SpaGFT is a graph Fourier transform for tissue module identification from spatially resolved transcriptomics

Yuzhou Chang[1, 2,†], Jixin Liu[3,†], Anjun Ma[1,2], Zihai Li[2], Bingqiang Liu[3,4*], and Qin Ma[1,2,*]

[1] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH 43210, USA.
[2] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.
[3] School of Mathematics, Shandong University, Jinan 250100, China.
[4] Shandong National Center for Applied Mathematics, Jinan, Shandong, Jinan 250100, China.
[†] These authors contributed equally
[*] Corresponding Authors

The tissue module (TM) was defined as an architectural area containing recurrent cellular communities executing specific biological functions at different tissue sites. However, the computational identification of TMs poses challenges owing to their various length scales, convoluted biological processes, not well-defined molecular features, and irregular spatial patterns. Here, we present a hypothesis-free graph Fourier transform model, SpaGFT, to characterize TMs. For the first time, SpaGFT transforms complex gene expression patterns into simple, but informative signals, leading to the accurate identification of spatially variable genes (SVGs) at a fast computational speed. Based on clustering the transformed signals of the SVGs, SpaGFT provides a novel computational framework for TM characterization. Three case studies were used to illustrate TM identities, the biological processes of convoluted TMs in the lymph node, and conserved TMs across multiple samples constituting the complex organ. The superior accuracy, scalability, and interpretability of SpaGFT indicate that it is a novel and powerful tool for the investigation of TMs to gain new insights into a variety of biological questions.

A tissue module (TM) is a critical concept for investigating molecular tissue biology based on molecule compositions and functions in either homogenous or heterogeneous tissues. However, there is no rigorous computational formulation for TM identification because of the following: (*i*) TMs exhibit a wide range of length scales, and the repository of TM spatial patterns is unknown; and (*ii*) the molecular features of a TM and the relevant feature crosstalk of convoluted TMs are not well-defined[1]. Among the to-be-discovered molecular features, a group of spatially variable genes (SVGs) can be used to represent and define TMs if they share recurrent and similar spatial expression patterns within one or across multiple datasets. Particularly, the prediction of SVGs can be fully enabled using spatially-resolved transcriptomics (e.g., 10X Genomics Visium and Slide-seqV2[2]), which simultaneously measures gene expression and spatial locations of spots within healthy or pathogenic tissues[3]. Existing SVG prediction methods are mainly hypothesis-driven and developed based on statistical frameworks (e.g., SpatialDE) or graph neural networks (e.g., SpaGCN)[1,4]. Although these methods exhibit good SVG detection performance, are equipped with rigorous statistical evaluation, and provide valuable biological insights, they exhibit two main limitations: (*i*) these methods can effectively identify certain well-defined patterns (e.g.,

radial hotspot, curve belt, or gradient streak), but they exhibit a lesser detection performance for irregular patterns, such as the T cell zone, B cell zone, or germinal center (GC) in the lymph node[1]; and (*ii*) although existing tools exhibit a competitive SVG identification accuracy with sacrificing scalability (e.g., SpatialDE and SPARK work well for Visium data), the accuracy decreases if a tool significantly improves the efficiency (e.g., SPARK-X)[5] of datasets with a large number of spots/cells[6,7].

To solve these challenges, we proposed a hypothesis-free graph Fourier transform framework (GFT), named SpaGFT, for SVG and TM identification from spatial transcriptomics data. Our framework transforms obscure spatial gene expression patterns from the spatial domain to simple, informative, and quantifiable frequency signals in the Fourier domain. First, by taking advantage of Fourier domain signals, SVGs can be identified quickly and accurately without relying on the spatial pattern hypothesis. To demonstrate the superior performance and efficiency of the SpaGFT, 31 public datasets were used to compare the performance of SpaGFT to those of other state-of-the-art tools. Furthermore, SVGs with similar Fourier domain signal patterns can also be grouped into clusters, which are defined as TMs in our framework. We used three cases to explain the major applications and biological insights of the identified TMs from the gene-centric perspective. In the first case, we proposed a TM ID card to define TMs by showing the following: (*i*) the signature Fourier domain signal patterns; (*ii*) the corresponding SVGs with similar spatial patterns; (*iii*) the enriched biological functions of these TM-associated SVGs; and (*iv*) the relevant cell type annotations. In the second case, SpaGFT showed its capability of identifying short-length scale TMs and revealing convoluted biological processes among distinct TMs in human lymph nodes. Lastly, we used seven mouse samples from two anatomical views to (*i*) demonstrate the contribution of tissue motifs to understanding the 3D structures of the mouse brain from a 2D perspective; and (*ii*) conclude that conserved and connected TMs shared by multiple samples compose the basic functional units of complex organs. Overall, the results revealed that SpaGFT can accurately identify SVGs at a fast computational speed, and for the first time, provide a computational formulation and strong biological interpretation for TM identification from a gene-centric perspective.

Results

SpaGFT is a graph Fourier transform framework for SVG identification and TM characterization. SpaGFT generates a novel representation of gene expression and the corresponding spot graph topology in a Fourier space (**Fig. 1a**), which enables TM identification and enhances SVG prediction and interpretation. This transform does not rely on predefined spatial pattern[8] assumptions, which ensures its generalizability in identifying both well-defined and irregular SVG patterns across various datasets (**Fig. 1b**). Particularly, the core algorithm of SpaGFT projects spatial transcriptomics data on an orthogonal basis, known as Fourier modes (FM), which is represented in the increasing order of its frequency, with FM_1 having the lowest frequency (**Supplementary Fig. 1a**). A low-frequency FM contributes to a slow signal variation, which results in a more recognizable spatial pattern (**Supplementary Fig. 1b**). To project a specific gene, each FM exhibits a signal intensity associated with the spot graph topology and retains the diverse orthogonal basis of the oscillation patterns. The signal intensity can be used to identify SVGs effectively and efficiently in SpaGFT using the rule: a gene with a high intensity

of low-frequency FM signals compared to high-frequency FM signals is typically an SVG, whereas a gene with a low intensity of low-frequency FM signals indicates random expression patterns (**Supplementary Fig. 1c**). To determine TMs using SpaGFT, the low-frequency SVG FM signals are selected as features to identify SVG clusters using Louvain clustering (**Fig. 1c**). Spatial regions (a group of spots) with high SVG expression patterns are considered as one TM. Multiple downstream analyses and interpretations can be given to elucidate a TM, including Uniform Manifold Approximation and Projection (UMAP) visualizations, TM-specific SVG functional enrichment, low SVG expression enhancement, sub-TM identification, short length-scaled TM identification (e.g., the GC of lymph nodes at ~55 μm diameter spots), and tissue motif[9] (basic compartment of a specific tissue) identification across multiple samples. Notably, due to SpaGFT transforming SVG spatial expressions into FM signals, signal processing approaches can be used for complementary applications. For example, due to the dropout issue[10], some SVGs may have retained a low expression in a given TM, and SpaGFT offers an additional function the SVG enhancement (**Fig. 1d and Methods**): a low-pass filter enhances low-frequency FM signals and denoises high-frequency FM signals to form enhanced FM signals. The new signal will then recover the SVG spatial expression with an enhanced magnitude via inverse graph Fourier transform (iGFT).

