Multimillion cell self-supervised representation learning enables organ-scale tissue niche discovery

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

Abstract

Spatial transcriptomics (ST) offers unique opportunities to define the spatial organization of tissues and organs, such as the mouse brain. We establish a workflow for self-supervised spatial domain detection that is scalable to multimillion-cell datasets and analysis of organ-scale ST datasets. This workflow uses a self-supervised framework for learning latent representations of tissue niches. We use a novel encoder-decoder architecture, which we named CellTransformer, to hierarchically learn higher-order tissue features from lower-level cellular and molecular statistical patterns. CellTransformer is effective at integrating cells across tissue sections, identifying domains highly similar to ones in existing ontologies such as Allen Mouse Brain Common Coordinate Framework (CCF) while allowing discovery of hundreds of uncataloged areas with minimal loss of domain spatial coherence. CellTransformer advances the state of the art for spatial transcriptomics by providing a performant solution for the detection of fine-grained tissue domains from spatial transcriptomics data.

1. Introduction

Hierarchical spatial organization is ubiquitous in tissue and organ biology. Systematic, high-quality and high-dimensional phenotypic measurements of this organization via experimental tools such as spatial transcriptomics, multiplex immunofluorescence, and electron microscopy, are becoming available as large, open datasets. However, transforming this complex data into a useful representation can be difficult, even for fields with a wealth of prior knowledge, such as neuroanatomy.

Datasets such as the Allen Brain Cell Mouse Whole Brain (ABC-MWB) Atlas (Yao et al., 2023; Zhang et al., 2023), a multi-million cell single-cell RNA sequencing (scRNA-seq) and spatial (MERFISH) atlas, provide unprecedented opportunities 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 existing methods operate on large intermediate data structures such as pairwise distance matrices (Haviv et al., 2024; Zhou et al., 2023; Hu et al., 2021) precluding scaling into the millions or tens of millions.

We implement representation learning and clustering workflow focusing on cellular subgraphs. Our model learns to condition cell-type specific gene expression predictions using this neighborhood context token. The model thus learns to predict expression of cell types in arbitrary cell neighborhoods. We show representation allows for recovery of important anatomically plausible spatial domains while remaining computationally efficient.

We evaluate our pipeline, CellTransformer, on using the ABC-MWB dataset (3.9 million cells collected with a 500 gene MERFISH panel) (Yao et al., 2023) demonstrating its effectiveness in producing completely data-driven spatial domains of the mouse brain by comparing the results to the Allen Mouse Brain Common Coordinate Framework version 3 (CCFv3) (Wang et al., 2020). CCF is a consensus hand-drawn 3D reference space compiled from a large multimodal data corpus. Annotations feature labels at three levels of coarseness (from 25 regions at coarse-grain to 670 at fine-grain). Compared with other methods, CellTransformer excels at identifying spatial domains which are spatially coherent and biologically relevant. CellTransformer domains also reproduce known regional architecture.

[1]Department of Neurology, University of California, San Francisco, San Francisco, USA [2]Allen Institute for Brain Science, Seattle, USA [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, USA. Correspondence to: Reza Abbasi-Asl <Reza.AbbasiAsl@ucsf.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

2. Results

2.1. The CellTransformer architecture and domain detection workflow

CellTransformer is a graph transformer that we train to learn latent representations of single cell neighborhoods or subgraphs. The model is trained to condition masked expression predictions on tissue context. We restrict this graph to a small neighborhood subgraph of the whole-tissue-section graph by only modeling cells within a user-specified distance from a given center cell, which we refer to as a reference cell. We model each cell within this neighborhood as a token. Truncating neighborhoods using a fixed spatial threshold allows the network to account for the varying density of cells in space. Consequently, our framework incorporates two main facets of neuroanatomical classification, incorporating both cytoarchitecture (relative density and proximity) and molecular variation (cell type and RNA-level variation) in the data.

Cell tokens are generated by composing cell-type and gene expression information. Representations are then refined by a vanilla transformer encoder. Tokens within each neighborhood are then aggregated using a learned pooling operation to produce a single token representation of the entire tissue context. The model receives a new mask token representing the reference cell-type to predict its gene expression following the operation of several transformer decoder layers (Supplementary Figure S.1b). Importantly, during this process, only the mask token and the neighborhood representation can attend to each other. This operation captures a hierarchical encoding and decoding process where low level information (gene and cell type) is produced at the cell token level and aggregated into a high-level representation. This high-level representation is then used to conduct the reverse decoding process (prediction of gene expression from cell type and tissue context information).

At test time, we extract this neighborhood representation for each cell, concatenating across sections or animals, and use k-means clustering to identify discrete spatial domains (Supplementary Figure S.1c). We will use the term spatial domain to refer to the output of clustering on embeddings and cluster to refer to single-cell clusters transferred from the ABC-WMB single cell taxonomy.

2.2. Data-driven discovery of fine-grained spatial domains in the mouse brain using ABC-WMB

The ABC-WMB spatial transcriptomics dataset contains data from five mouse brains. One animal was processed by the Allen Institute for Brain Science with a 500 gene MERFISH panel and 53 coronal sections (Yao et al, 2023)(Yao et al., 2023) The remaining four other animals, generated in Zhang et al. (2023)(Zhang et al., 2023) were collected with a 1129 gene panel. Sections from two of these animals ("Zhuang 1", 147 sections; and "Zhuang 2", 66 sections) were sampled coronally. The other two animals in the dataset ("Zhuang 3", 23 sections; and "Zhuang 4", 3 sections) were sampled sagittally.

We first trained CellTransformer on the Allen 1 dataset, subsequently extracting embeddings for each cell's neighborhood, which we defined as a set of cells within a fixed size square around that cell. We then clustered these embeddings using k-means. At higher numbers of domain (higher k), we observed isolated cases with non-contiguous domains in the striatum and no where else in the brain. These domains appear to be biologically plausible and resemble those identified in (Ollivier et al., 2024). However, neuroanatomical standard is to study spatially contiguous regions. Therefore we also optionally introduced a smoothing step prior to k-means, which we applied to spatially smooth the embeddings. See Supplementary Note 1 for a discussion on the effects of smoothing on detected domains and of the striatal domains.

We generated domains at k=25, 354, and 670, to match the division, structure, and substructure annotations in CCFv3, displaying domains for four consecutive tissue sections (Figure 1). We also provide representative images of spatial clusters across the brain (28/53 sections) at different k in Supplementary Figures figs. S.2 to S.4. Low domain numbers such as k=25 broadly divide the brain into neuroanatomically plausible patterns, with subregions of striatum (dorsal and ventral marked in Figure 1) and cortical layers clearly visible. A comparison of cortical layers across these sections shows that CellTransformer domains at k=25 are well matched to CCF (Supplementary Figure S.5) and correctly identify major classes of layers (1, 2/3 4, 5, and 6) across somatosensory and somatomotor cortex. In particular, we point out the excellent correspondence of domains across tissue sections at k=25 across the entire dataset (Supplementary Figure S.2), with nearly perfect consistency across regions.

