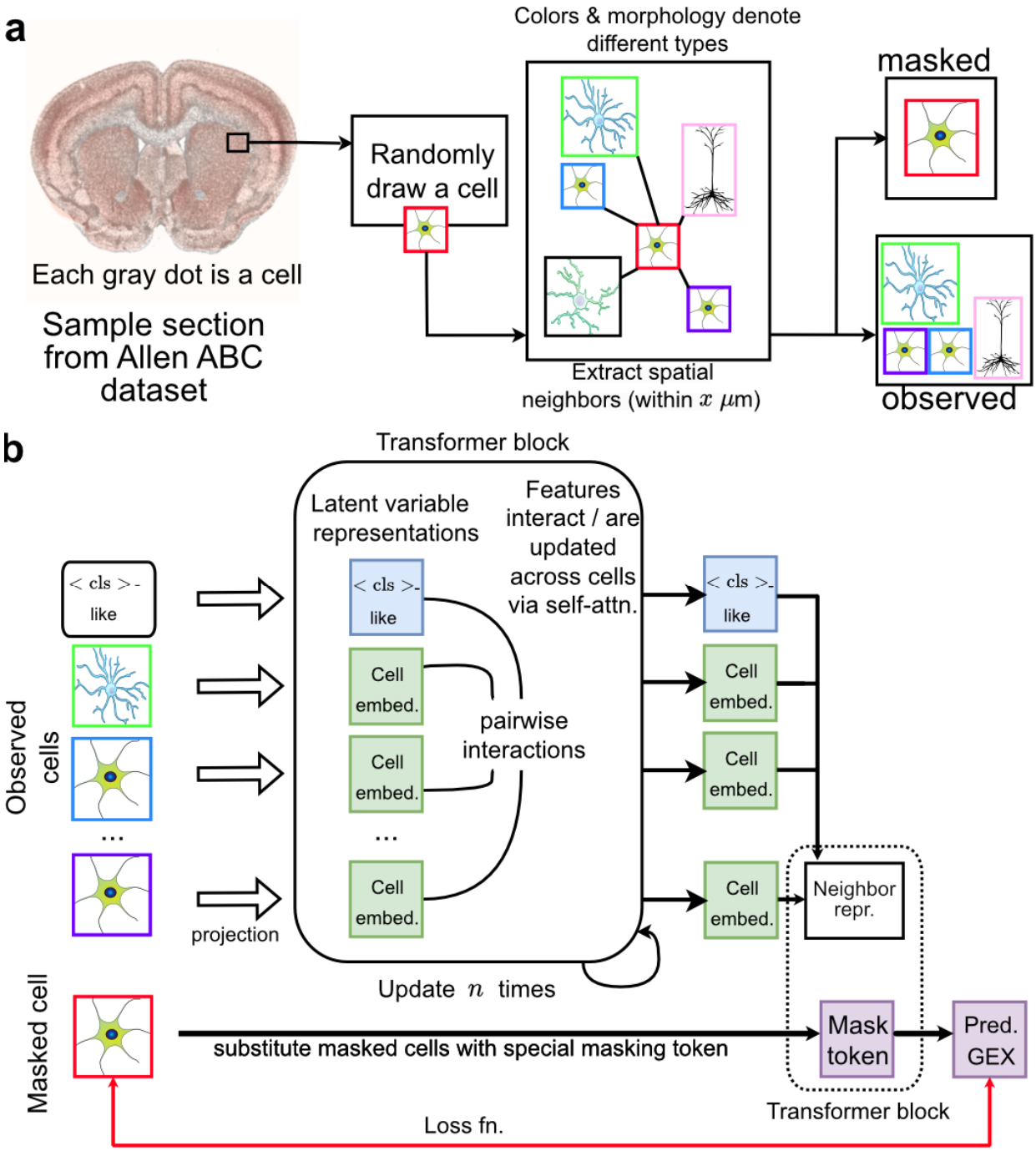


Figure 1. Overall training and architectural scheme for CellTransformer. (**a.**) During training, a single cell is drawn (we denote this the reference cell, boxed in red). We extract the reference cell's spatial neighbors and partition the group into a masked reference cell and its observed spatial neighbors. (**b.**) Initially, the model encoder receives

information about each cell and projects those features to *d*-dimensional latent variable space. Features interact across cells (tokens) through the self-attention mechanism, which is repeated *n* times. These per-cell representations and an extra token are then aggregated into a single vector representation, which we refer to as the neighborhood representation. This representation is concatenated to a mask token which is cell type specific and chosen to represent the type of the reference cell. A shallow transformer decoder (dotted lines) further refines these representations and then a linear projection is used to output parameters of a negative binomial distribution modeling of the MERFISH probe counts for the reference cell.

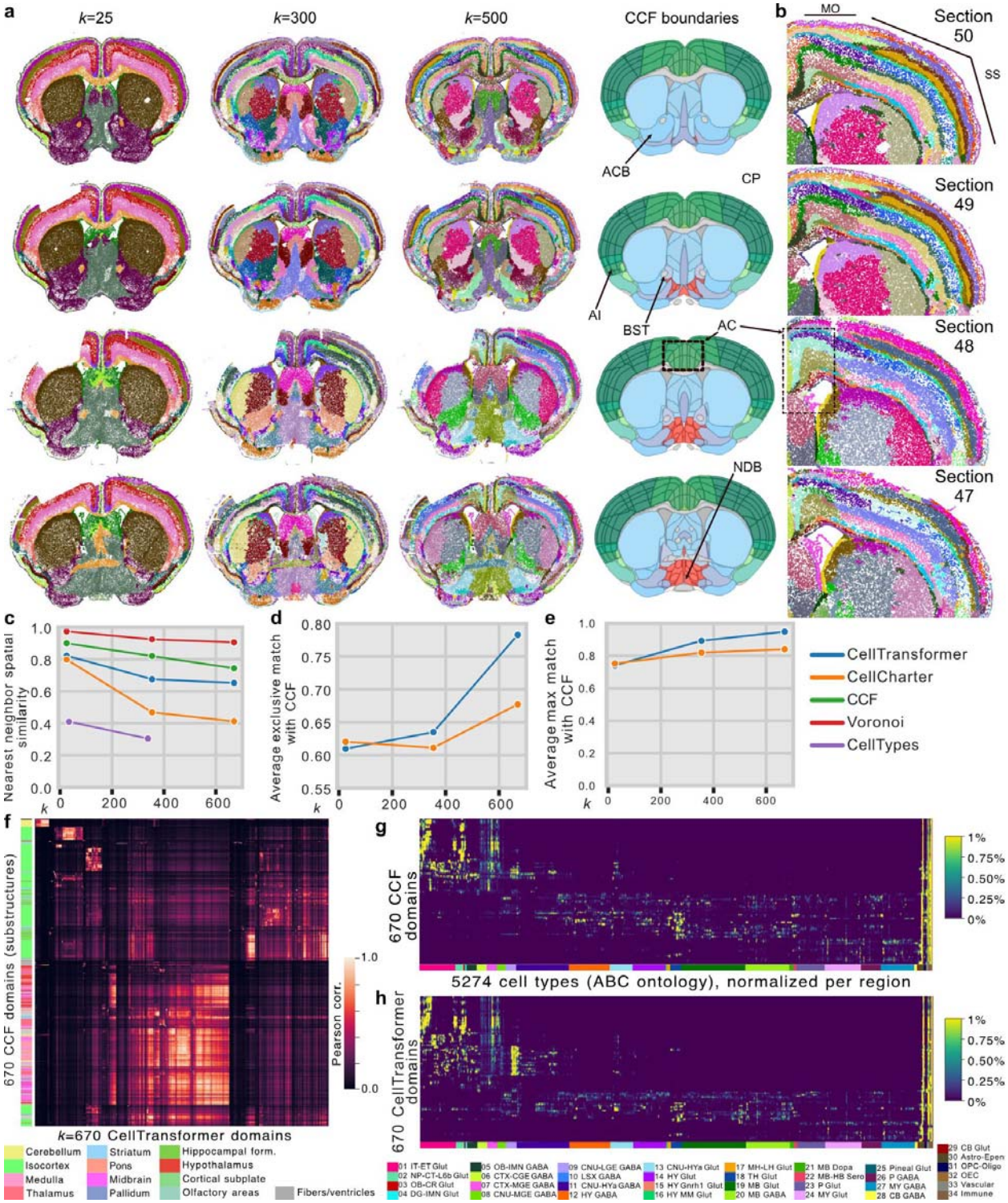


Figure 2. Representative images of spatial domains discovered using CellTransformer and comparison to CCF. (**a**.) Four sequential tissue sections (note that the inter-section distance is roughly 200 µm) from anterior to posterior. In the first three columns, each dot is a cellular neighborhood colored by the cluster label when *k*-means is conducted with *k* = 25, 300, and 500. Spatial domain labels are depicted with the same colors across sections. Fourth column shows CCF region registration to the same tissue section. The anterior cingulate (AC) is boxed. Caudoputamen (CP), nucleus accumbens (ACB), agranular insula (AI), bed nucleus (BST), and the diagonal band nucleus of the pallidum (NDB) are also labeled. (**b.**) Single hemisphere images of same tissue sections in **a.** at *k* = 500, zoomed in on cortical

layers of motor cortex (MO) and somatosensory cortex (SS). (**c.**) Comparison of cell type composition of automated CellTransformer and CellCharter[12] domains with CCF ones, averaging the Pearson correlation of each CCF region's cell type composition with its best match (nonexclusively) from CellTransformer, CellCharter, *k*-means on the spatial coordinates (labeled Voronoi, red) and the class and subclass cell type assignments (purple). (**d.**) Same as **c.** but quantifying a maximum similarity of any given CellCharter or CellTransformer region to any CCF region, exclusively. (**e.**) Comparison of CCF, CellCharter, and CellTransformer by quantifying number of nearest neighbors with the same spatial region label. (**f.**) Region-by-region Pearson correlation matrix comparing cell type composition vectors from 670 CCF regions (at substructure level) with 670 spatial domains from CellTransformer. The CCF regions are shown on the left with their structure annotations from CCF at division level. (**g.**) Cell type (cluster level) by region matrix for 670 CCF regions at substructure level. Rows are normalized to sum to 1. (**h.**) Cell type (cluster level) by region matrix for 670 CellTransformer regions. Rows are normalized to sum to 1. Colors along x-axis in both **g.** and **h.** show high level class annotations from ABC-MWB cell type ontology at class level. Rows of both **g.** and **h.** are grouped using clustering to produce approximately similar structure.