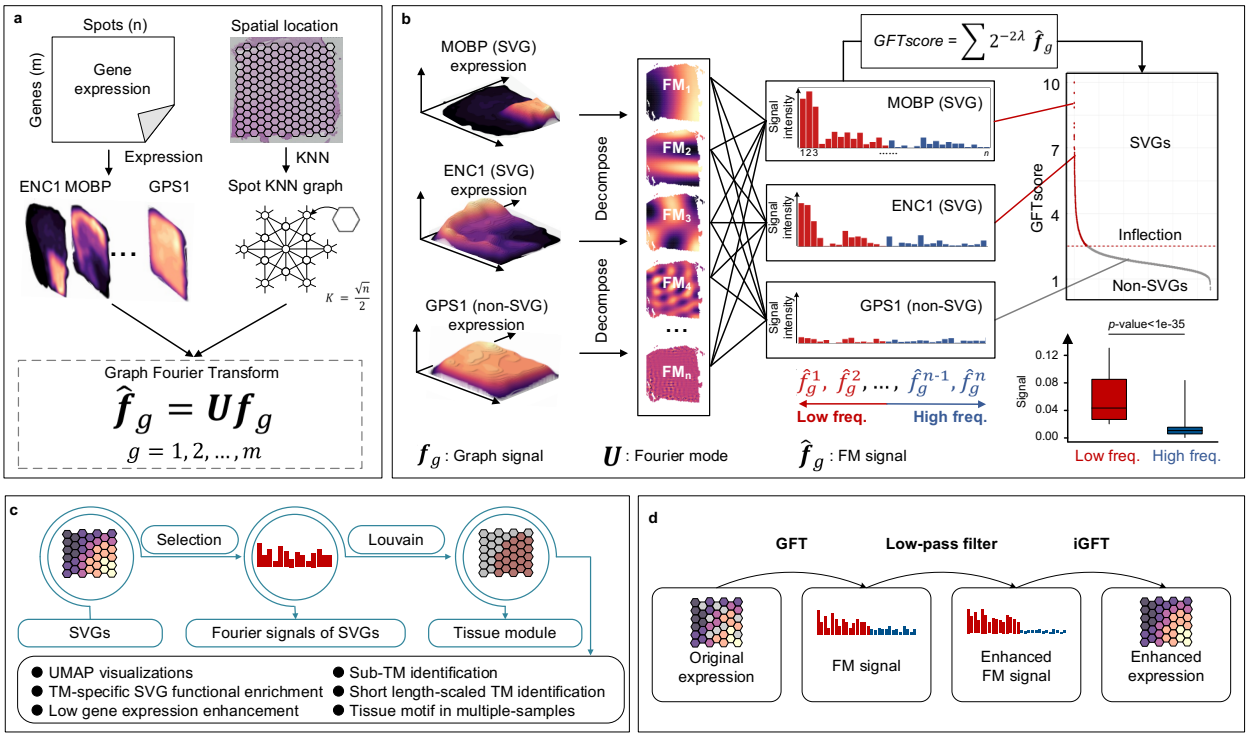

Fig.1 | Overview and validation of SpaGFT. a. SpaGFT considers a gene-spot expression count matrix ($m \times n$) and spatial locations as inputs. The spatial expression of three genes, *ENC1*, *MOBP*, and *GPS1*, are shown as examples. A KNN graph is generated by calculating the Euclidean distance among spots based on spatial locations between any two spots. Spot location is used to construct the spot graph using the KNN method, where *K* is half of the square root of n. By combining gene expressions and spot KNN graph, the graph signal $\boldsymbol{f}_g$ of gene g can be

projected to a series of FM U and transformed into a frequency signal $\hat{f}_g$ using a graph Fourier transform. **b.** The spatial expression of three genes, including two known SVGs (*MOBP* and *ENC1*) and one non-SVG (*GPS1*), are shown as examples. All genes can be decomposed into multiple FMs (a series of periodic signals with gradually faded patterns) and corresponding frequency signals. The FMs in the Fourier space can be separated into low-frequency (red) and high-frequency (blue) domains. For each gene, a *GFTscore* was designed to quantitatively measure the frequency signal intensity in the low-frequency domain. The threshold (inflection point) of the *GFTscore* was determined using the Kneedle algorithm, and the significance of a *GFTscore* (*p*-value) was determined using a non-parametric test. A gene is defined as an SVG (red dots) if its *GFTscore* is greater than the inflection point and its false discovery rate (FDR) adjusted *p*-value is less than 0.05. Additionally, for sample 151673, we observed an SVG with a significantly higher intensity of low-frequency FM signals than high-frequency FM signals (box plot in the right-bottom corner, with a *p*-value<$1e^{-35}$ by Wilcoxon rank-sum test). **c.** Workflow of TM identification in SpaGFT. **d.** The low SVG expression signal can be enhanced by a low-pass filter and iGFT using low-frequency FM signals.

SpaGFT accurately identifies SVGs in human and mouse brains. In this study, we collected 31 spatial transcriptome datasets from human and mouse brains from public domains, and the samples were sequenced using scales from two different spatial technologies (i.e., Visium measures ~55 μm diameter per spot and Slide-seqV2 measures ~10 μm diameter per spot [**Supplementary Table 1**]). Grid-search tests were conducted under a wide range of parameter combinations in all the benchmarking tools using three high-quality brain datasets. As no golden-standard SVG database is available, we collected 849 SVG candidates from the brain regions of mice and humans from five studies[11-15] and selected 458 genes as curated benchmarking SVGs based on cross-validation with the In Situ Hybridization (ISH) database of Allen Brain Atlas (**Supplementary Tables 2 and 3**, **Methods**). The SVG prediction performance was evaluated using six reference-based metrics, and the results revealed that SpaGFT outperformed the other five tools on the three datasets in terms of the Jaccard score and the other five metrics (**Fig. 2a**, **Supplementary Fig. 2a**, and **Supplementary Table 4**). It is essential to note that the computational speed of SpaGFT was two-fold faster than that of SPARK-X and hundreds-fold faster than those of the other four tools on the two Visium datasets. Although SpaGFT exhibited a slower performance than SPARK-X on the Slideseq-2 dataset, it exhibited a remarkably enhanced SVG prediction performance compared to SPARK-X (**Supplementary Table 5**). Based on the above grid-search result, we considered the parameter combination with the highest median Jaccard scores across the three datasets as the default parameter in SpaGFT. Subsequently, the performance of SpaGFT on an additional 28 independent datasets was compared to those of the other five tools (all using their default parameters) to test its applicability and robustness. The results revealed that it achieved the best performance among the investigated tools on all six metrics (**Fig. 2b, Supplementary Table 6**). In addition, the SVG prediction performance without the above curated benchmarking SVGs was evaluated using Moran's I and Geary's C statistics (two reference-free evaluation metrics, **Methods**), and the results revealed that the overall performance of SpaGFT was lower than that of MERINGUE (second best) because Moran's I was implemented in MERINGUE's model for SVG prediction (**Supplementary Fig. 2b**).

Two classical markers in the hippocampus and cortical region in **Fig. 2c** show SpaGFT's ability to identify SVGs that are detectable by other tools[16,17], and **Fig.2d** shows its ability to identify unique SVGs (**Supplementary Table 7**), which were validated using the ISH database (**Supplementary Fig. 3**) and reported by previous studies[18-21]. For example, the *Calb2* gene encoding calretinin was employed as one of the traditionally used markers to categorize interneurons[18]. *Hcrt* was associated with controlling sleepiness[19]. *Gal* has been implicated in many behavioral processes, including anxiety, and thus represents a potential target for novel strategies aimed the pharmacological treatment of depression and anxiety disorders[20]. In addition, *Asb4* was associated with obesity[21]. Furthermore, to demonstrate and visualize the strength of the FM signals for distinguishing SVGs and non-SVG patterns, we projected the top 50 low-frequency FM signals on a two-dimensional UMAP space and compared them to those of a UMAP that utilizes the top 50 principal components (**Methods**). The results revealed that SVGs identified by SpaGFT were distinguishably separated from non-SVGs on the FM-based UMAP with a linear boundary, whereas SVGs were irregularly distributed on the principal components analysis (PCA)-based gene UMAP (**Fig. 2e**). This indicates that the FM signal is a better low-dimensional representation for characterizing SVG patterns, which lays a solid foundation for TM identification and interpretation.