At k=354, anterior-posterior subdivisions emerge such as the presence of layer 4 in the motor cortex(Yao et al., 2023) (Figure S.5). Historically, the mouse motor cortex was thought to lack a granular layer 4, however recently, MERFISH, transcriptomic and epigenomic studies have confirmed its existence1,15,16. At k=100 and k=354, we find a domain corresponding to Layer 4 in the somatosensory cortex which clearly extends to layer 4 in the motor cortex.

At k=670, the cortical layers identified at lower resolution are further partitioned into superficial, intermediate, and deep strata within several layers. We visualize cortical layers across sections in depth (Figure 1b), showing CellTransformer not only identifies fine superficial-deep structure within cortical layers but also preserves the boundary be-

Table 1. Model performance (Allen 1 dataset). CCF labels are human generated; CT is CellTransformer. Bold indicates best performance.

METRIC	k	CT	CT (-SMOOTH)	CELLCHARTER	SPIRAL	K-MEANS	CCF
NORMALIZED MUTUAL INFORMATION (NMI)	25	**0.540**	**0.540**	0.304	0.365	0.392	–
	354	0.609	**0.609**	0.527	0.407	0.481	–
	670	0.642	**0.643**	0.566	0.438	0.525	–
ADJUSTED RAND INDEX (ARI)	25	**0.280**	0.268	0.111	0.181	0.074	–
	354	**0.134**	0.129	0.106	0.080	0.102	–
	670	**0.101**	0.100	0.082	0.055	0.089	–
MAX COMPOSITION SIMILARITY	25	0.741	**0.843**	0.752	0.726	0.747	–
	354	0.893	**0.895**	0.819	0.767	0.726	–
	670	**0.949**	0.944	0.840	0.829	0.746	–
SPATIAL SMOOTHNESS	25	**0.822**	0.819	0.798	0.077	0.826	0.899
	354	0.675	**0.680**	0.468	0.020	0.477	0.819
	670	0.652	**0.656**	0.412	0.016	0.402	0.743
PROPORTION DISCRETE	25	0.920	**1.0**	0.880	0.240	**1.0**	–
	354	0.698	**0.992**	0.147	0.011	0.0	–
	670	0.667	**0.964**	0.085	0.009	0.0	–

tween somatosensory and motor cortex (marked in thick black dotted lines in Figure 1b). Taken together these results showed that CellTransformer robustly describes previously known anatomical structures.

We compared CellTransformer to several other workflows to capture spatial coherency and multiresolution neuroanatomical annotations in CCF at the division, structure, and substructure levels. For comparison, we used two recent methods, CellCharter (Varrone et al., 2024) and SPIRAL (Guo et al., 2023) that are scalable to millions of cells as benchmarks. Additionally, we implemented a machine learning baseline that employs k-means clustering on cellular neighborhoods (represented as cell type count vectors). To quantify the relative effect of smoothing on CellTransformer results we report both unsmoothed and smoothed metrics.

We first computed the normalized mutual information and adjusted rand indices relative to the CCF label. CellTransformer performs better than alternatives at every resolution. To further quantify the similarity of detected domains with CCF annotations, we compared the cell type composition of domains using cell type calls from the ABC-WMB taxonomy. We chose the subclass cell type level, extracting for each domain and for each method a 338-long cell-type vector. We calculated the Pearson correlation of cell type composition vectors computed using the CCF regional annotations at division (25), structure (354) and substructure (670) levels against those of the various methods at the corresponding number of spatial domains. First, for each data-driven domain, we computed the maximum correlation to any CCF domain at the same CCF annotation resolution averaging these numbers across domains. CellTransformer outperforms other methods at mid-granularity and fine-granularity (Table 1).

To quantify the spatial coherence of domains, for each cell we identified its nearest 100 spatial neighbor cells. We then quantified the proportion of neighbor cells within the same spatial domain label as the starting cell (Table 1). CellTransformer outperforms CellCharter (58.2% better spatial coherence at 670 domains) and SPIRAL (4091.2%). CellTransformer also outperforms the machine learning baseline based on k-means clustering on cellular neighborhoods (61.9% better spatial coherence). For reference, we include the CCF parcellation as an upper bound of hand-drawn, human generated annotation. From these hand drawn annotations, we computed a cutoff to categorize data-driven domains as spatially discrete. We first averaged the spatial smoothness across each CCF domain at k=670 and then computed the median smoothness (0.633). We then repeated the same process for each of the data driven domains, using the same cutoff. CellTransformer produces significantly more discrete regions than comparator methods. Interestingly, the unsmoothed embeddings perform better on average, which we attribute to the fact that our smoothing procedure is isotropic and may not respect fine laminar structure in the brain.

2.3. Multi-animal analysis

In order to investigate CellTransformer's ability to integrate across animals, we trained a new model from scratch on the (Zhang et al., 2023) MERFISH data, which uses an 1129 gene panel and is split over four animals, with both coronal (Zhuang 1 and 2) and sagittal sections (Zhuang 3 and 4). We computed embeddings for each neighborhood as in the previous analysis and performed k-means, concatenating representations for all mice and sections. Spatial domains in sequential tissue sections appeared highly concordant across