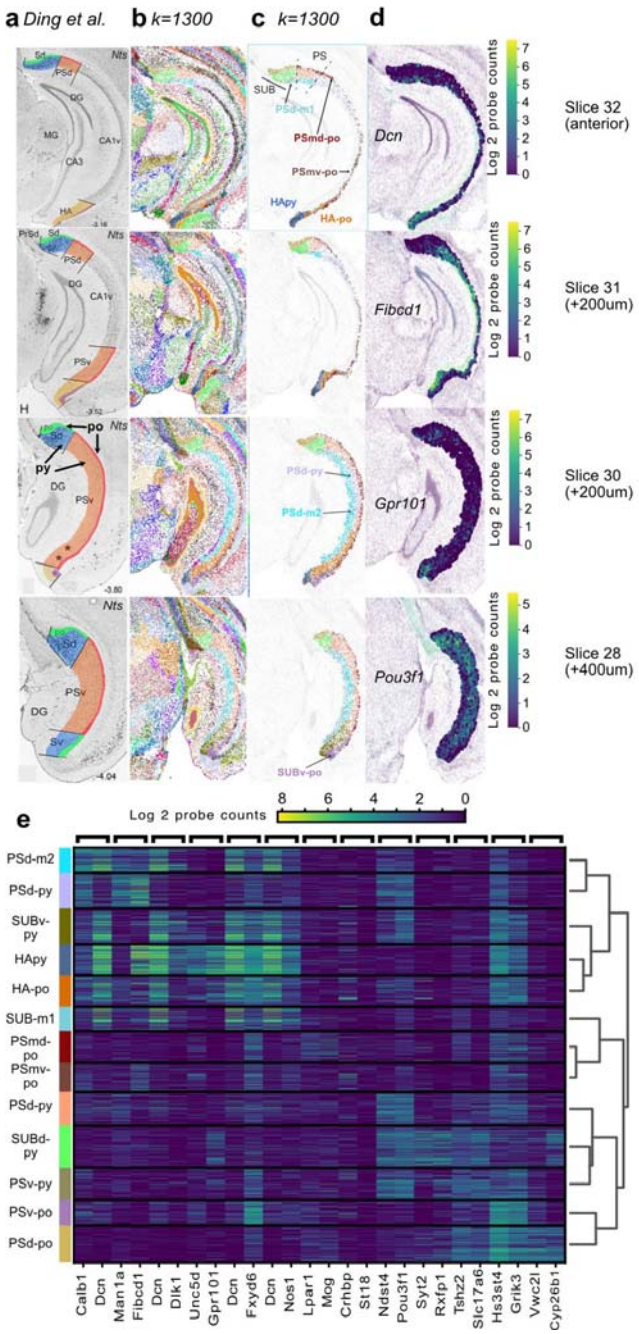


Figure 3. Comparison of CellTransformer domain sets with *k*=1400 with a comprehensive region set found in Ding et al. (2020), reproduced with permission from authors. (**a.**) Representative images reproduced from Ding et al. of region boundaries in prosubiculum (PS), subiculum (SUB) or hippocampal amygdala transition area (HATA). Abbreviations: m/v/d: medial, dorsal or ventral; m: molecular layer; py: pyramidal layer; po: polymorphic layer. (**b.**) Images from hippocampal formation across 4 sequential tissue sections (anterior to posterior). Each dot is a cellular neighborhood colored using cluster labels with *k*=1300 (across the entire brain). (**c.**) Same as **b.** but only showing cells inside PS, SUB, and HATA. (**d.**) Gene expression patterns visualized at the corresponding tissue section, where only cells within PS/SUB/HATA are shown. Units are in log 2 probe counts. (**e.**) Gene expression heatmap of identified subregions, with approximate anatomical annotation. Dendrogram from hierarchical clustering in gene expression space is shown to the right.

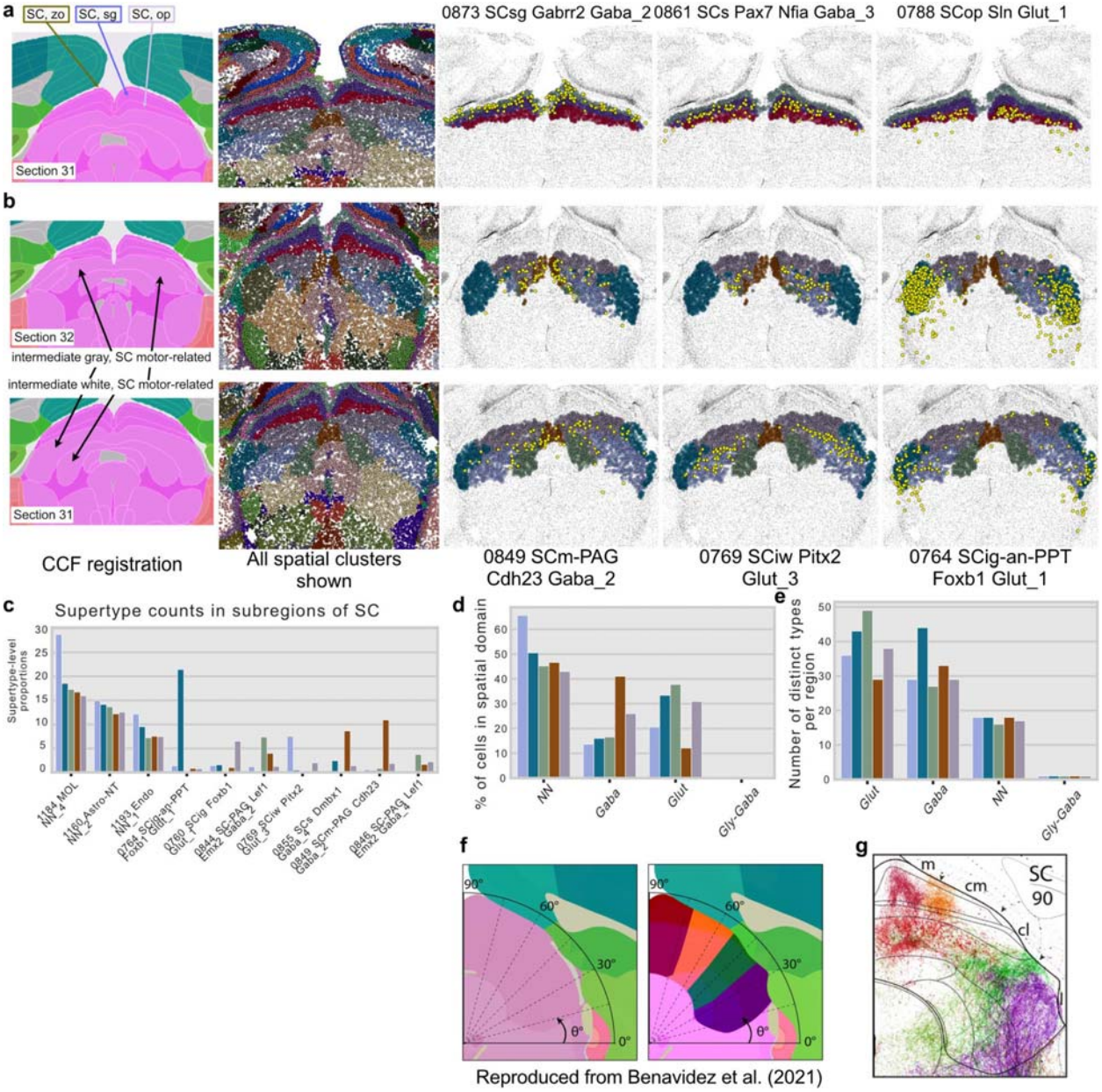


Figure 4. Examination of putative dorsal and ventral subregions of superior colliculus using CellTransformer. (**a.**) Putative subregions of sensory layers of superior colliculus in tissue section 32. CCF registration in the first column, with the zonal, superficial gray, and optic layers labeled by the color of their CellTransformer cluster in third, fourth, and fifth columns. The second column shows all cells in the field of view with color labels from their spatial domain from CellTransformer at *k*=1300. Third column shows only cells inside the zonal, superficial gray, and optic layers of superficial colliculus with supertypes of cells specific to those areas shown in yellow. (**b.**) Sequential tissue sections (32: anterior, 31: posterior) shown similarly to (**a.**) but visualizing subregions of the intermediate gray and intermediate white layers, which are shown in black arrows in the CCF registered annotation image. (**c.**) Proportions of different supertype level cell types for top-ten most abundant types in different spatial domains. Colors refer to the same spatial cluster label in (a) and (b). (**d.**) Barplot of the percentage of cells for a given neurotransmitter type found in a given region. (**e.**) Number of unique cell types at supertype level in ABC-MWB ontology found in each region. (**f.**) Figure reproduction with permission from Benavidez et al. (2021)[20], showing proposed schema of layer-specific anterograde connectivity patterns distributed across medial-lateral zones (labeled red, orange, green and purple) of superior colliculus. (**g.**) Figure reproduction from Benavidez et al. (2021) showing example anterograde connectivity patterns from regions projecting to superior colliculus.