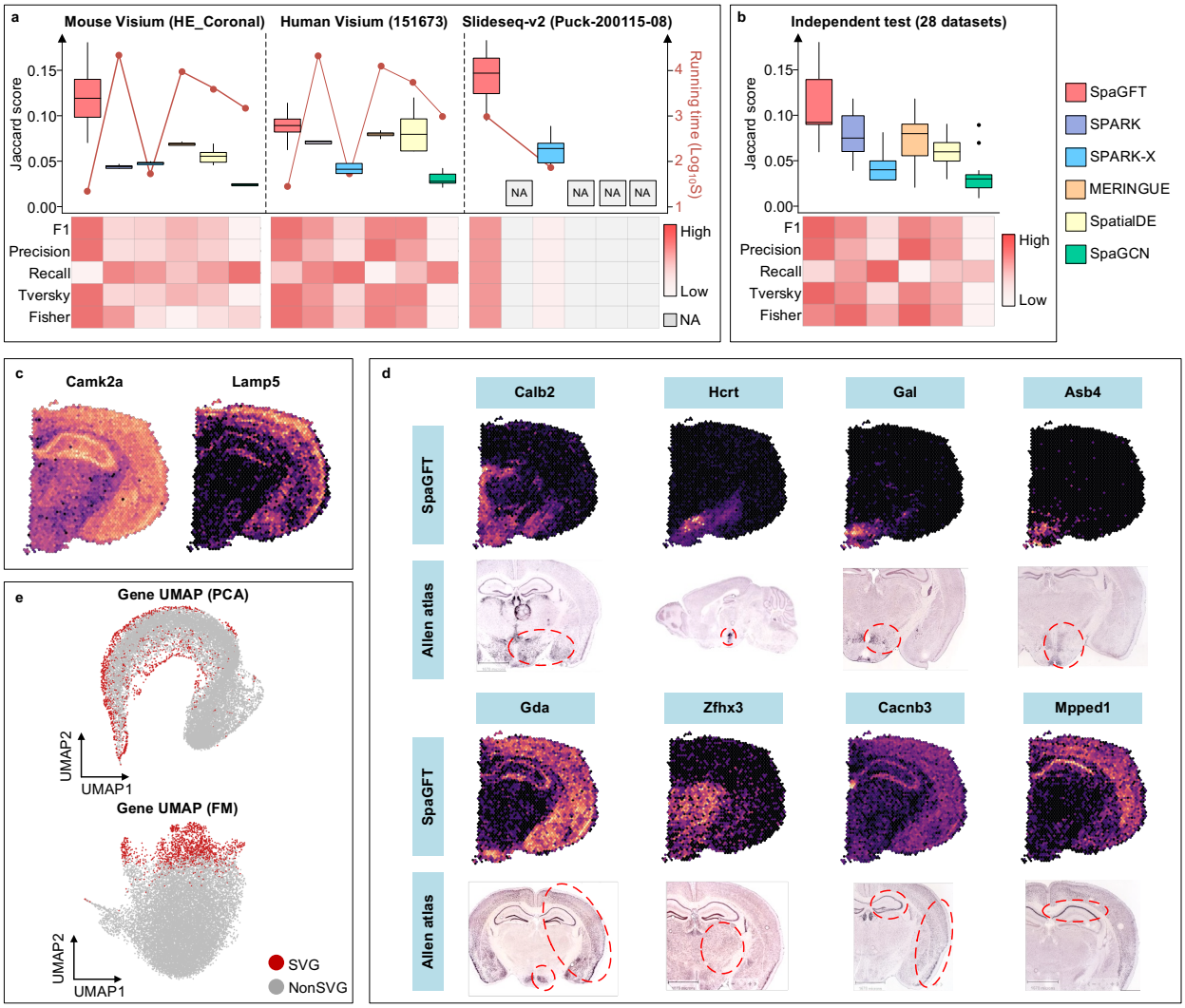

Fig.2 | SVG identification performance comparison. a. The SVG prediction evaluation of SpaGFT was compared to those of the five benchmarking tools in terms of the Jaccard similarity score. To evaluate the robustness of the tool, a grid-search method was used on all tools and three datasets (HE-coronal, 151673, and Puck-200115-08) under different parameter combinations. The running time (seconds with log-transformation) of each tool is represented as the line graph. In addition, the median of five additional matrices (i.e., F1 score, precision, recall, Tversky index, and odds ratio of Fisher's exact test) on all parameter combinations for each tool is also shown as heatmaps. **b.** We selected the parameter combination showing the highest median Jaccard scores among all three benchmark datasets as the default parameter in SpaGFT. Using such parameter selection, the SVG prediction performance of SpaGFT on additional 28 independent datasets was compared to those of the five benchmark tools using their own default parameters. The black line in each box indicates the median Jaccard score of all the 28 datasets. **c.** Two SVGs in the mouse brain coronal plane (proved by Allen Brain Atlas) that can be simultaneously identified within the top 100 SVGs by SpaGFT, SpaGCN, SPARK-X, MERINGUE, and SpatialDE. The spatial map results from SpaGFT are shown with a brighter color, which represents higher expressions. **d.** Spatial map visualization and Allen Brain Atlas ISH data of

eight SVGs that are uniquely identified by SpaGFT, showing distinct patterns in the spatial domain. Red circles in the ISH data indicate the expression region of the mouse brain. **e.** Comparison of the UMAPs obtained using the top 50 principal components (PCs) (left) and the top 50 FMs (right) of the Mouse Visium data (HE-coronal, 2702 spots). The principal component analysis (PCA) dimensions were generated directly from the gene-spot expression matrix using PCA analysis in Scanpy. Red dots indicate the 1,456 SVGs identified by SpaGFT using default settings, whereas the grey color suggests non-SVGs.

SpaGFT characterizes TM in the mouse brain based on the gene-centric perspective. We believe a rigorous computational formulation for TM identification is non-travail, and a clear TM definition should be multi-angled and involve multi-omics. From a spatial transcriptomics perspective, we characterize a specific TM using an ID card, including a spatial expression map, TM digital map, transformed FM signals, associated SVG list, and underlying biological pathways. We applied SpaGFT to identify TMs by determining SVG clusters, which share similar signal patterns in the FM space. Taking the HE-coronal data as an example, seven SVG clusters were identified from a total of 1,456 SVGs, which corresponded to seven TMs (TM 1–7) (**Fig. 3a**). Particularly, we revealed the ID card for each identified TM (**Fig. 3b** and **Supplementary Figs. 4-9**). The top four SVGs (e.g., *Ctxn1*, *Ngef*, *Hpcal4*, and *Tspan7*) were selected to support the spatial expression pattern of TM 1. The enrichment of ontologies, pathways, and transcriptional factors was also performed for TM-associated SVGs to elucidate the underlying biological process of the TM[16,22,23]. Using this identified TM pattern, SpaGFT enhanced the expression signal of the SVGs using a signal processing method followed by a low-pass filter and iGFT (**Fig. 1d**). For example, *JunB* is a validated SVG in the cornu ammonis field 1 (CA1) region (**Supplementary Fig. 10a**) with a regulatory role on memory[24]. However, the rank of the *JunB* GFT score was 214th among 275 SVGs in TM 1, indicating that the strength of the SVG signal in the spatial domain was lower than that of the other SVGs (**Supplementary Table 8**). Moreover, *JunB* could be enhanced by SpaGFT to obtain a more distinguishable pattern, and the granularity of the enhanced gene expression signals also increased (figure looks sharper) when the number of selected low-frequency FMs was used (**Fig. 3c**). Next, we observed that TMs remain sub-patterns. For example, TM 3 could be further clustered into four sub-TM patterns (**Fig. 3d**), which corresponded to preferred regions in the hippocampus, hypothalamus, cortical subplate, and thalamus. Similar sub-structures were also identified in the other six TMs and were associated with corresponding brain structures (**Supplementary Figs. 10** and **11**).

We further investigated the cell composition of the TMs and corresponding sub-TMs. Based on the deconvolution result of 59 mouse brain cell types from the cell2location framework (**Supplementary Table 9**)[25], we observed that TM 1, TM 2, TM 5, and their sub-TMs were composed of similar cell compositions (shown in the red rectangular box), including major neuronal cells, and supported the anatomical structures of the mouse cerebral cortex. This cellular composition could also be validated by TM-associated SVGs. For example, *Gad1* (TM 1-associated SVG), *Vip* (TM 2-associated SVG), and *Snap25* (TM 5-associated SVG) are classic markers of GABAergic neurons[14]. In addition, *C1ql3* (TM 1-associated SVGs), *Slc17a7* (TM 2-associated SVGs), and *Arf5* (TM 5-associated SVGs)[12,14] are known markers of glutamatergic neurons[26]. Similarly, TM 3, TM 4, and the corresponding sub-TMs were enriched with major

inhibitory neurons (**Fig. 3e**). *Calb2* (TM 3-associated SVG) and *Pvalb* (TM 4-associated SVG) are documented markers of inhibitory neurons[14]. TM 6 and TM 7 mainly contained cell types from the white matter and thalamus region (**Fig. 3f and Supplementary Fig. 12**). For example, *Mog* (TM 6-associated SVG) and *Tcf7l2* (TM 7-associated SVG) are the markers of oligodendrocyte and thalamocortical neurons [14,27]. Interestingly, although the sub-TM 4 of TM 6 was classified as a TM 6-alike category, the tissue module pseudo-expression and cell-type distribution supported that it belonged to the caudoputamen region (CP), which was a distinct region from the other sub-TMs, including sub-TM 1, sub-TM 2, and sub-TM 3 (**Fig. 3g**). For example, *Meis2*-positive inhibitory neurons enriched in the CP were also enriched in this region[25], and the CP regional marker *Adora2a* was also the TM-associated SVGs (sub-TM 4 of TM 6, i.e., TM 6_4)[28]. By investigating the TM-associated SVGs, their enriched biological functions, and cell type compositions, a TM can be defined and characterized from both the genetic perspective (i.e., how to define SVGs and functions in a specific region) and cellular perspective (i.e., what is the cell type composition in a specific region). In conclusion, these results demonstrated the ability of SpaGFT to identify, characterize, and interpret TMs based on SVGs, and it is complementary to the current state-of-the-art deconvolution tool for further TM interpretation[10].