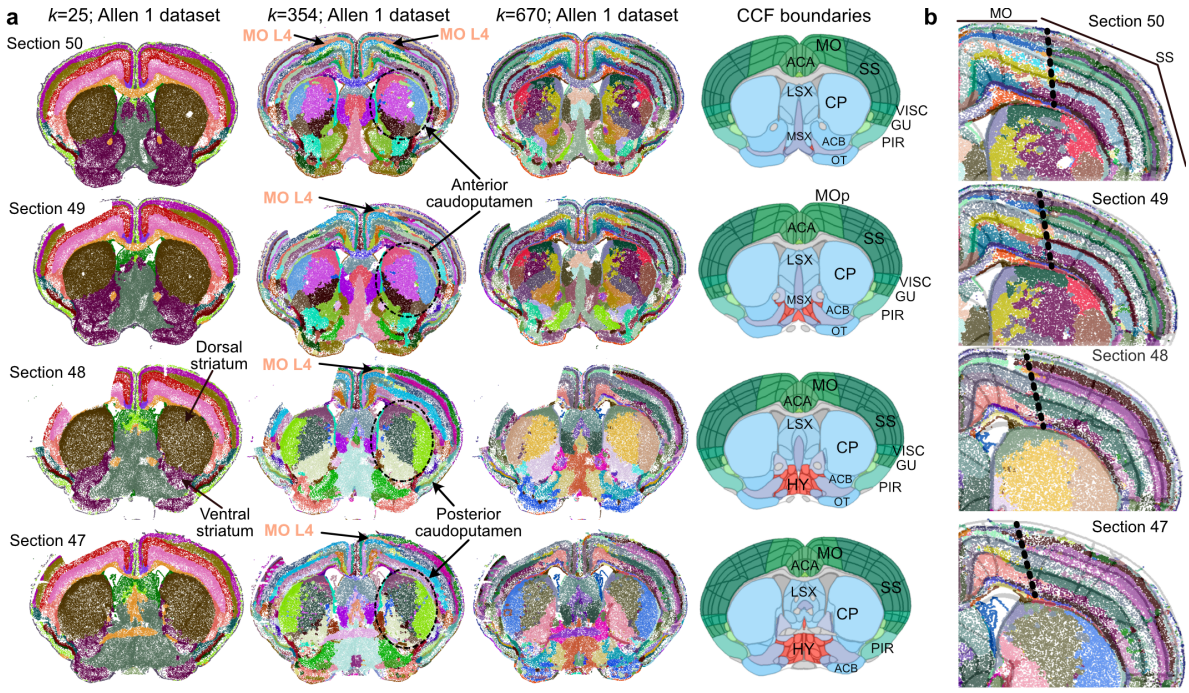

Figure 1. Representative images of spatial domains discovered using CellTransformer on the Allen 1 dataset (53 coronal sections and 500 gene MERFISH panel1) and comparison to CCF. (**a**.) Comparison of automated domain discovery at three resolutions (first three columns) and CCF (last). Each dot is a cell and within columns are colored by *k*-means cluster. (**b**.) Single hemisphere images of same tissue sections in **a.** domains fit at *k*=670, zoomed in on cortical layers of motor cortex (MO) and somatosensory cortex (SS). CCF boundaries are shown in semi-transparent lines, with the boundary between SS and MO outlined in larger black dotted lines.

all four mice (Supplementary Figure S.6) at the 50-domain resolution. We used 50 domains to facilitate clear visualization of the domains across animals with relatively few colors. Coronal and sagittal sections across mice clearly corresponded anatomically. Cortical layers were highly consistent across animal and section orientation. Structures that appear in the coronal view can be readily identified in the sagittal sections. For example the hippocampal formation (blue) is well delineated in sections 088 for Zhuang 1, section 044 for Zhuang 2, and across displayed sections of Zhuang 3 and Zhuang 4. Despite a relatively low number of cells in mouse 4 (162,579 cells versus more than 1.5 million for each of the other animals), nearly all spatial domains observed for Zhuang 4 are present in other animals.

We quantified the robustness of CellTransformer domains in a multi-animal context across and within Zhuang 1-4 datasets. We ran clustering and identified domains at the three values of k: 25, 333, 630. These k values correspond to three CCF resolution levels reported by registration in (Zhang et al., 2023) (note the number of domains differs due to registration differences). For each k value, we counted the number of domains observed in all four animals. We also repeated this analysis without data for Zhuang 4, which contains far fewer cells than the datasets from other animals (Figure 6b). We find that even at high resolution (630 domains), 93.3% domains were found in each mouse, showing high consistency of CellTransformer domains across datasets. With the Zhuang 4 included, at 630 domains, 80.0% domains were found in every animal.

Discussion

This work demonstrates that a self-supervised, cellular neighborhood based strategy for tissue niche can be used to replicate fine-grained human annotations in large spatial transcriptomics datasets. The representations learned in our model can be clustered to identify finer-scale spatial domains directly from local cellular and molecular information alone, but can produce biologically plausible domains. These domains are also spatially consistent both within and across tissue sections and even over multiple animals.

Impact Statement

Our work addresses a critical bottleneck in spatial transcriptomics analysis that has limited organ-scale tissue niche discovery. We enable data-driven domain detection at resolutions finer than CCF, potentially identifying previously

unidentified brain areas. This capability may contribute to the acceleration systematic neuroanatomical studies of the mammalian brain in both health and disease and has implications for a variety of applications where spatial cellular relationships are under study. Our method also may be used to study individual-level variation in neuroanatomy at the cellular level, which historically has been difficult to study.

References

Guo, T., Yuan, Z., Pan, Y., Wang, J., Chen, F., Zhang, M. Q., and Li, X. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biol.*, 24(1):241, October 2023.

Haviv, D., Remšík, J., Gatie, M., Snopkowski, C., Takizawa, M., Pereira, N., Bashkin, J., Jovanovich, S., Nawy, T., Chaligne, R., Boire, A., Hadjantonakis, A.-K., and Pe'er, D. The covariance environment defines cellular niches for spatial inference. *Nat. Biotechnol.*, pp. 1–12, April 2024.

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., and Li, M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods*, 18(11): 1342–1351, November 2021.

Ollivier, M., Soto, J. S., Linker, K. E., Moye, S. L., Jami-Alahmadi, Y., Jones, A. E., Divakaruni, A. S., Kawaguchi, R., Wohlschlegel, J. A., and Khakh, B. S. Crym-positive striatal astrocytes gate perseverative behaviour. *Nature*, 627(8003):358–366, March 2024.

Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A., and Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nat. Genet.*, 56(1):74–84, January 2024.

Wang, Q., Ding, S.-L., Li, Y., Royall, J., Feng, D., Lesnar, P., Graddis, N., Naeemi, M., Facer, B., Ho, A., Dolbeare, T., Blanchard, B., Dee, N., Wakeman, W., Hirokawa, K. E., Szafer, A., Sunkin, S. M., Oh, S. W., Bernard, A., Phillips, J. W., Hawrylycz, M., Koch, C., Zeng, H., Harris, J. A., and Ng, L. The allen mouse brain common coordinate framework: A 3D reference atlas. *Cell*, 181 (4):936–953.e20, May 2020.

Yao, Z., van Velthoven, C. T. J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., Baker, K., Baker, P., Barkan, E., Bertagnolli, D., Bhandiwad, A., Bielstein, C., Bishwakarma, P., Campos, J., Carey, D., Casper, T., Chakka, A. B., Chakrabarty, R., Chavan, S., Chen, M., Clark, M., Close, J., Crichton, K., Daniel, S., DiValentin, P., Dolbeare, T., Ellingwood, L., Fiabane, E., Fliss, T., Gee, J., Gerstenberger, J., Glandon, A., Gloe, J., Gould, J., Gray, J., Guilford, N., Guzman, J., Hirschstein, D., Ho, W., Hooper, M., Huang, M., Hupp, M., Jin, K., Kroll, M., Lathia, K., Leon, A., Li, S., Long, B., Madigan, Z., Malloy, J., Malone, J., Maltzer, Z., Martin, N., McCue, R., McGinty, R., Mei, N., Melchor, J., Meyerdierks, E., Mollenkopf, T., Moonsman, S., Nguyen, T. N., Otto, S., Pham, T., Rimorin, C., Ruiz, A., Sanchez, R., Sawyer, L., Shapovalova, N., Shepard, N., Slaughterbeck, C., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Valera Cuevas, N., Vance, S., Wadhwani, K., Ward, K., Levi, B., Farrell, C., Young, R., Staats, B., Wang, M.-Q. M., Thompson, C. L., Mufti, S., Pagan, C. M., Kruse, L., Dee, N., Sunkin, S. M., Esposito, L., Hawrylycz, M. J., Waters, J., Ng, L., Smith, K., Tasic, B., Zhuang, X., and Zeng, H. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624 (7991):317–332, December 2023.