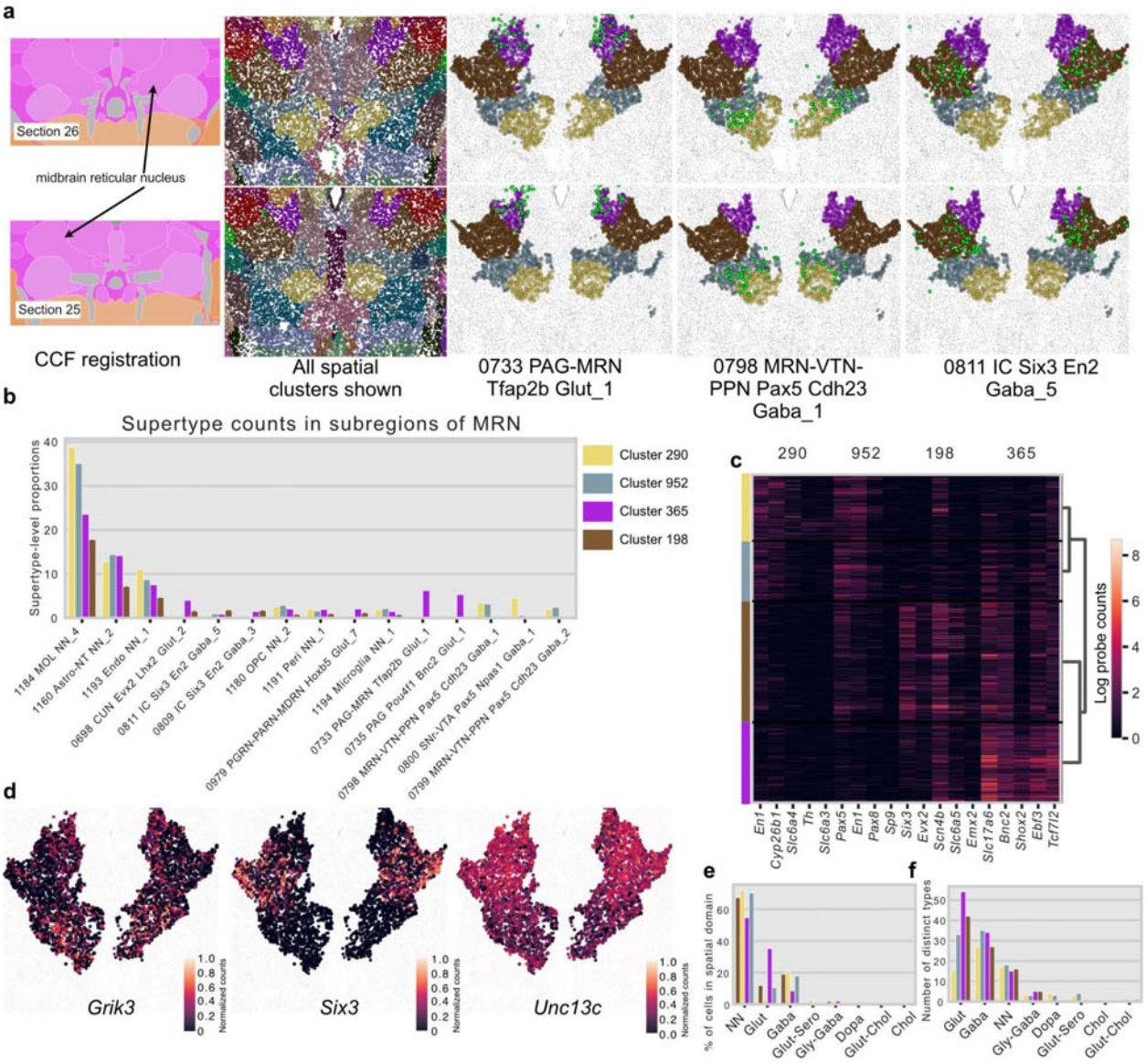


Figure 5. Subregions of midbrain reticular nucleus discovered using CellTransformer. (**a.**) Sequential tissue sections (26 and 25, anterior to posterior). First column: CCF registration with approximate location of midbrain reticular nucleus shown with arrows. Note that registration is not exact and can differ across hemispheres. Second column: all cells in field of view, with color from spatial clusters determined with *k*=1300. The rest of the columns show only cells located in the MRN, and each column shows a different supertype level cell type in green. (**b.**) Supertype level cell type proportions for top ten most abundant types across the MRN subregions. (**c.**) Top 5 differentially expressed genes across regions. (**d.**) Gene expression gradients across tissue section 25 for *Grik3*, *Six3*, and *Unc13c*, showing specificity for each of the putative MRN subregions. Intensity of color is 0-1 normalized after log scaling raw probe counts. Each dot is a cell, and the color shows the relative transcript count. We show only cells within the subregions to make it visually easier to distinguish the relevant cells. (**e.**) Bar plot of the percentage of cells for a given neurotransmitter type found in each domain. (**f.**) Number of unique cell types at supertype level in ABC-MWB ontology found in each domain.

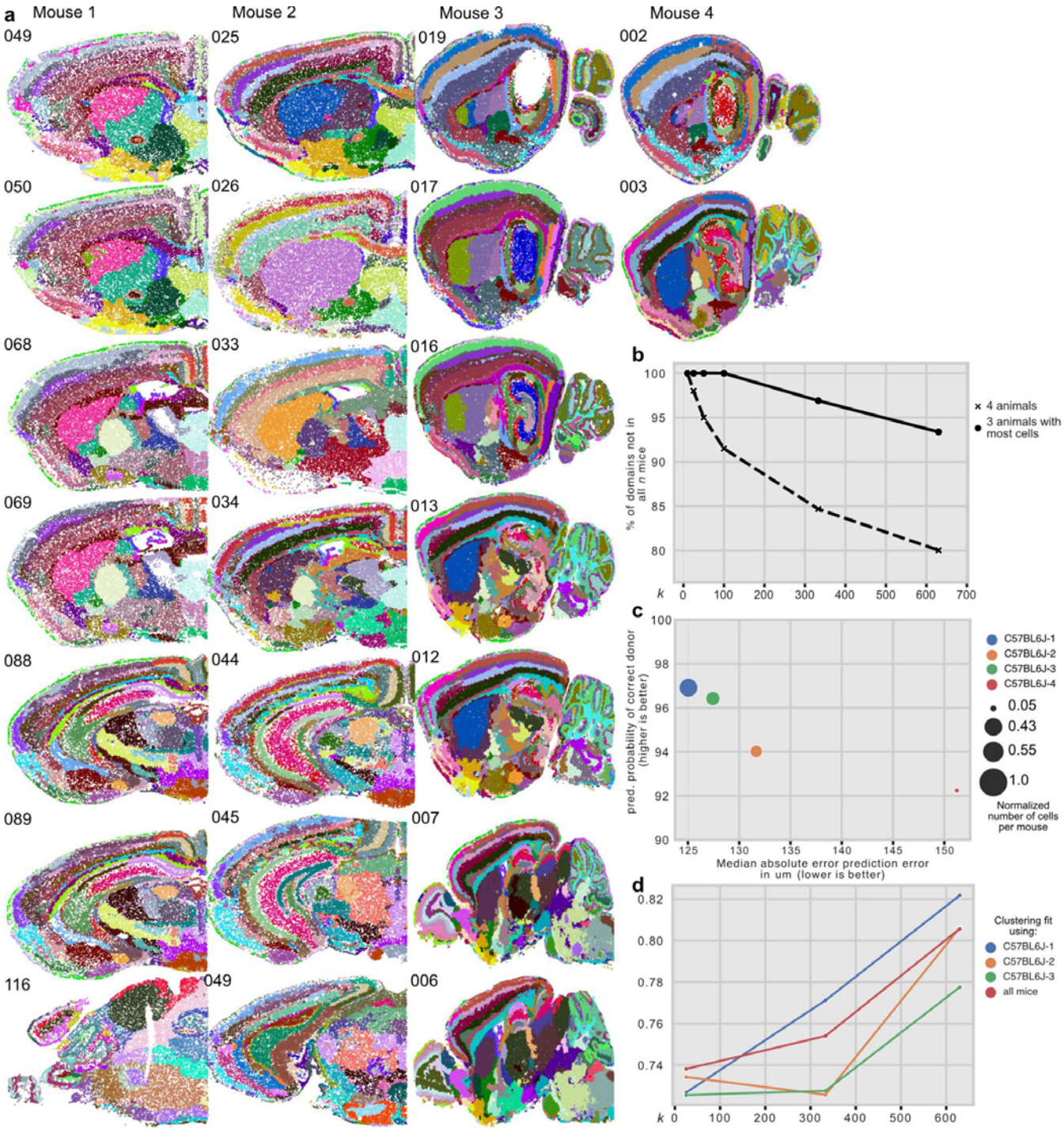
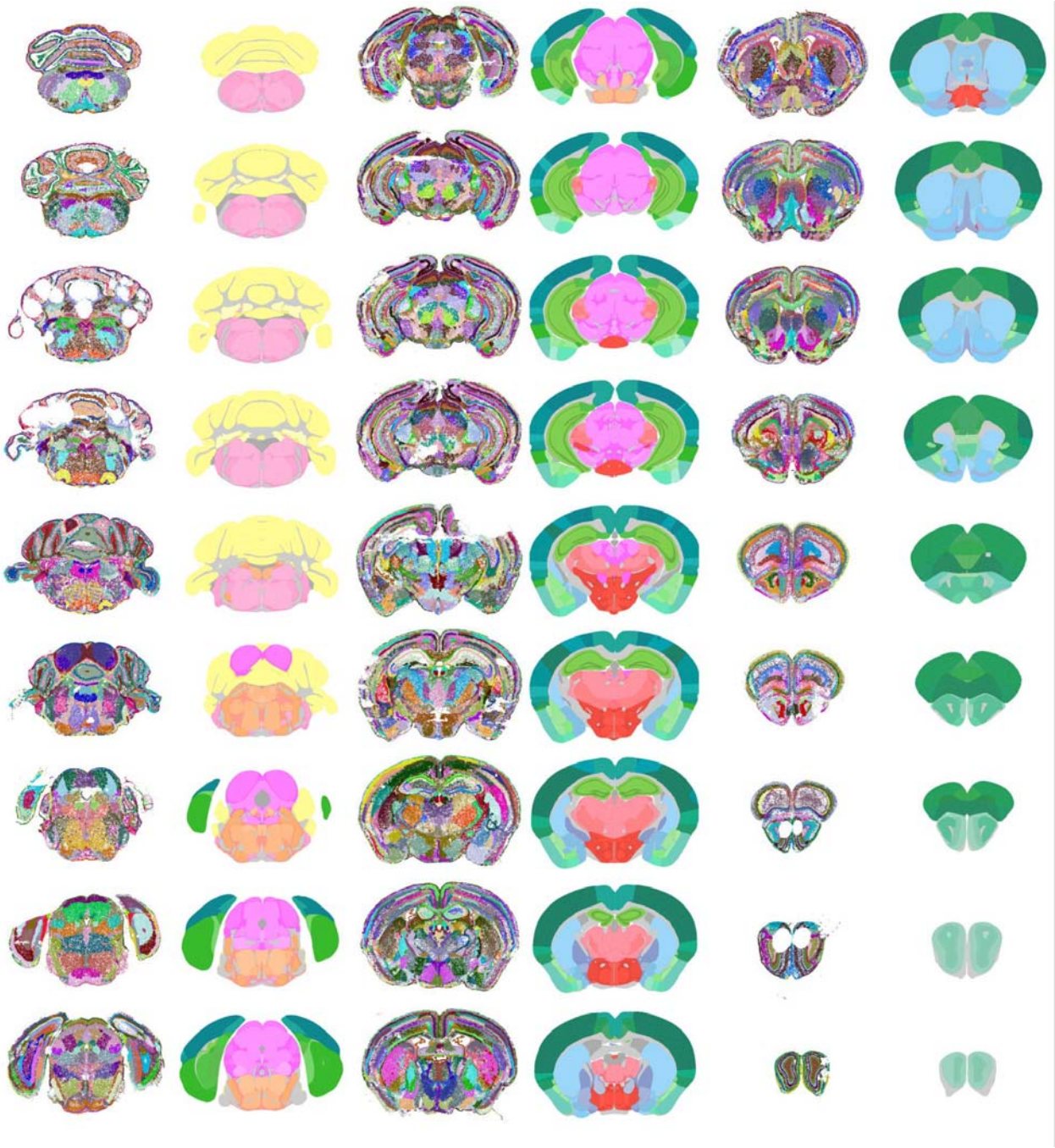