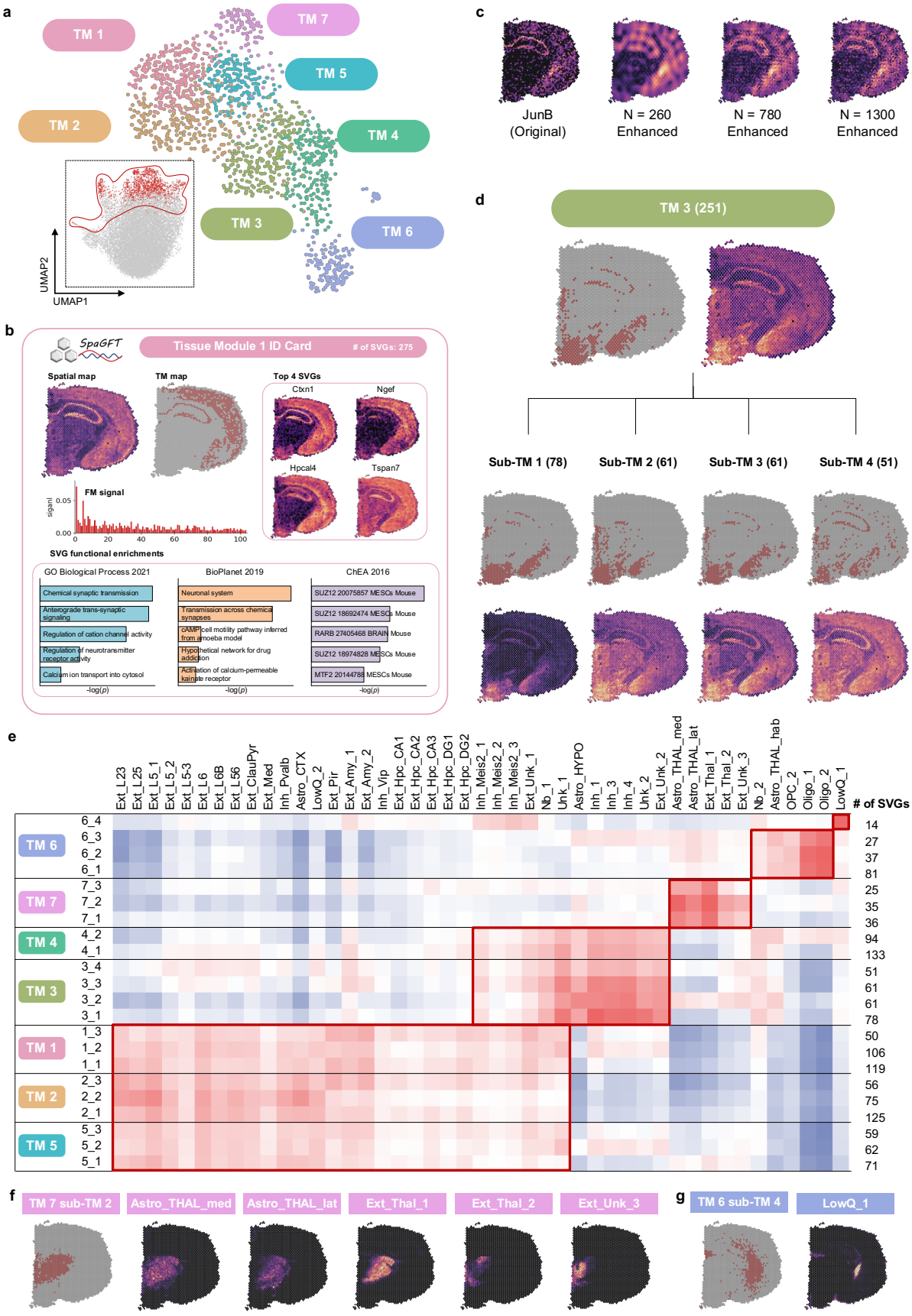

Fig.3 | TM identification in mouse brain. a. All 1,456 SVGs identified in the Mouse Visium data (HE-coronal) were grouped into seven clusters, which represent seven TMs (TM 1–7), and it is shown in a UMAP space, which is shown in Fig.1g with all genes located in the left-bottom corner (red: SVGs, grey: non-SVGs). **b**. An ID card is created to display fundamental information of each TM. Here, we use TM 1 as an example. The spatial map shows the pseudo-expression of TM1 with 275 SVGs, where brighter color indicates higher pseudo-expressions. The TM map is a binarized pseudo-expression map with an expression cutoff of 85 percentile. The low-frequency FM signals of TM 1 is displayed below. The spatial maps of the top four SVGs ranked by their *GFTscore* from high to low are shown on the right. Functional enrichment tests of the 275 SVGs are performed on three databases (i.e., GO Biological Process 2021, BioPlanet 2019, and ChEA 2016) via Enrichr R package to provide insights on the functional and regulation information enriched in TM 1. **c.** SpaGFT can enhance the low SVG expression signal of JunB (an SVG in TM1) using an inverse graph Fourier transform (iGFT) for low-frequency FM signals. The spatial maps of JunB using original expression and enhanced expression are shown. **d.** Four sub-TMs were identified in TM 3 by re-clustering SVGs in TM 3. Each sub-TM possesses a group of unique SVGs, which exhibits varying spatial expression patterns. The number in the parenthesis indicates gene numbers in each sub-TM. **e**. The heatmap visualizes the TM-cell type matrix, where a row represents a sub-TM and a column represents a specific cell type. An element in this matrix represents the Pearson correlation coefficient between the proportion of a cell type and the pseudo-expression of a sub-TM across all the spots. A red color block in the heatmap indicates a high association between the corresponding cell type and sub-TM. **f** and **g**. The figures showcase cell type composition and distribution of TM 7 sub-TM 2 and TM 6 sub-TM 4, respectively.

SpaGFT identifies short-length scale TMs and the crosstalk among convoluted TMs in a human lymph node sample. Lymph node belongs to the secondary lymphoid organ, containing T cell zones, B follicles, and a GC (a short length-scale TM under a ~55 μm resolution)[25]. We applied SpaGFT to lymph node Visium data[29] to investigate the organization of the three functional regions and their convoluted crosstalk at the molecular level. SpaGFT identified 1,490 SVGs, leading to nine TMs (**Fig. 4a** and **Supplementary Table 10**). The cell proportion of 34 cell types in this Visium data were predicted using cell2location (**Supplementary Table 11**). Each TM (and sub-TMs) was correlated with a set of cell types based on cell proportions (**Fig. 4b**). Particularly, TM 6 was highly correlated with six GC signature cell types defined by cell2location, including T follicular helper cells (T_CD4_TfH_GC), follicular dendritic cells (FDC), pre-GC B cells (B_GC_prePB), cycling B cells (B_cycling), dark zone B cells (B_GC-DZ), and light zone B cells (B_GC-LZ). TM 7 was highly correlated with eight T cell related cell types, and TM 8 was highly correlated with five B cell related cell types. Altogether, 143 SVGs were associated with TM 6, including *PCNA*, *CDK1*, and *CDC20,* which were marker genes of cell proliferation in the GC enriched in the cell cycle pathway[30,31] (**Fig. 4c**). In addition, TM 7 exhibited a higher proportion of the seven T cell types and 132 SVGs, including several T cell zone markers relevant to T cell survival, such as *CD3E, IL7R, CCR7, and CCL19*[32]. The pathway analysis result showed that the 132 SVGs were enriched in the T cell activation and differentiation pathway (**Supplementary Fig. 13**). Additionally, a B cell-enriched niche was identified in TM 8, where several B cell markers (e.g., *CD19, CD79B, and CR2*) and relevant pathways (e.g., antigen processing and presentation)

were identified (**Fig. 4c**)[33]. Therefore, we defined TMs 6, 7, and 8 as GC, T cell zone, and B follicle, respectively. We visualized the three TM locations and found that they were spatially close to each other, indicating potential convoluted functions among these three TM regions (**Fig. 4d**).

To further reveal the crosstalk among these three regions, we projected spots (assigned to all three regions) to the Barycentric coordinate (the equilateral triangle in **Fig. 4e and Supplementart Table 12**) to display the distribution differences of cell type components and the abundance of spots in the three TMs and adjacent regions between the GC and T cell zone or B follicles. There were 174 spots assigned to the interactive region between the GC and B follicle, and the region was indicated by the local spatial map (**Fig. 4f**). Furthermore, the B follicle-associated SVG, *CXCL13,* supports one of the major lymph node functions for B cell maturation and antibody production [34,35]. The 66 spots aligned from the interactive region between the GC and T cell zone also showed another convoluted collaboration (**Fig. 4g**). This could be supported by the T cell zone-associated SVG, *ICOS*, which was one of the signature follicular helper T cell genes for GC formation and high-affinity antibody development[36-38]. Overall, we reasoned that SpaGFT could be used to identify short length-scale TMs and interpret the crosstalk among convoluted TMs to support complex biological processes[39,40]. Future studies will address if SpaGFT can be used to discern functionally specific TMs associated with effective immune responses (such as in the case of vaccination) and/or pathology (such as in the case of cancer metastasis to the lymph node)[41] .