Zhang, M., Pan, X., Jung, W., Halpern, A. R., Eichhorn, S. W., Lei, Z., Cohen, L., Smith, K. A., Tasic, B., Yao, Z., Zeng, H., and Zhuang, X. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624(7991):343–354, December 2023.

Zhou, X., Dong, K., and Zhang, S. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nat. Comput. Sci.*, 3(10):894–906, October 2023.

A. Appendix.

A.1. Supplementary Methods

A.1.1. Data preprocessing

Allen Institute for Brain Science dataset preprocessing We downloaded the log-transformed MERFISH probe counts and metadata for the Allen Institute for Brain Science animal ("Allen 1") from the Allen Institute public release (https://alleninstitute.github.io/abc_atlas_access/intro.html) access for ABC-MWB.The Allen 1 dataset is composed of 53 coronal sections. The MERFISH probe set included 500 genes. Serial sections were collected at 200 μm intervals. We used the taxonomy from the "20231215" data release. Allen 1 is composed of 3,737,550 cells. 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.

For the Zhuang datasets, data were downloaded from the "20230830" data release from the Allen Institute ABC-MWB public data release. Two animals ("Zhuang 1" and "Zhuang 2") were collected with coronal sections. The other two animals ("Zhuang 3" and "Zhuang 4") were collected sagittally. Serial sections for Zhuang 1 (female) were collected at 100 μtextm intervals, while serial sections for the other animals (all male) were collected at 200 μm intervals. The size of the MERFISH probe set included 1129 genes. Zhuang 1 and Zhuang 2 consist of 2,846,909 cells and 1,227,409 cells, respectively. Zhuang 3 and Zhuang 4 consist of 1,585,844 cells and 162,579 cells, respectively. We transformed the (x, y) coordinates of each cell into microns instead of mm as provided. Otherwise, the data were used as-is for neural network training.

A.1.2. Cellular neighborhood construction

We consider cells in the same neighborhood as a reference cell if the distance between them is within a box of fixed size. For all MERFISH datasets we used a box width of 85 μm.

A.1.3. 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 to perform this operation on GPU (`cuml.KMeans`), with arguments `n_init=3`, `oversampling_factor=3`, and `max_iter=1000`.

A.1.4. 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, smoothing was performed with a Gaussian filter with 40 micron full-width at half maxima (sigma of 12.01 microns).

A.1.5. Model fitting on the Allen 1 dataset

We used an 80%-20% train-test split proportion (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.

A.1.6. Model fitting on Zhuang datasets

We perform training from scratch without transfer. We trained for 40 epochs with the same settings as for the Allen 1 with the exception of adapting projections to 1129 genes instead of 500.

A.1.7. 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. For each region derived from one of the tested models, we compute two quantities: the best match (maximum value of Pearson correlation, non-exclusively) to any CCF (Figure 2d) or an exclusive match (using the linear sum assignment algorithm) to pair the regions from either set one-to-one (Figure 2e). We then

computed the average Pearson correlation across the paired matches as the metric. We use scipy.optimize's implementation to solve the linear sum problem.

A.1.8. 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 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 the text.

A.1.9. SPIRAL

To run SPIRAL we generated edge sets for 40μm, 85μm, and 170μm neighborhood radii. SPIRAL requires supervision on single-cell types so for this we use the subclass cell type levels. We trained models across neighborhood sizes for 1 epoch and then chose the neighborhood size with best performance (170μm) and trained this model to saturation (10 epochs). SPIRAL uses four objective functions so to assess saturation we averaged them. We note that SPIRAL does not use a training and testing set split in their training, making it difficult to assess an optimal stopping point. For the k=354 and k=670 domain discovery analyses the SPIRAL clustering pipeline produced an out-of-memory error and we instead used our own pipeline with k-means on SPIRAL embeddings.

A.1.10. 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.

A.1.11. Zhuang lab dataset per-animal CCF comparison

We contrast two methods of extracting spatial domains from the four animals in the Zhuang lab dataset1. 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 four (Zhuang 1, 2, 3, and 4) mice together. We also fit a k-means model to the embeddings 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 their regional cell type composition vectors.

A.2. Supplementary Note 1

When scaling up the number of regions past 500 in the Allen 1 dataset, we observed that almost all spatial clusters were spatially smooth except for a recurring pattern in the striatum. We plot (Supplementary Figure S.7) 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 S.7) forming a large proportion of cells.

We sought to understand whether these spatial clusters might be biologically relevant or somehow related to noise. A recent paper, Ollivier et al. (2024) 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 11e 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 may have biological relevance.

However, to simplify downstream analyses and conform with neuroanatomical conventions we applied a simple smoothing operation (see Supplementary Methods), which removed this spatial cluster in successive clustering operations. We used a very small smoothing window (12 micron sigma, or 40 micron full-width at half-maxima) and found the order of ranked methods and their relative performance changes were not significantly affected.