Figure 6. Investigation into performance of CellTransformer on all four Zhuang lab animals. (**a.**) Representative images of all four mice arranged by column. The section number for each mouse is shown in the upper left of each image. Note that mouse 4 only had three sections. For each image, each dot is a cell neighborhood and colors come from a spatial clustering with *k*=330 (number of CCF regions at structure level), fit by concatenating embeddings across mice. (**b.**) Quantification of number of per-mouse specific spatial clusters, computed by clustering at different *k* and computing the number of clusters found for all mice (4 animals) and for the three mice with the most cells per mouse (animals, 1, 2, and 3). Note that because the sections are sampled irregularly some brain areas will have marginally higher coverage in one brain or another. (**c.**) Quantification of subject-level information present in embeddings using linear regression. The median absolute prediction error (x-axis) quantifies accuracy in predicting

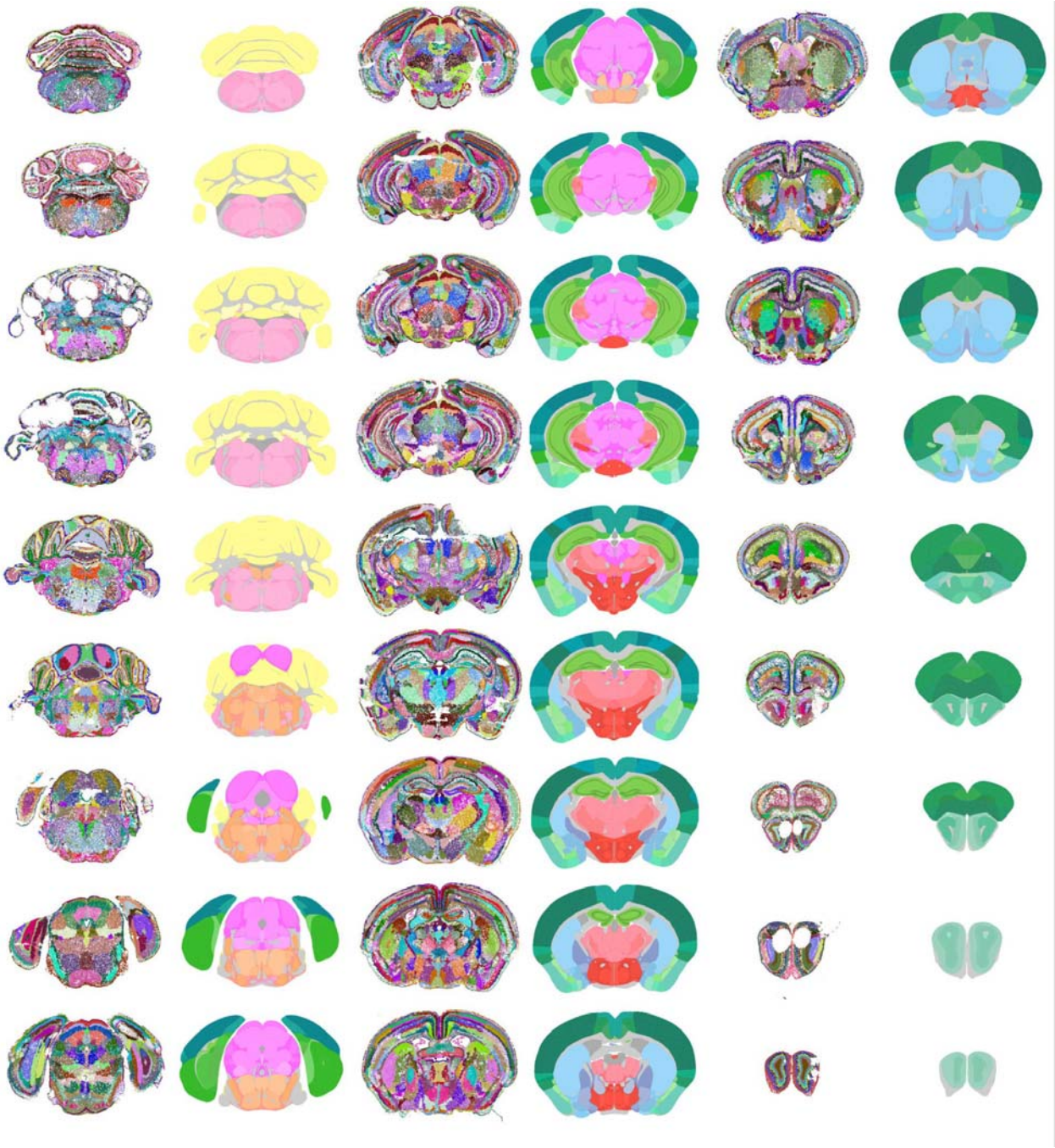
the (x, y, z) coordinates of a neighborhood from its embeddings. The y-axis quantifies accuracy when predicting mouse identity from embeddings using logistic regression. Values are averaged across instances per mouse. (**d.**) Average correlation of the cell type composition of brain regions computed CellTransformer to CCF regions, computed using the Hungarian matching algorithm (exclusively matching regions from one set to the other). Dotted lines indicate results when fitting using all three mice with >1M cells together. Solid lines indicate results when computing spatial clustering on each mouse in isolation.

Supplementary Figures

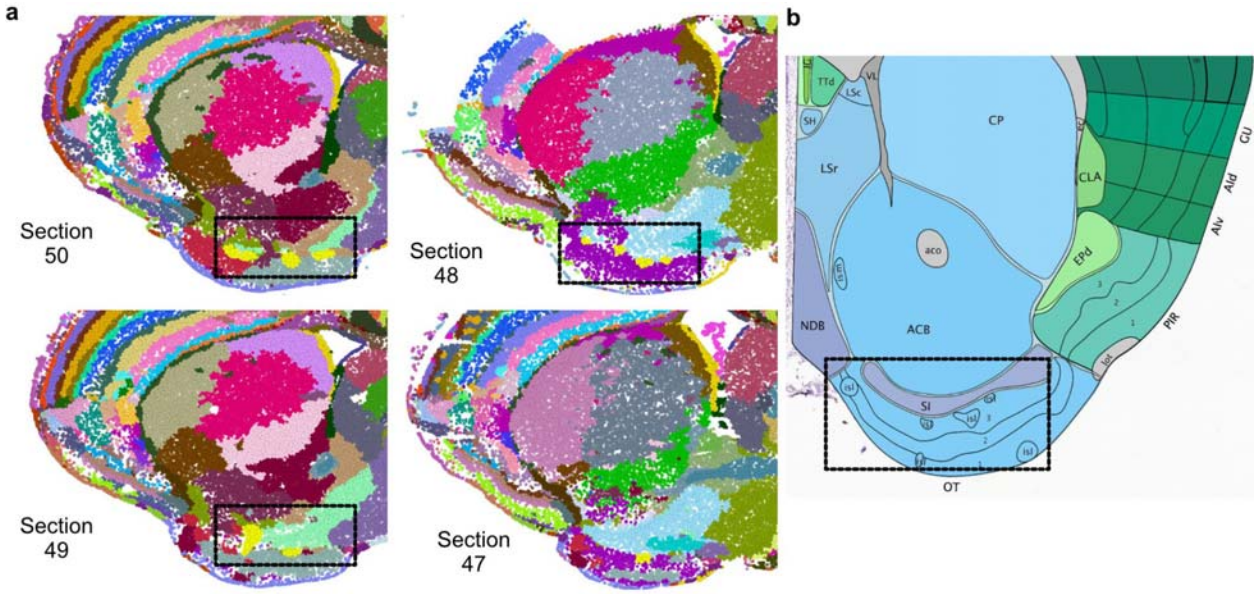
Supplementary Figure 1. Roughly half of the sections in the Allen Institute Brain Science mouse, approximately every other section. For each section color labels come from *k*-means with *k*=100 clusters and the cartoon image to the right is the CCF registered region set.