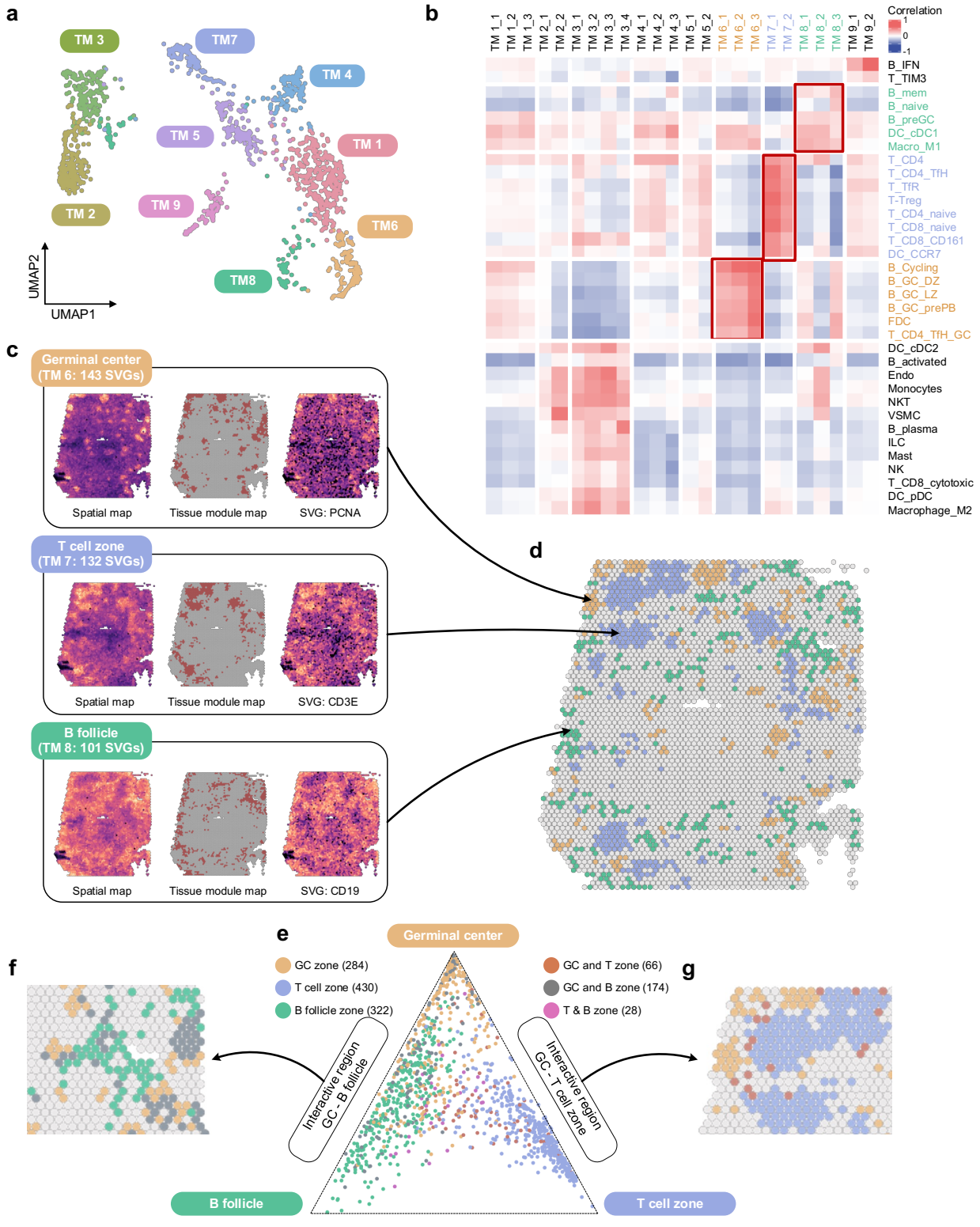

Fig.4 | Case study of a human lymph node to demonstrate short-length scale TM identification. a. UMAP visualization of nine SVG clusters in the Human lymph nodes data identified by SpaGFT using default parameters. **b.** The heatmap visualizes the transposed TM-

cell type matrix defined at **Fig. 3e**. According to the transposed TM-cell type matrix, TM 6, TM 7, and TM 8 correspond to the GC, T cell zone, and B follicle, respectively. **c.** The figure demonstrates the pseudo-expression and binary TM of GC, T cell zone, B follicle, and corresponding signature genes. **d**. By combining all three TMs on the graph, different colors correspond to spots in each TM, and the spots overlapped in the three TMs, and non-assigned spots were colored grey. **e**. The barycentric coordinate plot shows cell-type components and the abundance of spots in interactive and functional regions. If the spot is closer to the vertical of the equilateral triangle, the cell type composition of the spot tends to be signature cell types of the functional region (**Methods**). The spots were colored by functional region and interactive region categories. **f**. The spot distribution of spots from a local spatial map of the interactive region between GC and B follicle. **g**. The spot distribution of spots from a local spatial map of the interactive region between GC and T cell zone.

SpaGFT reveals the 3D structures of the cerebrum, hypothalamus, and white matter in terms of tissue motifs in the mouse brain. Tissue motif is a newly computational concept for investigating the tissue organization and collaboration of fundamental structures[9]. In our study, we extended the definition of a tissue motif to include a conserved tissue structure across multiple samples of a complex organ (e.g., mouse brain), and hypothesized that the conserved TMs, defined as a group of TMs representing the fundamental structure of the same organ, should have similar SVG components regardless of sampling strategies or sources. Seven mouse brain samples were collected from the Visium website and one independent study[29,42], including two anatomical planes (i.e., sagittal and coronal planes). The four samples in the sagittal plane were obtained from frozen fresh samples. Regarding the three samples from the coronal plane, HE-coronal and GSM5519054 were frozen fresh samples from different sources and sampling locations, and IF-FFPE was preserved in formalin and paraffin (**Fig. 5a** and **Supplementary Table 1**). Using the default parameter settings, SpaGFT identified 67 TMs among the seven samples (**Supplementary Table 13**). The 67 TMs were grouped into 14 clusters using the Louvain algorithm based on their associated SVGs (**Methods**). If two TMs contain similar components of TM-associated SVGs, they are typically grouped into the same cluster and represent the same fundamental structure, even if they were from different anatomical views.

As a result, we focused on the three colored TM clusters (TM clusters 1, 2, and 3 in **Fig. 5b**), each of which contained conserved TMs from at least six samples (**Supplementary Table 14**). First, we investigated the cell components of each TM using cell2location, and the results showed that TM clusters 1, 2, and 3 were highly correlated with excitatory neurons, inhibitory neurons, and non-neuronal cells, respectively (**Fig. 5c**). For example, TM cluster 3 was enriched with oligodendrocytes and was also in agreement with the white matter anatomical structure from the Allen Brain Atlas[16,26] (**Fig. 5d**). TM clusters 1 and 2 were highly correlated with the partial cerebrum and hypothalamus regions, respectively (**Supplementary Fig. 14**). Particularly, regarding TM cluster 3, *Mbp* and *Mobp* (white matter signature genes) were simultaneously captured by all seven conversed TMs[15], while the two genes were not conserved TM-associated SVGs of the cerebrum and the hypothalamus regions. We concluded that conserved TMs forming one TM cluster typically contained conserved cell types and reflected the organ structure in the 3D view (**Fig. 5d**).

We then defined TM clusters 2 and 3 as tissue motif 1 and all three TM clusters as tissue motif 2. Based on the spot label assigned by the TM clusters (**Fig. 5e and Supplementary Table 15**), we found that tissue motif 1 co-occurred and was conserved in all seven samples, which was not affected by sample status and sampling strategies. The overlapped spots between TM clusters 2 and 3 indicated the convolution of elements in tissue motif 1, reflecting a potential collaboration between the hypothalamus region and white matter[43]. Compared with tissue motif 1, tissue motif 2 was a complex conserved structure in the mouse brain that was repeated in six samples rather than seven samples. Furthermore, TM cluster 1 (enriched with excitatory neurons) showed a strong association with TM cluster 2 (enriched with inhibitory neurons), indicating the neuronal circuit activity of inhibitory and excitatory neurons in the hypothalamus region[44,45]. In addition, TM cluster 3 displayed possible collaborations with either TM cluster 1 or 2, indicating potential connectivity among the partial cerebrum, hypothalamus region, and white matter[43,46]. Notably, the two identified tissue motifs could be observed from two anatomical views, which strengthened the claim that tissue motifs contributed to understanding the 3D structures of the cerebrum, hypothalamus, and white matter. Based on multiple anatomical views of mouse brain samples, our results demonstrated that SpaGFT provided a novel gene-centric perspective for investigating conserved TMs among multiple samples and their convolution, and helped to discover insights into the 3D functional structures in a complex organ.