A.3. Supplementary Figures

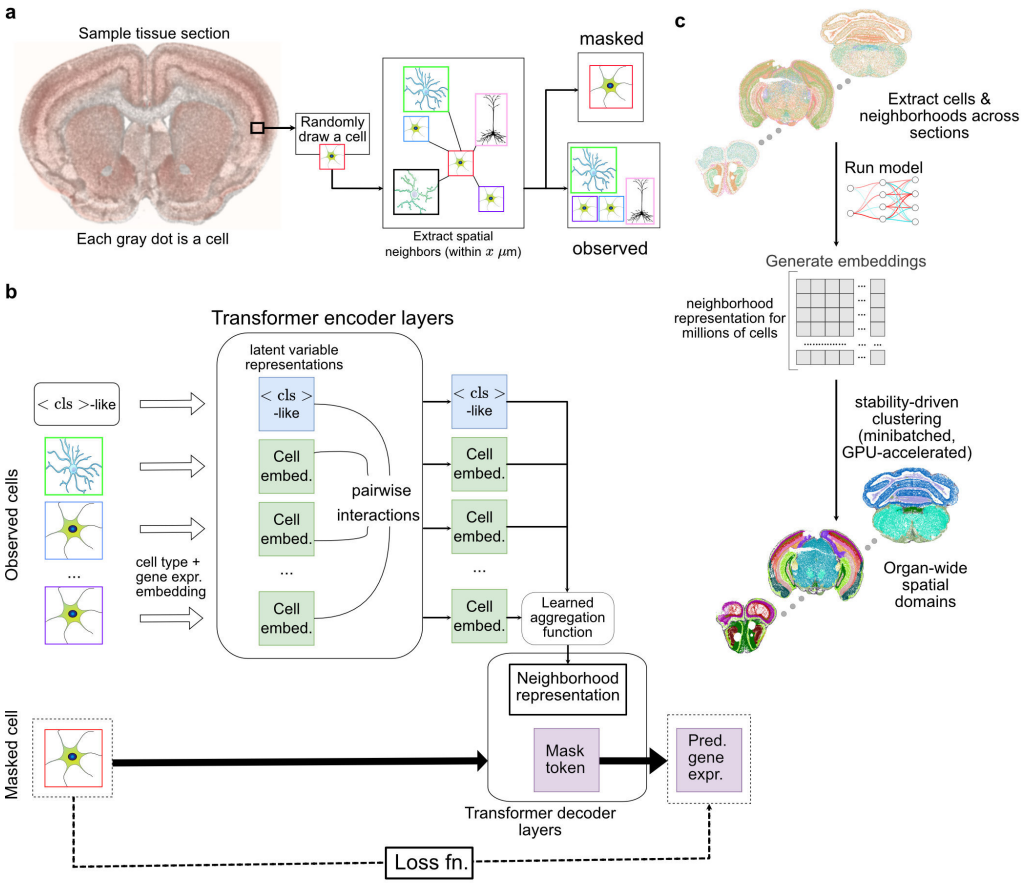

Figure S.1. Overall training and architectural scheme for CellTransformer. (**a**.) During training, a single cell is drawn (we denote this the reference cell, highlighted in red). We extract the reference cell's spatial neighbors and partition the group into a masked reference cell and its observed 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. These per-cell representations and an extra token acting as a register 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. (**c**.) Once the model is trained, we compute embeddings (one for each neighborhood/reference-cell pairing) and concatenate these embeddings within the tissue section datasets and across tissue sections. Concatenating embeddings across tissue sections produces region discovery at organ level. We then cluster these embeddings using k-means to discover tissue domains across sections.

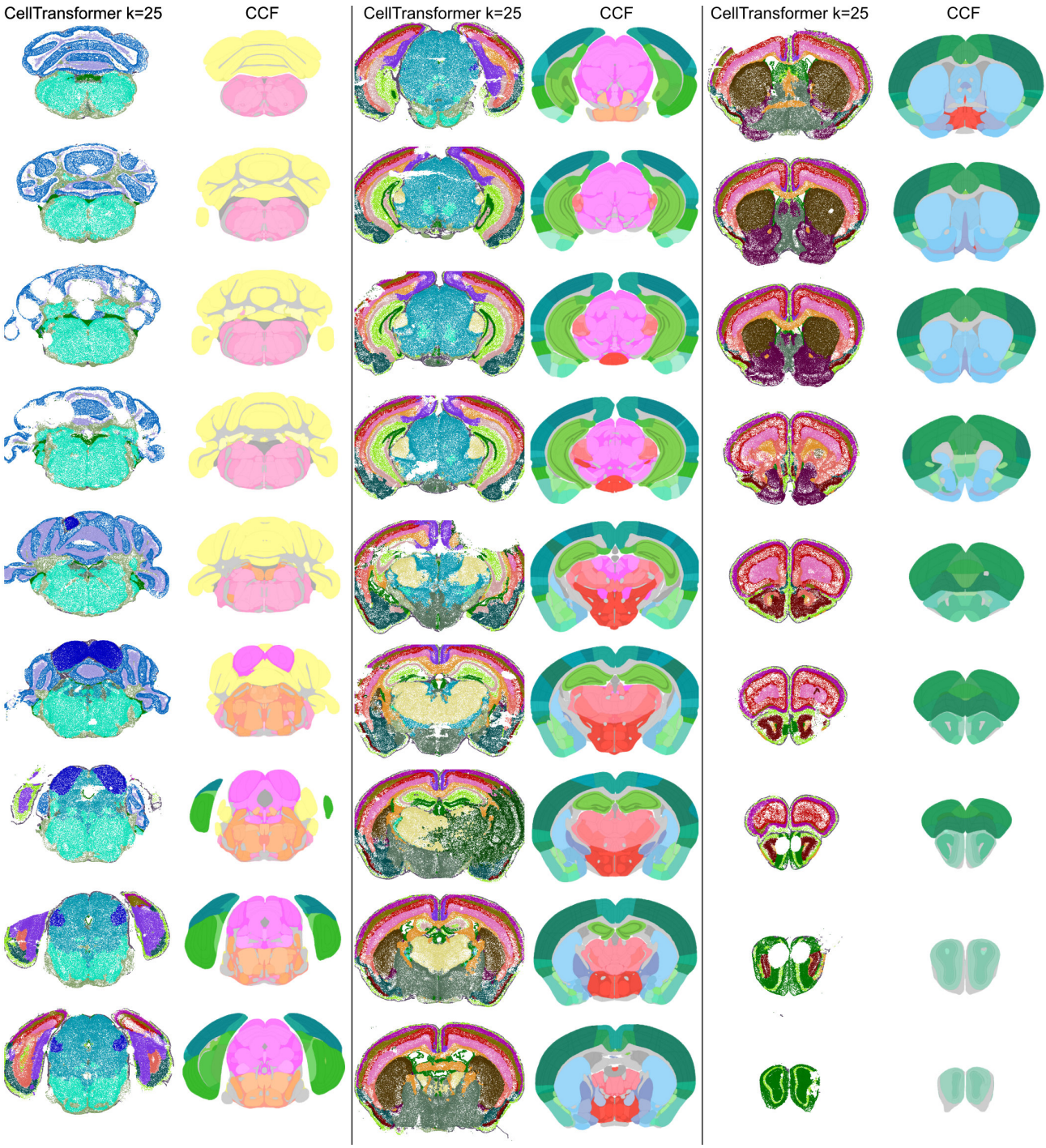

Figure S.2. CellTransformer spatial domains (left) and the corresponding CCF annotations (right) organized in 3 columns for roughly half of the sections in the Allen 1 dataset, approximately every other section. CellTransformer domains were calculated at k=25 clusters.