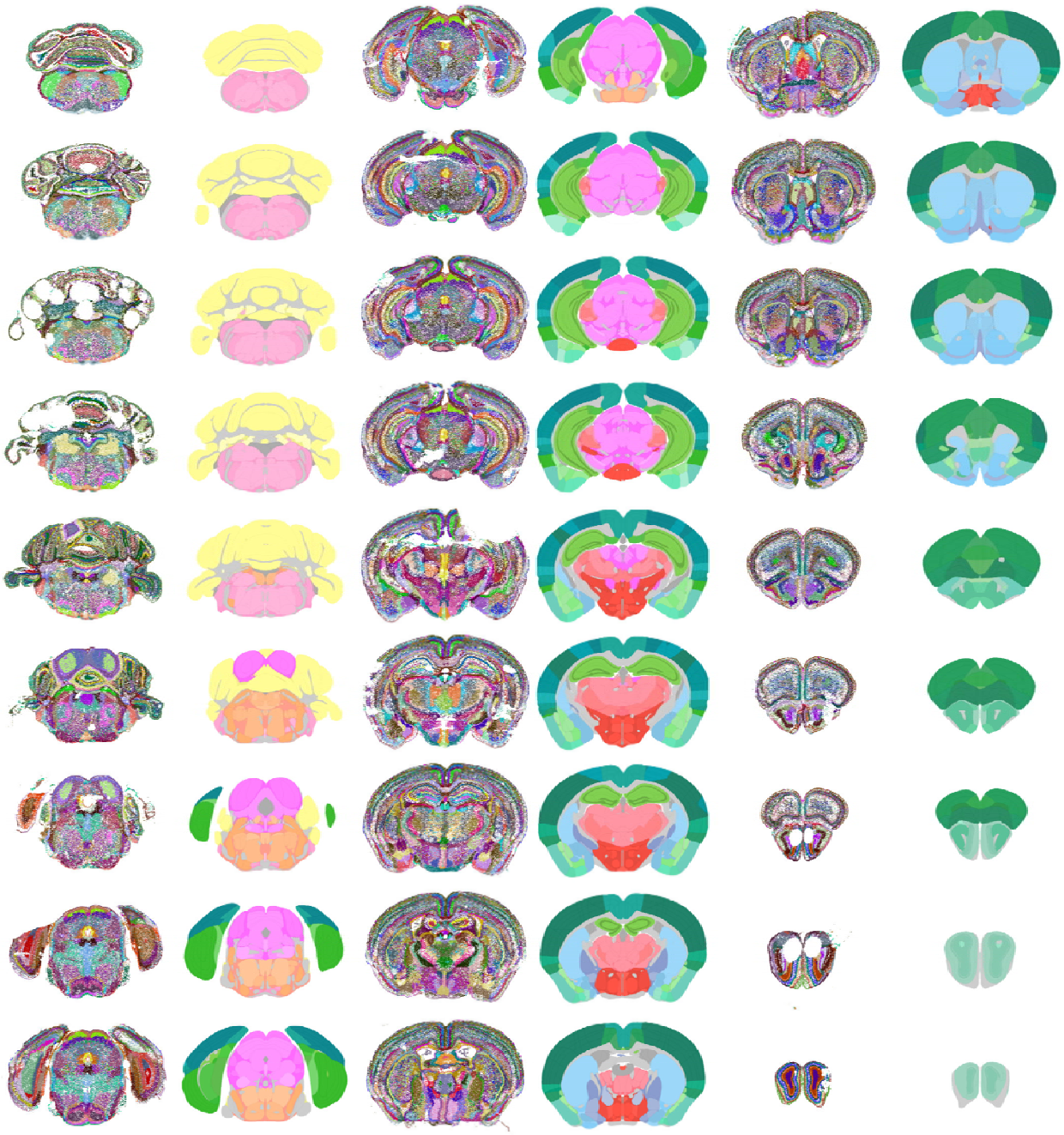
Supplementary Figure 2. Roughly half of the sections in the Allen Institute Brain Science mouse, approximately every other section. For each section color labels come from *k*-means with *k*=1000 clusters and the cartoon image to the right is the CCF registered region set.



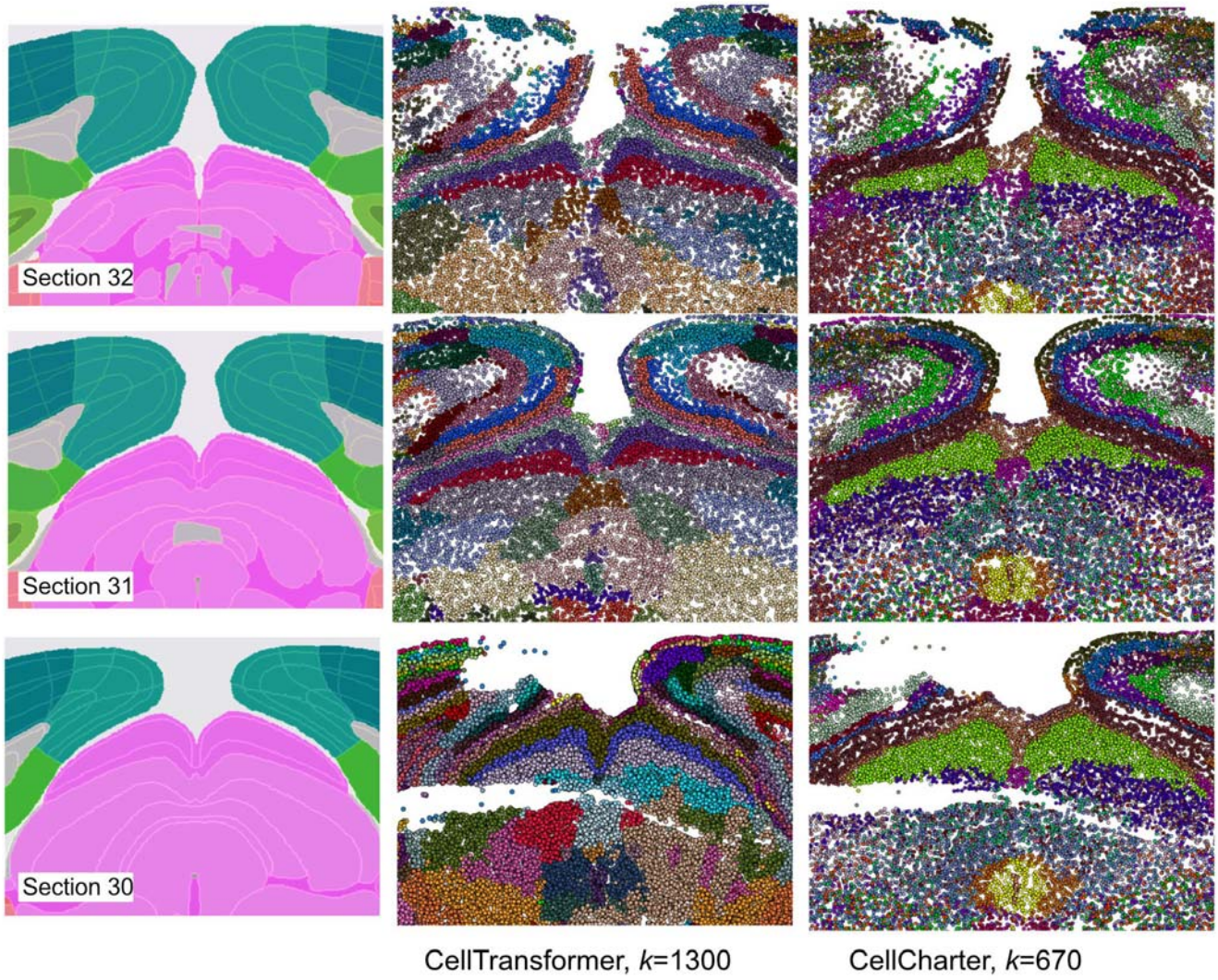
Supplementary Figure 3. Roughly half of the sections in the Allen Institute Brain Science mouse, approximately every other section. For each section color labels come from *k*-means with *k*=1300 clusters and the cartoon image to the right is the CCF registered region set.