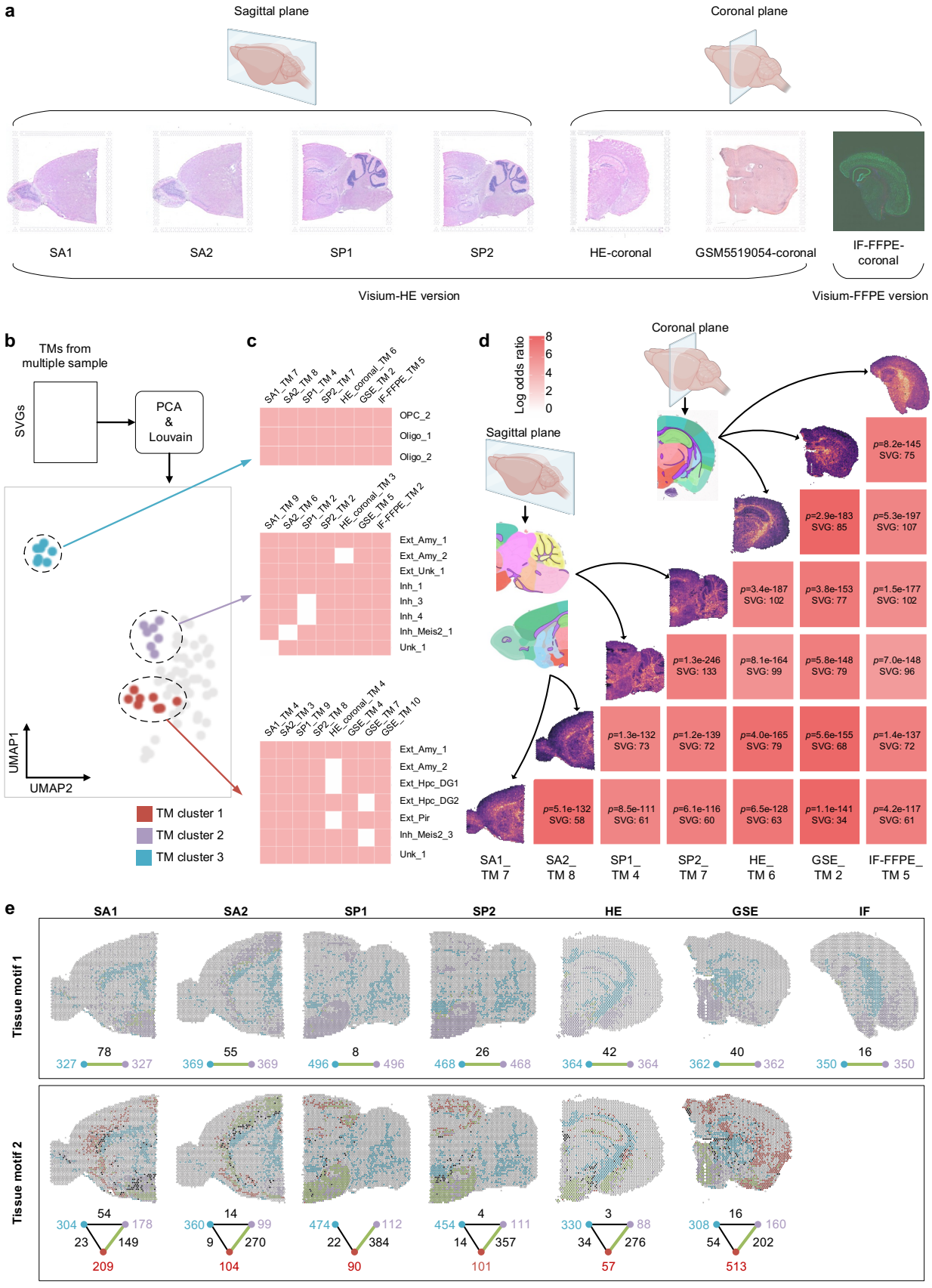

Fig.5 | Case study for tissue motifs in mouse brain based on TM clustering. a. The figure demonstrates seven sample sources, including sagittal (Sagittal Anterior 1, Sagittal Anterior 2, Sagittal posterior 1, Sagittal posterior 2) and coronal planes (HE-coronal, GSM5519054-coronal, and IF-FFPE-coronal). **b**. The pipeline of TM cluster (Methods) identification. UMAP showcases the results of clustering 67 TMs identified by seven samples. **c**. Three heatmaps show the binarized TM-cell type matrix, indicating consistent cell types shared within three TM clusters, where a red-color block means cell type existence in the corresponding TM and a white-color block means non-existence. **d**. Interpretation of TMs from multiple samples with similar SVG components. Seven samples are used to demonstrate the commonality of SVG-similar TMs in multiple samples. Heatmap color indicates the log-odds ratio of the Fisher exact test. The *p*-value (Benjamini–Hochberg adjusted) and the number of shared SVGs between the two samples are shown on the heatmap. White matter anatomical structure is derived from Allen Brain Altas, and was indicated by the purple color. **e**. The figure demonstrates two conserved tissue motifs shared by multiple samples. The spatial map indicates spot localization where a spot is colored according to TM clusters assignment (brown for TM cluster 1; purple for TM cluster 2; and blue TM cluster 3). The tissue motif below each spatial map demonstrates the colocalization of TM clusters. A node represents one specific TM cluster, and the value of the node means the number of spots in the corresponding sample of a TM cluster. An edge will be added if there are existing overlapped spots between the two nodes. The weight of an edge is the number of overlapped spots between the two nodes. The green edge denotes the edge with the largest weight in one tissue motif.

Discussion

We present SpaGFT as a fast and accurate SVG identifier and a novel computational formulation for TM characterization using spatial transcriptome data. For the first time, SpaGFT introduced a graph Fourier transform ideology to transform complex spatial gene expression signals into informative FM signals from a gene-centric perspective. The benchmarking results of 31 spatial transcriptome data revealed that SpaGFT achieved superior SVG detection performances compared to existing tools, indicating that the FM signals can effectively capture gene expression signals spatially and distinguish SVGs from non-SVGs. In addition, TMs defined by SVG clusters in SpaGFT were confirmed to maintain diverse TM-associated biological processes, and we demonstrated that SpaGFT can effectively complement the cell/spot-centric tool (e.g., cell2location) for investigating molecular tissue biology. Moreover, three case studies provided biological insights from the TM ID card and demonstrated the capability of identifying short length-scale TMs, convoluted TM collaborations, and fundamental elements constituting the complex organ in the 3D structure. Furthermore, the tissue motif concept was originally proposed using high-resolution spatial proteomic data at the cellular level[9], defined as basic structural units (a small region containing simple cell types), and played an important role in propagating biological signals (e.g., molecular diffusion or cellular movement) to support organ functions. We extended this concept to spatial transcriptomics data and demonstrated that TMs conserved in multiple samples could form tissue motifs that allowed us to study convoluted collaborations among TMs and the 3D structure of functional regions (i.e., white matter and hypothalamus region in the mouse brain) from multiple anatomical views.

Overall, SpaGFT is a computational framework geared towards the accurate identification and characterization of a TM, which may significantly enhance our understanding of molecular tissue biology. However, there is still room for improving prediction performance and understanding the TM mechanism. First, although the SpaGFT computation speed is very competitive, it can be further improved by reducing the computational complexity from $O(n^2)$ to $O(n \times log(n))$ using fast Fourier transform algorithms[47,48]. Second, the alteration of the spot graph and TM topology represents a potential challenge in identifying TMs across spatial samples from different tissues or experiments, which results in diverse FM signal spaces and renders the FM signals incomparable. This is similar to the "batch effect" issue in multiple single-cell RNA sequencing (scRNA-seq) integration analyses[49].

SpaGFT bridges the gap left by existing SVG prediction methods and provides a method for investigating molecular tissue biology from the gene-centric perspective. In the future, we expect that SpaGFT could potentially be used for spatial multi-omics data harmonization and integration by discovering conserved spatial FM signal patterns of metabolic, proteomic, morphogenetic, and epigenetics in nature in both healthy and pathological state. Meanwhile, there is an increasing need for building connections between spatial spots using multi-omics at the single-cell level[1]. Based on the SpaGFT framework, it is feasible to decompose the graph signals to match the spots with single cells using the graph Fourier transform to align spatial TMs with single cells. Such an alignment can provide further insight into understanding the underlying gene regulatory networks in TMs and facilitate the identification of cell-cell communications using the spatial information within a TM or between TMs.

References

1 Palla, G., Fischer, D. S., Regev, A. & Theis, F. J. Spatial components of molecular tissue biology. Nature Biotechnology, doi:10.1038/s41587-021-01182-1 (2022).

2 Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature Biotechnology 39, 313-319, doi:10.1038/s41587-020-0739-1 (2021).

3 Method of the Year 2020: spatially resolved transcriptomics. Nature Methods 18, 1-1, doi:10.1038/s41592-020-01042-x (2021).

4 Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. Nature Reviews Molecular Cell Biology, doi:10.1038/s41580-022-00466-x (2022).

5 Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biology 22, 184, doi:10.1186/s13059-021-02404-0 (2021).

6 Liao, J., Lu, X., Shao, X., Zhu, L. & Fan, X. Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. Trends in Biotechnology, doi:10.1016/j.tibtech.2020.05.006 (2020).

7 Lewis, S. M. et al. Spatial omics and multiplexed imaging to explore cancer biology. Nature Methods 18, 997-1012, doi:10.1038/s41592-021-01203-6 (2021).

8 Ricaud, B., Borgnat, P., Tremblay, N., Gonçalves, P. & Vandergheynst, P. Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs. Comptes Rendus Physique 20, 474-488, doi:https://doi.org/10.1016/j.crhy.2019.08.003 (2019).

9 Bhate, S. S., Barlow, G. L., Schürch, C. M. & Nolan, G. P. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. Cell Systems 13, 109-130.e106, doi:https://doi.org/10.1016/j.cels.2021.09.012 (2022).

10 Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nature Methods, doi:10.1038/s41592-022-01480-9 (2022).

11 Ortiz, C. et al. Molecular atlas of the adult mouse brain. Sci Adv 6, eabb3446, doi:10.1126/sciadv.abb3446 (2020).

12 Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. Nature 573, 61-68, doi:10.1038/s41586-019-1506-7 (2019).

13 Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72-78, doi:10.1038/s41586-018-0654-5 (2018).

14 Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience 19, 335-346, doi:10.1038/nn.4216 (2016).

15 Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience 24, 425-436, doi:10.1038/s41593-020-00787-0 (2021).

16 Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168-176, doi:10.1038/nature05453 (2007).

17 Gouwens, N. W. et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. Nature Neuroscience 22, 1182-1195, doi:10.1038/s41593-019-0417-0 (2019).

18 Camillo, D., Levelt, C. N. & Heimel, J. A. Lack of functional specialization of neurons in the mouse primary visual cortex that have expressed calretinin. Front Neuroanat 8, 89-89, doi:10.3389/fnana.2014.00089 (2014).

19 McGregor, R., Wu, M.-F., Barber, G., Ramanathan, L. & Siegel, J. M. Highly Specific Role of Hypocretin (Orexin) Neurons: Differential Activation as a Function of Diurnal Phase, Operant Reinforcement versus Operant Avoidance and Light Level. The Journal of Neuroscience 31, 15455, doi:10.1523/JNEUROSCI.4017-11.2011 (2011).

20 Barrera, G. et al. One for all or one for one: does co-transmission unify the concept of a brain galanin “system” or clarify any consistent role in anxiety? Neuropeptides 39, 289-292, doi:https://doi.org/10.1016/j.npep.2004.12.008 (2005).

21 Vagena, E. et al. ASB4 modulates central melanocortinergic neurons and calcitonin signaling to control satiety and glucose homeostasis. Science Signaling 15, eabj8204, doi:doi:10.1126/scisignal.abj8204 (2022).

22 Sunkin, S. M. et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic acids research 41, D996-D1008, doi:10.1093/nar/gks1042 (2013).