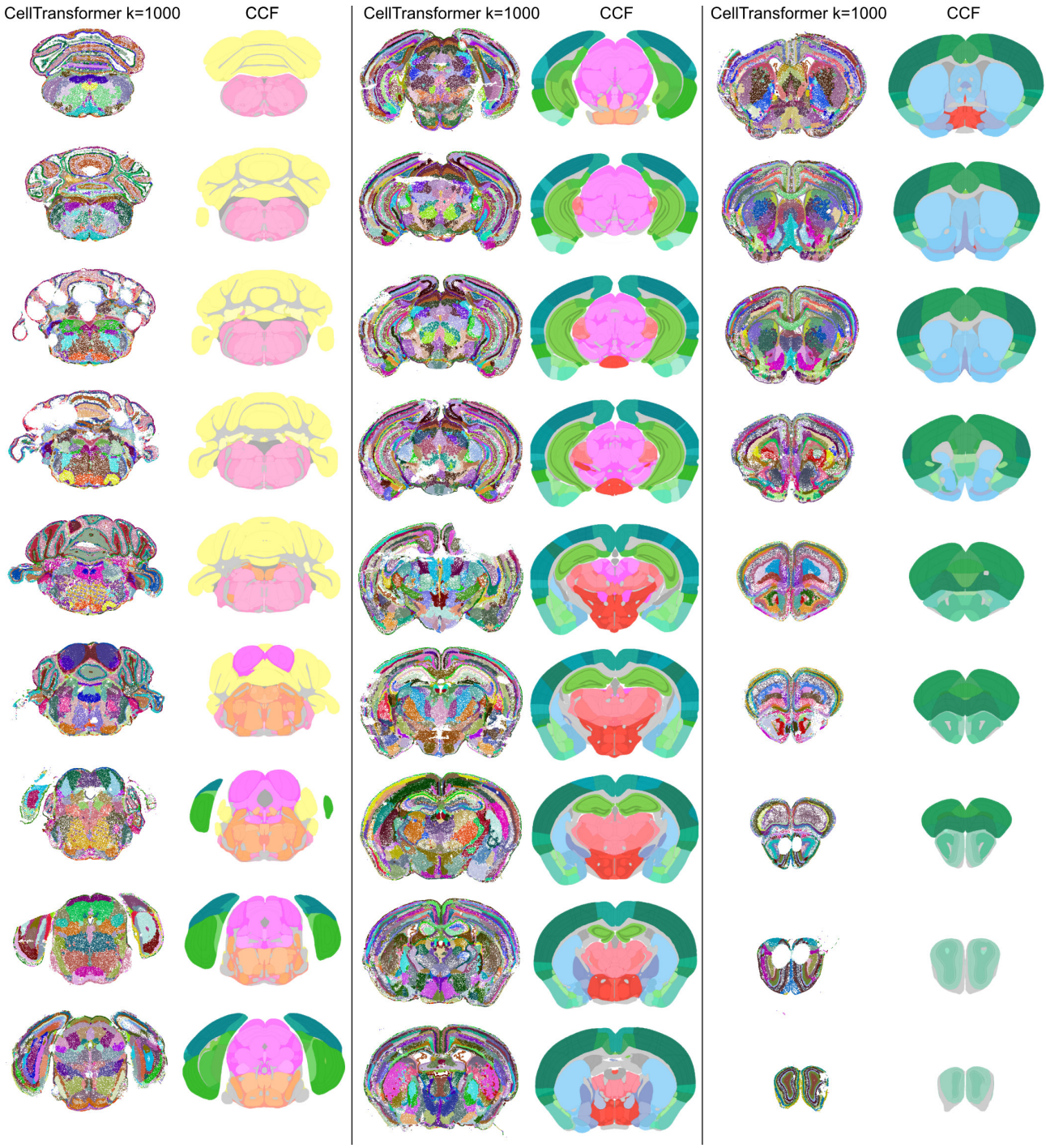

Figure S.3. CellTransformer spatial domains (left) and the corresponding CCF annotations (right) organized in 3 columns for roughly half of the sections in the Allen 1 dataset, approximately every other section. CellTransformer domains were calculated at k=1000 clusters.

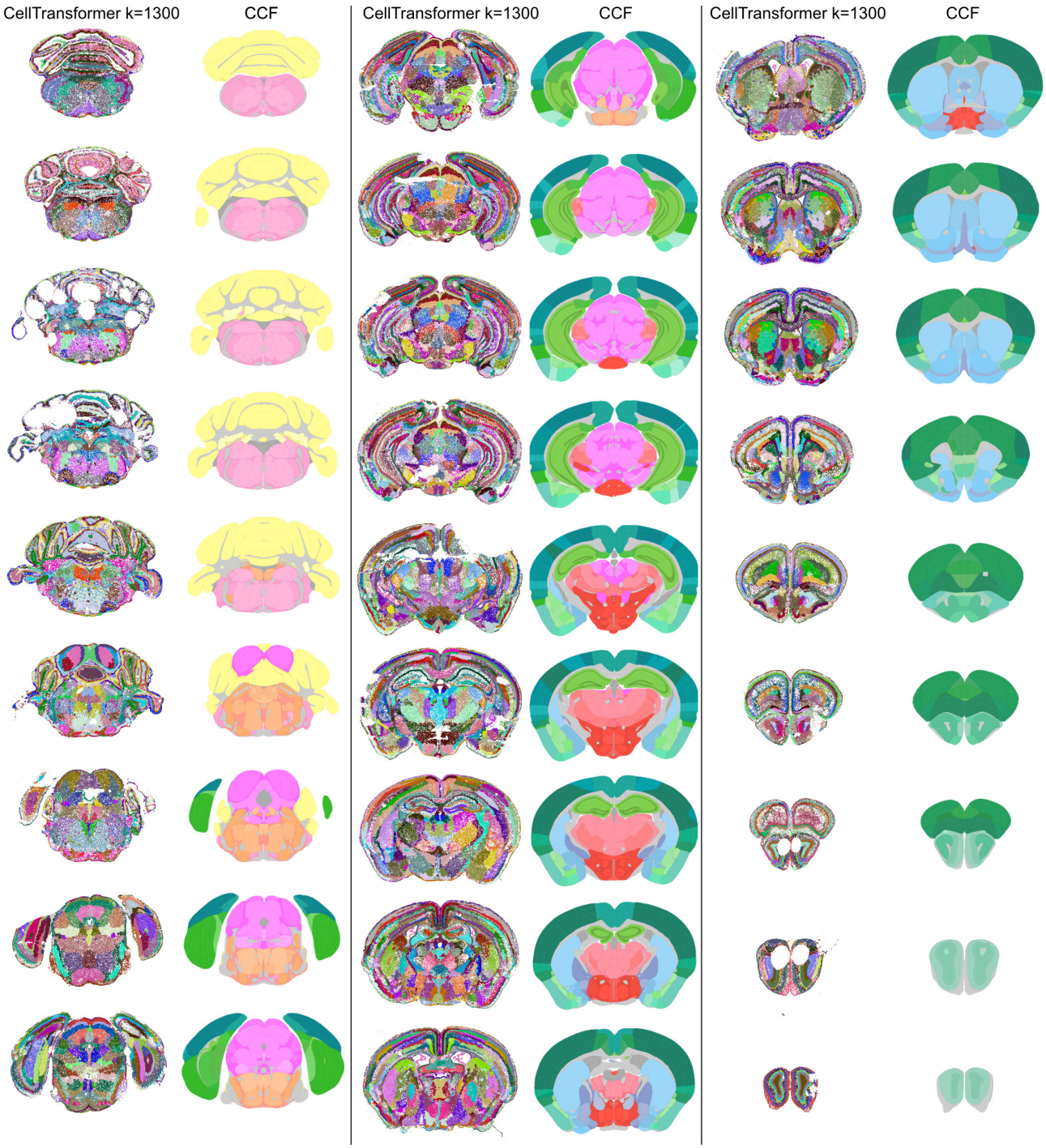

Figure S.4. CellTransformer spatial domains (left) and the corresponding CCF annotations (right) organized in 3 columns for roughly half of the sections in the Allen 1 dataset, approximately every other section. CellTransformer domains were calculated at k=1000 clusters.