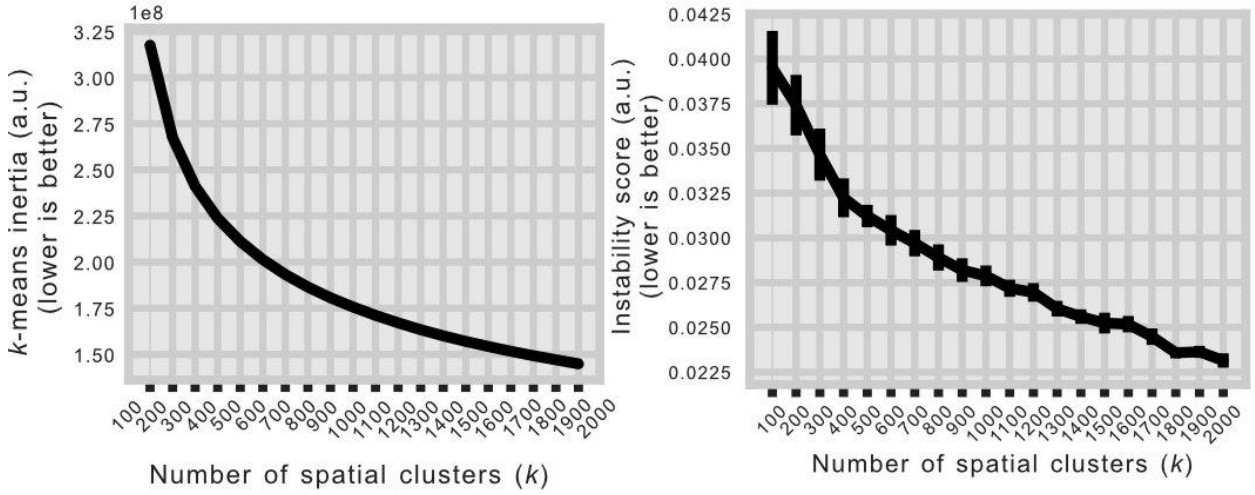
Supplementary Figure 4. (**a.**) Left: spatial cluster assignments (k=500). Right: approximate CCF position (ARA 43/132, P56 Coronal). Black boxes indicate Islands of Calleja and red boxes indicate Major Islands of Calleja. (**b.**) Zoom in on the left hemisphere of sections 50, 49, 48, and 47 (50 is most anterior section), showing segmentation into fine-grained subregions.



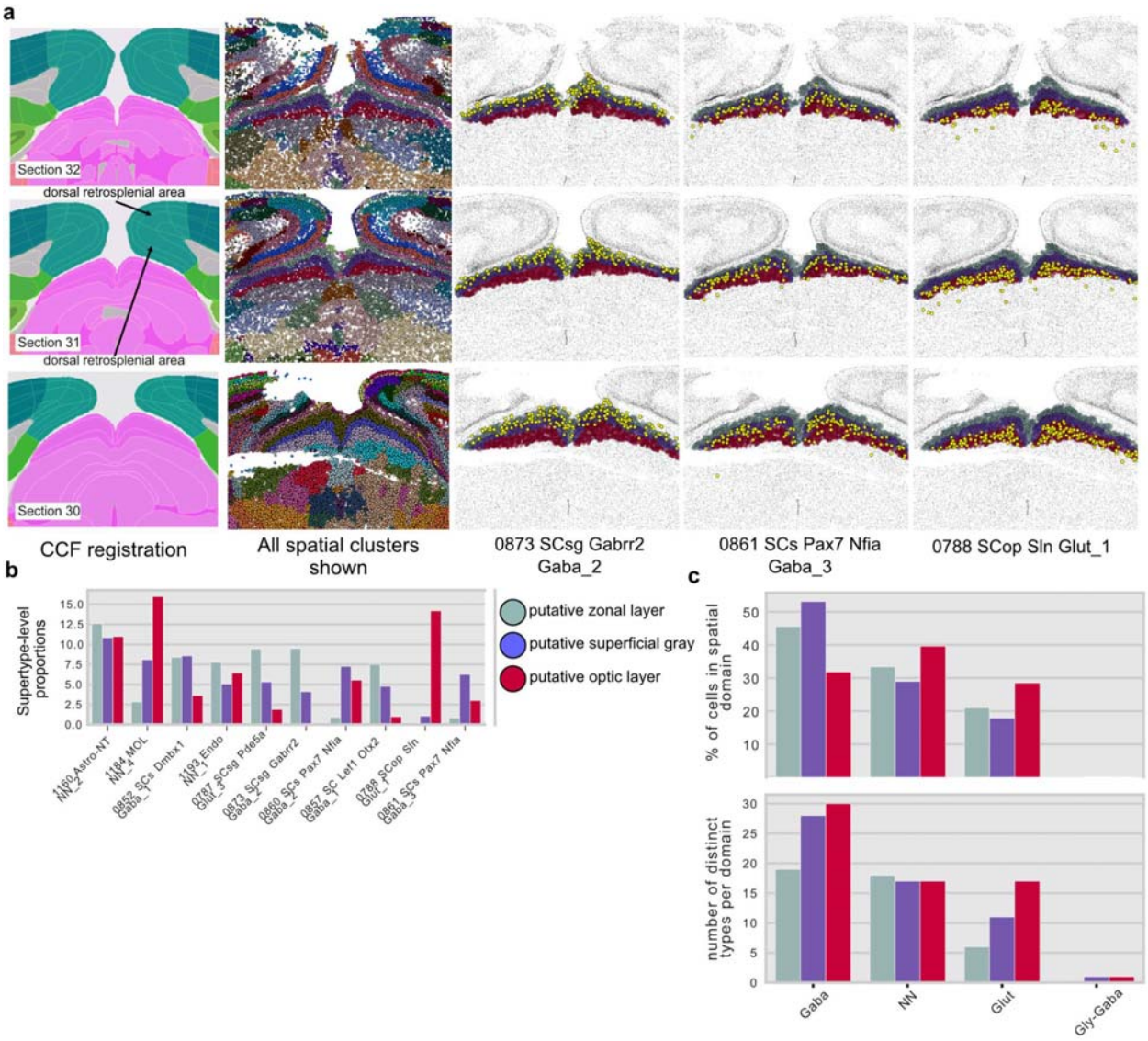
Supplementary Figure 5. Roughly half of the sections in the Allen Institute Brain Science mouse, approximately every other section. For each section color labels come from CellCharter's Gaussian mixture model implementation with *k*=670 clusters and the cartoon image to the right is the CCF registered region set.



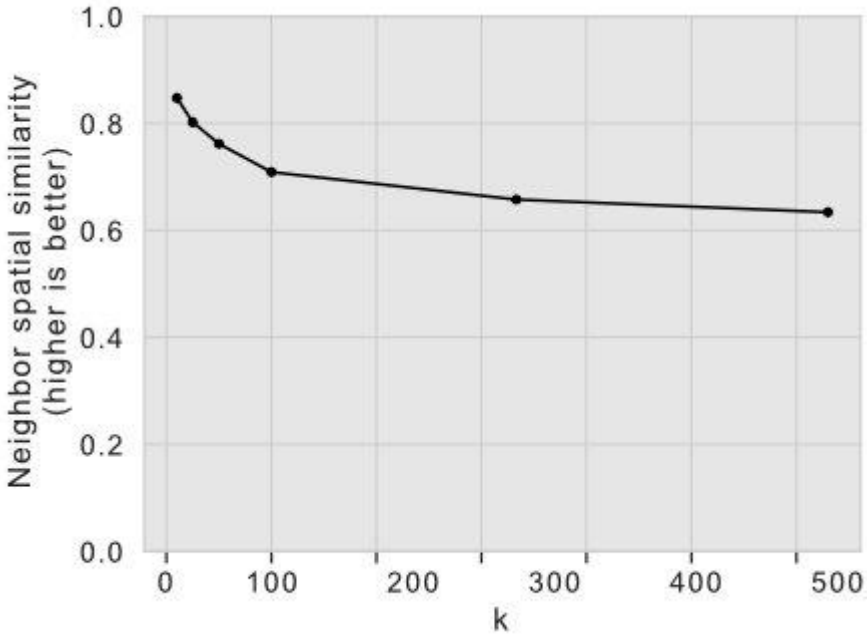
Supplementary Figure 6. Comparison of spatial domains in midbrain. Left column shows approximate CCF registration. Middle column CellTransformer at *k*=1300 and the right column shows CellCharter with 670 Gaussians. The general performance in outlining cortical layers is similar, however in the midbrain even at half the number of clusters, CellCharter loses spatial coherence compared with CellTransformer. In particular CellCharter also only identifies two layers of superior colliculus, whereas multiple are present in CellTransformer domains.