23 Desai, D. & Pethe, P. Polycomb repressive complex 1: Regulators of neurogenesis from embryonic to adult stage. J Cell Physiol 235, 4031-4045, doi:10.1002/jcp.29299 (2020).

24 Strekalova, T. et al. Memory retrieval after contextual fear conditioning induces c-Fos and JunB expression in CA1 hippocampus. Genes, Brain and Behavior 2, 3-10 (2003).

25 Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nature Biotechnology, doi:10.1038/s41587-021-01139-4 (2022).

26 Erö, C., Gewaltig, M.-O., Keller, D. & Markram, H. A Cell Atlas for the Mouse Brain. Frontiers in Neuroinformatics 12, doi:10.3389/fninf.2018.00084 (2018).

27 Bakken, T. E. et al. Single-cell and single-nucleus RNA-seq uncovers shared and distinct axes of variation in dorsal LGN neurons in mice, non-human primates, and humans. eLife 10, e64875, doi:10.7554/eLife.64875 (2021).

28 Ichise, M. et al. Leucine-Rich Repeats and Transmembrane Domain 2 Controls Protein Sorting in the Striatal Projection System and Its Deficiency Causes Disturbances in Motor Responses and Monoamine Dynamics. Frontiers in molecular neuroscience 15, 856315-856315, doi:10.3389/fnmol.2022.856315 (2022).

29 Genomics, X. Spatial Gene Expression Datasets, <https://www.10xgenomics.com/resources/datasets/> (2020).

30 Klein, U. et al. Transcriptional analysis of the B cell germinal center reaction. Proceedings of the National Academy of Sciences of the United States of America 100, 2639-2644, doi:10.1073/pnas.0437996100 (2003).

31 Holmes, A. B. et al. Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome. J Exp Med 217, doi:10.1084/jem.20200483 (2020).

32 Link, A. et al. Fibroblastic reticular cells in lymph nodes regulate the homeostasis of naive T cells. Nature Immunology 8, 1255-1265, doi:10.1038/ni1513 (2007).

33 Medaglia, C. et al. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. Science 358, 1622-1626, doi:10.1126/science.aao4277 (2017).

34 Qi, H., Egen, J. G., Huang, A. Y. C. & Germain, R. N. Extrafollicular Activation of Lymph Node B Cells by Antigen-Bearing Dendritic Cells. Science 312, 1672-1676, doi:doi:10.1126/science.1125703 (2006).

35 Havenar-Daughton, C. et al. CXCL13 is a plasma biomarker of germinal center activity. Proceedings of the National Academy of Sciences 113, 2702-2707, doi:doi:10.1073/pnas.1520112113 (2016).

36 Crotty, S. T follicular helper cell differentiation, function, and roles in disease. Immunity 41, 529-542, doi:10.1016/j.immuni.2014.10.004 (2014).

37 Qi, H., Cannons, J. L., Klauschen, F., Schwartzberg, P. L. & Germain, R. N. SAP-controlled T–B cell interactions underlie germinal centre formation. Nature 455, 764-769, doi:10.1038/nature07345 (2008).

38 Vinuesa, C. G., Linterman, M. A., Yu, D. & MacLennan, I. C. M. Follicular Helper T Cells. Annual Review of Immunology 34, 335-368, doi:10.1146/annurev-immunol-041015-055605 (2016).

39 Fillatreau, S. & Gray, D. T cell accumulation in B cell follicles is regulated by dendritic cells and is independent of B cell activation. The Journal of experimental medicine 197, 195-206, doi:10.1084/jem.20021750 (2003).

40 Pae, J., Jacobsen, J. T. & Victora, G. D. Imaging the different timescales of germinal center selection. Immunol Rev 306, 234-243, doi:10.1111/imr.13039 (2022).

41 Reticker-Flynn, N. E. et al. Lymph node colonization induces tumor-immune tolerance to promote distant metastasis. Cell 185, 1924-1942.e1923, doi:10.1016/j.cell.2022.04.019 (2022).

42 Buzzi, R. M. et al. Spatial transcriptome analysis defines heme as a hemopexin-targetable inflammatoxin in the brain. Free Radic Biol Med 179, 277-287, doi:10.1016/j.freeradbiomed.2021.11.011 (2022).

43 Lemaire, J.-J. et al. White matter connectivity of human hypothalamus. Brain Research 1371, 43-64, doi:https://doi.org/10.1016/j.brainres.2010.11.072 (2011).
44 Belousov, A. B., O'Hara, B. F. & Denisova, J. V. Acetylcholine becomes the major excitatory neurotransmitter in the hypothalamus in vitro in the absence of glutamate excitation. J Neurosci 21, 2015-2027, doi:10.1523/jneurosci.21-06-02015.2001 (2001).
45 Mickelsen, L. E. et al. Single-cell transcriptomic analysis of the lateral hypothalamic area reveals molecularly distinct populations of inhibitory and excitatory neurons. Nature Neuroscience 22, 642-656, doi:10.1038/s41593-019-0349-8 (2019).
46 Bassett, D. S., Brown, J. A., Deshpande, V., Carlson, J. M. & Grafton, S. T. Conserved and variable architecture of human white matter connectivity. NeuroImage 54, 1262-1279, doi:https://doi.org/10.1016/j.neuroimage.2010.09.006 (2011).
47 Le Magoarou, L., Gribonval, R. & Tremblay, N. Approximate fast graph fourier transforms via multilayer sparse approximations. IEEE transactions on Signal and Information Processing over Networks 4, 407-420 (2017).
48 Lu, K.-S. & Ortega, A. Fast graph Fourier transforms based on graph symmetry and bipartition. IEEE Transactions on Signal Processing 67, 4855-4869 (2019).
49 Lähnemann, D. et al. Eleven grand challenges in single-cell data science. Genome Biology 21, 31, doi:10.1186/s13059-020-1926-6 (2020).

Online Methods

We introduce Spatial Graph Fourier Transform (SpaGFT) to identify SVGs and characterize TMs based on spatial transcriptomics data. The core concept of SpaGFT is to transform spatial gene expressions into a kind of frequency signals in Fourier space. The main framework of SpaGFT includes three major steps: graph signal transform, SVG identification, and TM characterization.

Graph signal transform

K-nearest neighbor (KNN) Graph construction. Given a gene expression matrix containing n spots, including their spatial coordinates and m genes, SpaGFT calculates the Euclidean distances between each pair of spots based on spatial coordinates first. In the following, an undirected graph $G = (V, E)$ will be constructed, where $V = \{v_1, v_2, \dots, v_n\}$ is the node set corresponding to n spots; E is the edge set while there exists an edge e_{ij} between v_i and v_j in E if and only if v_i is the KNN of v_j or v_j is the KNN of v_i based on Euclidean distance, where $i, j = 1, 2, \dots, n$; and $i \neq j$. Note that, all the notations of matrices and vectors are bolded, and all the vectors are treated as column vectors in the following description. An adjacent binary matrix $\boldsymbol{A} = (a_{ij})$ with rows and columns as n spots is defined as:

$$a_{ij} = \begin{cases} 1, & e_{ij} \in E \\ 0, & \text{else.} \end{cases} \quad (1)$$

A diagonal matrix $\boldsymbol{D} = diag(d_1, d_2, \dots, d_n)$, where $d_i = \sum_{j=1}^{n} a_{ij}$ represents the degree of v_i.

Fourier mode calculation. Using matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, a Laplacian matrix $\boldsymbol{L}$ can be obtained by

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}. \quad (2)$$

The Laplacian matrix $\boldsymbol{L}$ can be decomposed using spectral decomposition

$$\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathbf{T}} \quad (3)$$

$$\boldsymbol{\Lambda} = \boldsymbol{diag}(\lambda_1, \lambda_2, \dots, \lambda_n),$$
$$\boldsymbol{U} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_n),$$

where the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of $\boldsymbol{L}$ with $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, and the columns of $\boldsymbol{U}$ are the unit eigenvector of $\boldsymbol{L}$. $\boldsymbol{\mu}_k$ is the k^{th} Fourier mode (FM), $\boldsymbol{\mu}_k \in \mathbb{R}^n$, $k = 1, 2, \dots, n$, and the set $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k\}$ is an orthogonal basis for the linear space (**Supplementary Figs. 1a and 1b**). For $\boldsymbol{\mu}_k = (\mu_k^1, \mu_k^2, \dots, \mu_k^n)$, where μ_k^i indicates the value of the k^{th} FM on node v_i, the smoothness of $\boldsymbol{\mu}_k$ reflects the total variation of the k^{th} FM in all mutual adjacent spots, which can be formulated as

$$\frac{1}{2}\sum_{v_i \in V}\sum_{v_j \in V} a_{ij}\left(\mu_k^i - \mu_k^j\right)^2. \tag{4}$$

The form can be derived by matrix multiplication as

$$\begin{aligned}
\frac{1}{2}\sum_{v_i \in V}\sum_{v_j \in V} a_{ij}\left(\mu_k^i - \mu_k^j\right)^2 &= \frac{1}{2}\left[\sum_{v_i \in V} d_i\left(\mu_k^i\right)^2 - 2\sum_{v_i \in V}\sum_{v_j \in V} a_{ij}\mu_k^i\,\mu_k^j + \sum_{v_j \in V} d_j\left(\mu_k^j\right)^2\right] \\
&= \sum_{v_i \in V} d_i\left(\mu_k^i\right)^2 - \sum_{v_i \in V}\sum_{v_j \in V} a_{ij}\mu_k^i\,\mu_k^j \\
&= \boldsymbol{\mu}_k^T \boldsymbol{D}\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \boldsymbol{A}\boldsymbol{\mu}_k \\
&= \boldsymbol{\mu}_k^T \boldsymbol{L}\boldsymbol{\mu}_k \\
&= \lambda_k
\end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_k^T$ is the transpose of $\boldsymbol{\mu}_k$. According to the definition of smoothness, if an eigenvector corresponds to a small eigenvalue, it indicates the variation of FM values on adjacent nodes is low. The increasing trend of eigenvalues corresponds to an increasing trend of oscillations of eigenvectors; hence, the eigenvalues and eigenvectors of $\boldsymbol{L}$ are used as frequencies and FMs in our SpaGFT, respectively. Intuitively, a small eigenvalue corresponds to a low-frequency FM, while a large eigenvalue corresponds to a high-frequency FM.