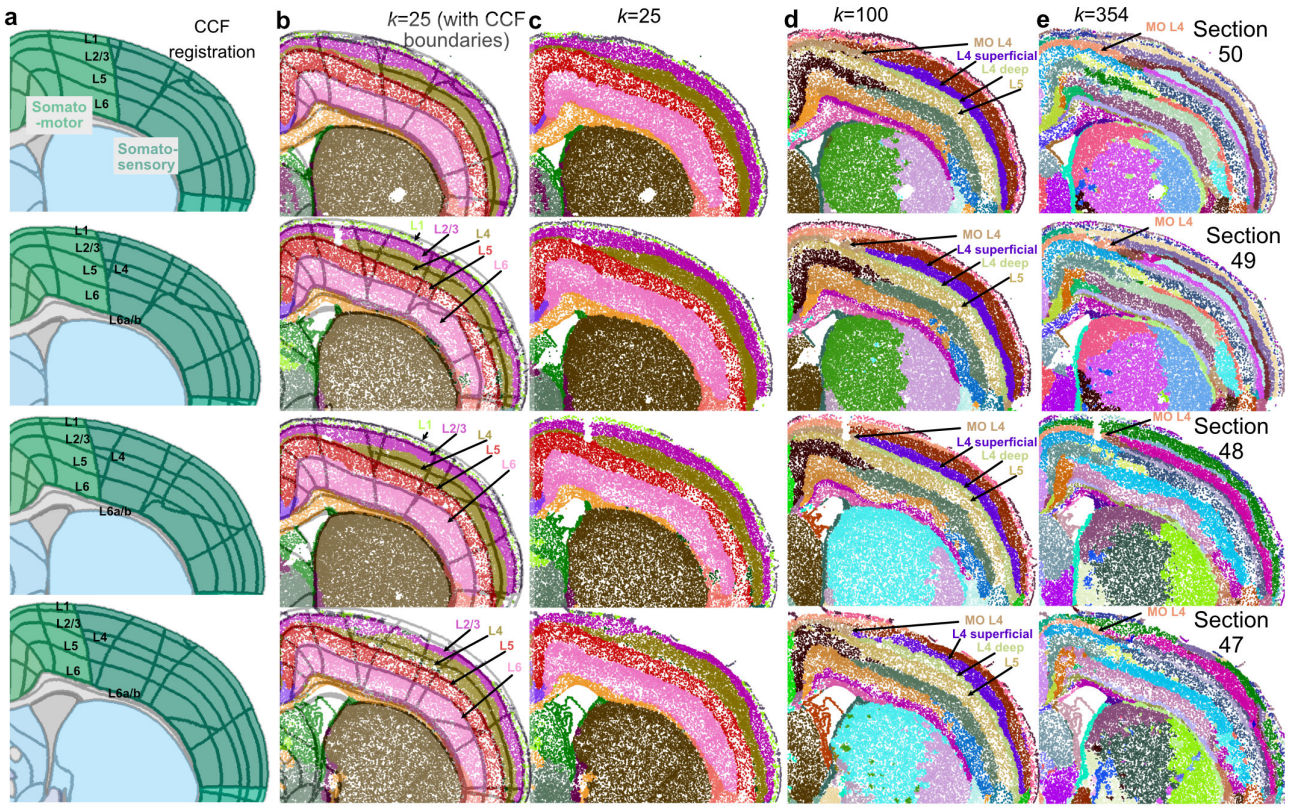

Figure S.5. Four sequential sections of the Allen 1 dataset (200 µm sampling interval between sections1) displayed with CellTransformer labels at varying resolution alongside CCF registration results. MO: motor cortex. (**a**.) CCF registration of four sequential sections shown in Fig. 1. cortical layers are marked based on CCF annotations. (**b**.) k=25 spatial domains with CellTransformer shown with regional boundaries from CCF in light gray. Putative cortical layers are annotated, showing CellTransformer replicates known cortical layers. (**c**.) 25 domains shown without CCF annotations to facilitate visualization. (**d**.) Same sections now shown with 100 domains to help show the transition from coarse (25 domains) to fine (100 domains). Sublayers of cortex are identified including layer 4 in motor cortex which transcriptomic studies have verified but has been difficult to identify using histological approaches. (**e**.) 354 domain zoom in on the same sections, showing consistency of layer 4 motor cortex detection as well as an anterior-posterior subdivision across motor and somatosensory cortical layers and clear distinction of cortical layers that lie within motor and somatosensory areas.