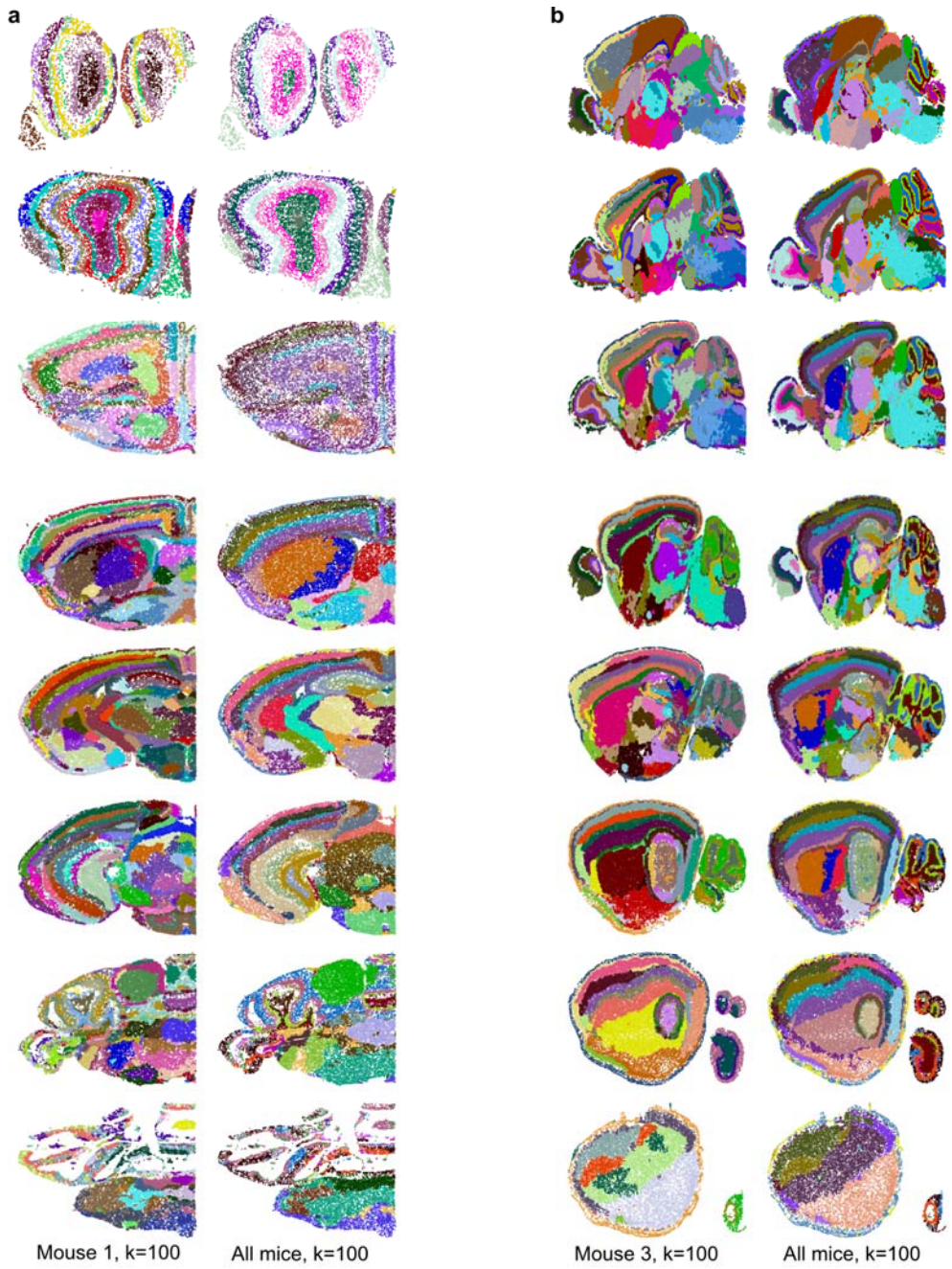
Supplementary Figure 7. Right: instability scores (see **Methods**) calculated for different clustering solutions. Error bars are standard deviation. Left: inertia (sum of squares errors for each cluster centroid) calculated for different clustering solutions. Error bars (standard deviation) are calculated but not visible due to scale.



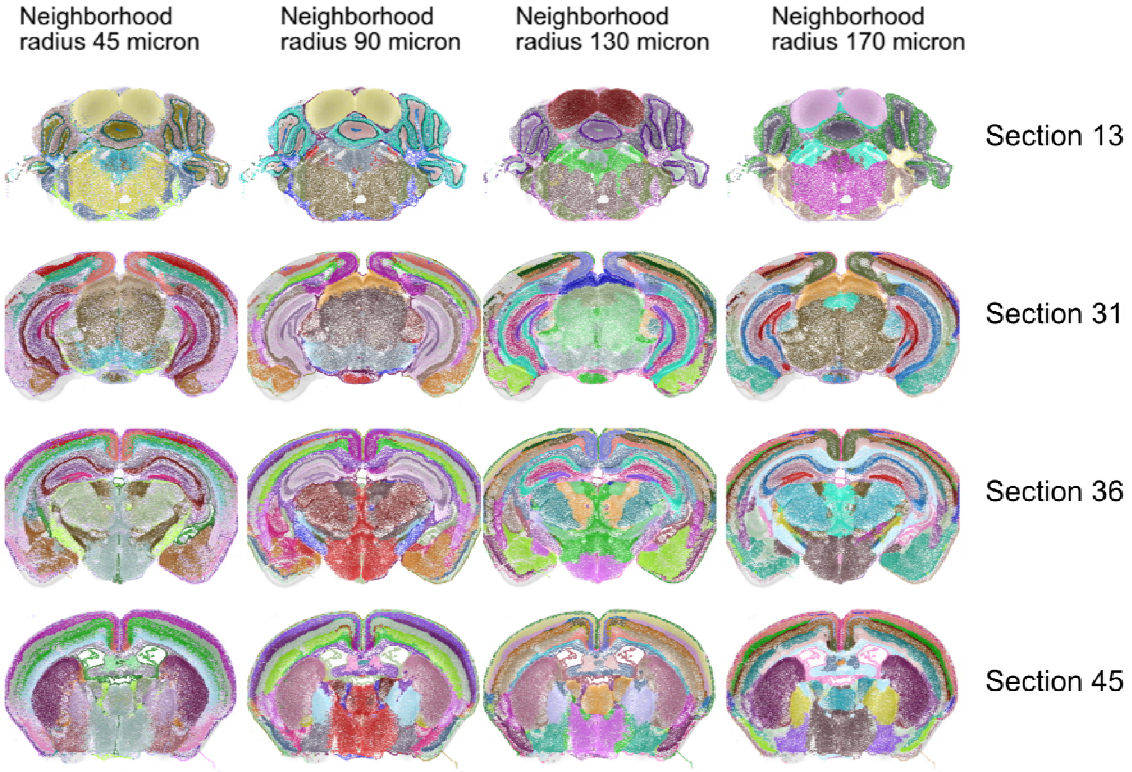
Supplementary Figure 8. Comparison of best fit spatial domains from CellTransformer with layers of the superior colliculus, sensory related area. (**a.**) Sequential tissue sections (32, 31, 30, from anterior to posterior) showing in first column CCF registration and borders of relevant areas. Second column: all cells in field of view, colored by spatial domain from CellTransformer. Third column: only visualizing cells inside our putative matches for the zonal, superficial gray, and optic layers in the superior colliculus. Also visualized is the 0879 SCsg Pde5a Glut_1 cell type (supertype-level) in yellow. Fourth column: same as third but visualizing the 0865 SCs Pax7 Nfia Gaba_3 cell type. Fifth column: same as third and fourth but visualizing the 0882 SCop Sln Glut type. (**b.**) Bar chart of cell type abundance (as a percentage) for top ten most abundant types across the putative subregions. (**c.**) Bar chart of per-region percentages of GABA and glutamatergic neurons and non-neuronal types.



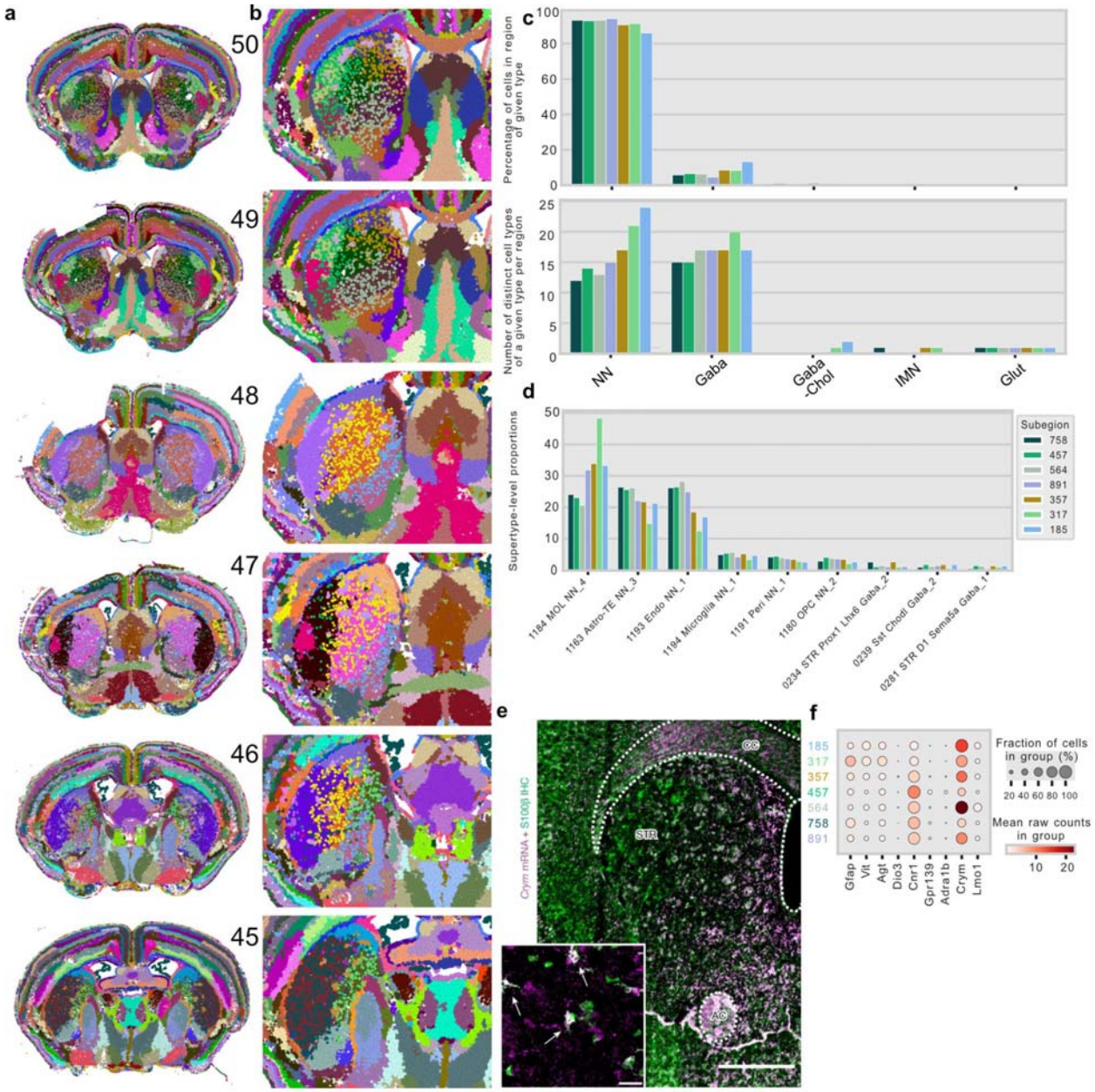
Supplementary Figure 9. Results of quantitative comparison of CCF regions and CellTransformer regions at equivalent number of regions. (**a.**) Spatial smoothness of spatial clusters as measured using a nearest-neighbors approach, computed by clustering the concatenated latent variables for neighborhoods in the Zhuang lab datasets. (**b.**) Result of comparing clustering across all animals versus individually. Solid lines indicate the match of computed spatial clusters with CCF when spatial clusters are computed at the per-animal level. Dashed area indicates standard deviation. For the analysis comparing the concatenated dataset clustering solution, we extract cell type by region matrices per animal.



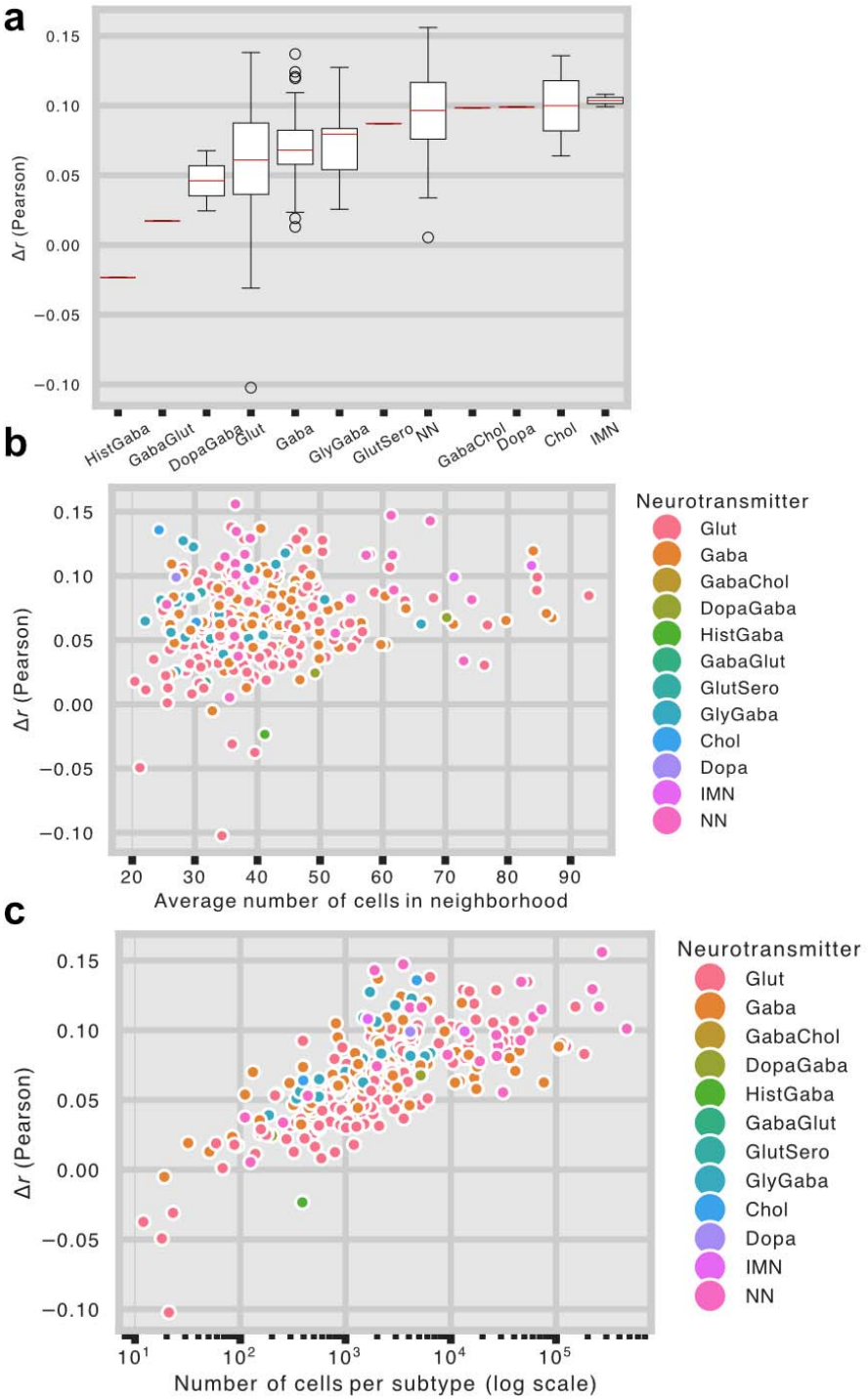
Supplementary Figure 10. Representative images of a spatial clustering solution run per-mouse on animals 1 and 3. (**a.**) Clustering results for Zhuang lab animal 1, comparing (left column) clustering with only one animal of embeddings and with all mice together (right column). (**b.**) Clustering results for Zhuang lab animal 3, comparing (left column) clustering with only one animal of embeddings and with all mice together (right column).



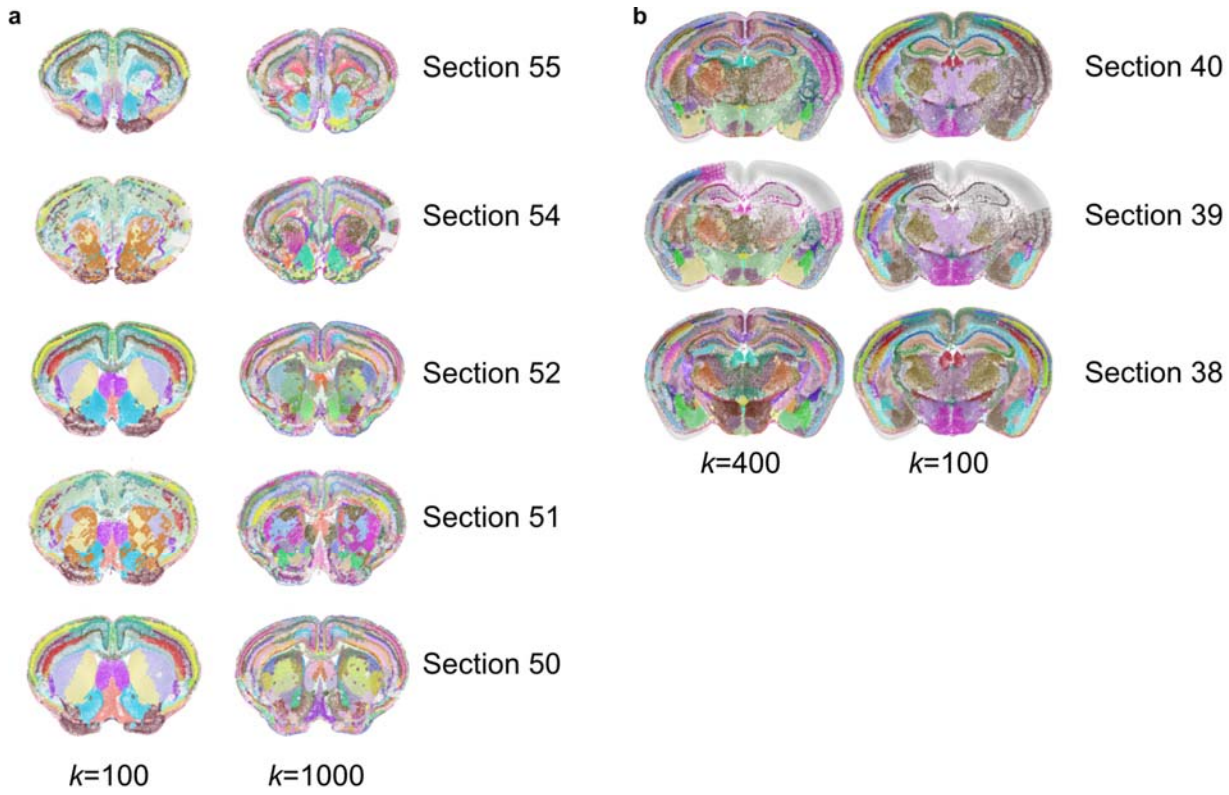
Supplementary Figure 11. Representative images of spatial clustering from CellTransformer models trained with different neighborhood radius parameters.



Supplementary Figure 12. Representative images of spatial clustering from CellTransformer models with *k*=1300. (**a.**) Sequential tissue sections (50 is most anterior) showing smoothness of spatial domains across and within tissue sections as well as consistent appearance of an irregular spatial pattern inside caudoputamen. (**b.**) Zoom in on the striatum for the same tissue sections. (**c.**) Plots showing percentage of cell types of different neurotransmitter for the non-uniform spatial clusters as well as the distribution of unique cell types of a given neurotransmitter type. (**d.**)



Supplementary Figure 13. (**a.**) Increase in Pearson correlation from per-cell type (at subclass level) average expression. Red lines show medians per distribution. (**b**.) Scatterplot of increase in average Pearson correlation per subclass level cell type against average neighborhood size for reference cells of that type. **c**. Scatterplot of increase in average Pearson correlation per subclass level cell type vs the number of cells of that type in log scale.



Supplementary Figure 14. Representative failure mode images. (**a.**) Section 54 consistently has a tear-like artifact; see preceding tissue section (55), which shows that similar clusters are discovered but are dominated by a noisy pattern. Section 51 has a similar phenomenon. (**b.**) Left-right asymmetry in section 39 and 40.