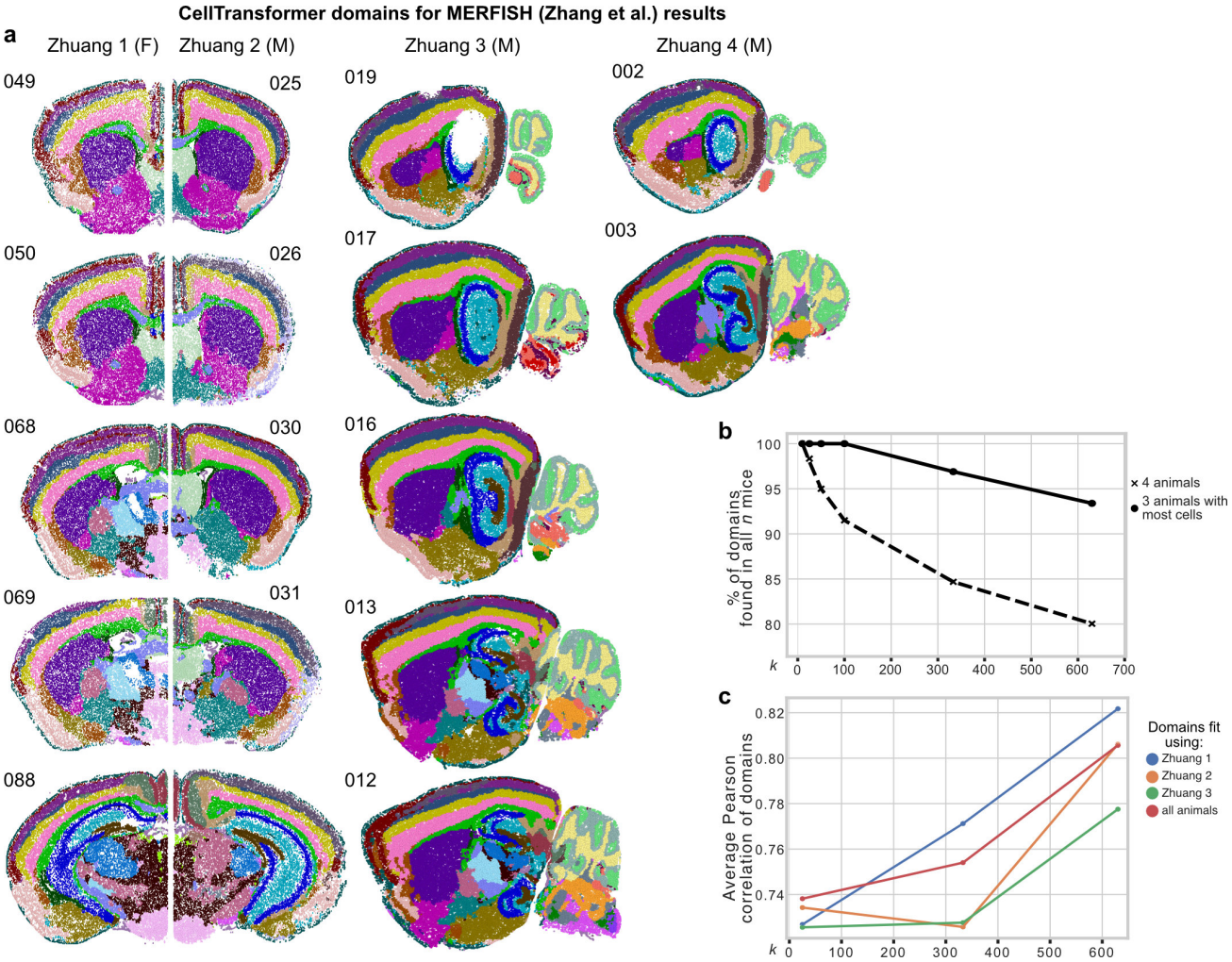

Figure S.6. Investigation into performance of CellTransformer on the Zhuang 1-4 datasets (239 sections, both coronal and sagittal, with a 1129 gene MERFISH panel (Zhang et al., 2023)). (a.) Representative images of all four mice arranged by column. The section number for each mouse is shown in the upper left of each image. Note that Zhuang 4 only had three sections. For each image, each dot is a cell neighborhood and colors come from a spatial clustering with k=50 (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 (Zhuang 1, 2, and 3). Note that because serial sections were collected at a higher frequency (100µm versus 200µm), different areas of the brain will have marginally higher coverage in one brain or another.

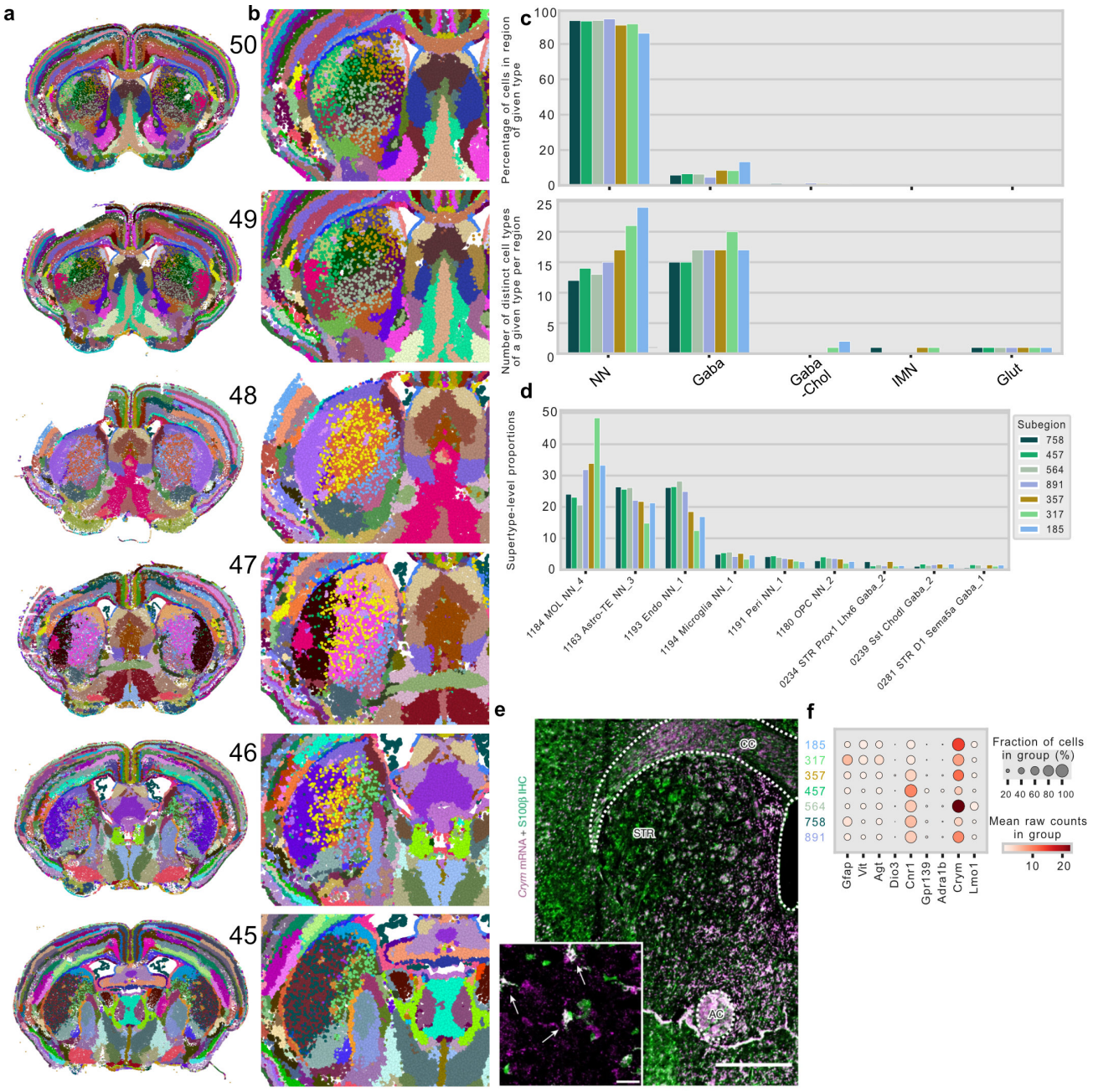

Figure S.7. Representative images of spatial clustering from CellTransformer models with k=1300 identified using the Allen 1 dataset. (a.) Sequential tissue sections (50 is most anterior) showing smoothness of spatial domains across and within tissue sections as well as consistent appearance of an irregular 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.) Supertype-level counts in putative subpopulations of caudoputamen. (e.) Reproduction with permission of results from (Ollivier et al., 2024). showing the distribution of *Crym* mRNA and its protein product (S100B), clearly identifying a medial population of *Crym*+ neurons which resembles the spatial pattern observed in clusters 758 and 457 (dorsoventral and *Crym*-). (f.) Dotplot of cell type expression proportions and mean counts per group (raw counts) in identified irregular caudoputamen areas.