Graph Fourier transform. The graph signal of a gene g is defined as $\boldsymbol{f}_g = \left(f_g^1, f_g^2, \dots, f_g^n\right) \in \mathbb{R}^n$, which is a n-dimensional vector and represents the gene expression values across n spots. The graph signal $\boldsymbol{f}_g$ is transformed to a frequency signal $\hat{\boldsymbol{f}}_g$ by

$$\hat{\boldsymbol{f}}_g = \boldsymbol{U}\boldsymbol{f}_g,\ \hat{\boldsymbol{f}}_g = \left(\hat{f}_g^1, \hat{f}_g^2, \dots, \hat{f}_g^n\right). \tag{6}$$

In such a way, $\hat{f}_g^k$ is the projection of $\boldsymbol{f}_g$ on FM $\boldsymbol{\mu}_k$, representing the contribution of FM $\boldsymbol{\mu}_k$ to graph signal $\boldsymbol{f}_g$, $k = 1, 2, \dots, n$.This Fourier transform harmonizes gene expression and its spatial distribution to represent gene g in the frequency domain. The details of SVG identification using $\hat{\boldsymbol{f}}_g$ can be found below.

SVG identification

We designed a *GFTscore* to quantitatively measure the randomness of gene expressions distributed in the spatial domain, defined as

$$GFTscore\left(f_g\right) = \sum_{k=1}^{n} 2^{-2\lambda_k}\,\tilde{f}_g^k, \tag{7}$$

where λ_k is the pre-calculated eigenvalue of $\boldsymbol{L}$, and the normalized frequency signal $\tilde{f}_g^k$ is defined as:

$$\tilde{f}_g^k = \frac{|\hat{f}_g^k|}{\sum_{i=1}^{n} |\hat{f}_g^i|}. \tag{8}$$

The gene with a high *GFTscore* tends to be a non-random distributed gene in the spatial domain, and vice versa. Therefore, all m genes are decreasingly ranked based on their *GFTscore* from high to low and denote these *GFTscore* values as $y_1 \geq y_2 \geq \cdots \geq y_m$. In order to determine the cutoff y_z to distinguish SVG and non-SVGs based on *GFTscore*, we applied the Kneedle algorithm[50] to search for the inflection point of a *GFTscore* curve described below. The *GFTscore* y_t of gene g_t is converted by $y_{c_t} = max\{y_1, y_2, \dots, y_m\} - y_t, t = 1,2, \dots, m$, where y_{c_t} is the converted value of y_t. Each point $(x_{c_t} = t,\ y_{c_t})$, where x_{c_t} is the rank number of y_{c_t}, is processed by a smoothing spline to preserve the curve shape and obtain $\left(x_{s_t}, y_{s_t}\right), t = 1,2, \dots, m$. Denote coordinate set $\mathcal{D}_s = \left\{\left(x_{s_t}, y_{s_t}\right) \middle| t = 1,2, \dots, m\right\}$ and can be normalized to corrdinate set $\mathcal{D}_n$ as follows:

$$\begin{aligned} \mathcal{D}_n &= \left\{\left(x_{n_t}, y_{n_t}\right) \middle| t = 1, 2, \dots, m\right\} \\ x_{n_t} &= \left(x_{s_t} - min(x_s)\right) / \left(max(x_s) - min(x_s)\right) \\ y_{n_t} &= \left(y_{s_t} - min(y_s)\right) / \left(max(y_s) - min(y_s)\right), \end{aligned} \tag{9}$$

where $min(x_s)$ and $max(x_s)$ are the minimum and maximum in $\left\{x_{s_1}, x_{s_2}, \dots, x_{s_m}\right\}$, respectively. Analogously, $min(y_s)$ and $max(y_s)$ are the minimum and maximum in $\left\{y_{s_1}, y_{s_2}, \dots, y_{s_m}\right\}$, respectively. In addition, let $\mathcal{D}_d$ represents the set of differences between the x- and y-values, and one has:

$$\mathcal{D}_d = \left\{\left(x_{d_t}, y_{d_t}\right) \middle| x_{d_t} = x_{n_t}, y_{d_t} = y_{n_t} - x_{n_t}, t = 1, 2, \dots, m\right\}. \tag{10}$$

In the following, the question of determining the cutoff y_z can be converted to the determination of the inflection point y_z if it satisfies $y_{d_{z-1}} < y_{d_z}$, $y_{d_{z+1}} < y_{d_z}$, and $y_{d_h} < T_z, h = z, z+1, \dots, m$, where

$$T_z = y_{d_z} - S\frac{s_{n_t} - s_{n_1}}{t - 1}. \tag{11}$$

In equation (11), S is a coefficient that can be used to determine the level of aggression for the inflection point.

A non-parametrical test is used for testing the difference between median values of low-frequency signals and high-frequency signals. Especially, the null hypothesis is that the median of low-frequency signals of a SVG is equal to or lower than the median of high-frequency elements. The alternative hypothesis is that the median of low-frequency signals of a SVG is higher than the median of high-frequency signals. The *p*-value of each gene is calculated based on *Wilcoxon* one-sided rank-sum test and then adjusted using the false discovery rate (FDR) method. Eventually, a gene with *GFTscore* higher than y_z and adjusted *p*-value less than 0.05 is considered as an SVG.

Visualization of frequency signal of SVGs in low-dimensional latent spaces

The novel SVG presentation $\hat{\boldsymbol{f}}_g$ is a simple and distinguishable one-dimensional vector. It can be visualized in two-dimensional space. First, frequency signals were computed by SpaGFT based on optimized parameters. Second, the top $2\sqrt{n}$ low-frequency signals were selected and then followed by $L1$ normalization method to normalize selected low-frequency signals.

TM identification and characterization

SVGs with similar patterns also have similar low-frequency signals in the frequency domain, which provides the fundamental basis of clustering. Louvain clustering method was applied to group SVGs based on the top $2\sqrt{n}$ low-frequency signals in the frequency domain. For a total number of p SVGs donated as $g_1, g_2, \dots, g_p$ in a SVG cluster, a pseudo-expression value $Pseudo^i$ for spot i can be calculated as

$$Pseudo^i = \sum_{l=1}^{p} log\left(1 + f_{g_l}^i\right), \tag{12}$$

where $i = 1, 2, \dots, n.$ The pseudo-expression value was further transformed into a binary value by

$$Binary^i = \begin{cases} 1, & \text{if } Pseudo^i > cutoff, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where the $cutoff$ is a given number of percentiles of the pseudo-expression across all spots. We define those spots with binary-expression as 1 as a TM, and the corresponding SVGs are TM-associated SVGs. The pseudo-expression and binary-expression profiles can be visualized in a spatial map and TM map, respectively. The low-frequency signals of TM are calculated using Pseudo-expression values by SpaGFT.

SVG signal enhancement

A SVG of a specific TM may suffer from low expression or dropout issues. To solve this problem, SpaGFT implemented the low-pass filter to enhance the SVG expressions. For a SVG with a measured expression value $\boldsymbol{f}_g \in \mathbb{R}^n$, we define $\bar{\boldsymbol{f}}_g \in \mathbb{R}^n$ as the expected expression value of this SVG, and $\boldsymbol{f}_g = \bar{\boldsymbol{f}}_g + \boldsymbol{\epsilon}_g$, where $\boldsymbol{\epsilon}_g \in \mathbb{R}^n$ represents noises. SpaGFT estimates an approximated $\boldsymbol{f}_g^\star$ to $\bar{\boldsymbol{f}}_g$ in the following way, resisting the noise $\boldsymbol{\epsilon}_g$. The approximation has two requirements (*i*) the enhanced signal (estimated gene expression) should be similar to the measured signals, and (*ii*) keep low variation within estimated gene expression to prevent inducing new noises. Therefore, the following optimization problem is proposed to find an optimal solution $\boldsymbol{f}_g^\star$ for $\bar{\boldsymbol{f}}_g$

$$\begin{aligned} \boldsymbol{f}_g^\star &= argmin_{\boldsymbol{f}}\left[||\boldsymbol{f} - \boldsymbol{f}_g||^2 + c\frac{1}{2}\sum_{\boldsymbol{v}_i \in V}\sum_{\boldsymbol{v}_j \in V} a_{ij}\left(f^i - f^j\right)^2\right] \\ &= argmin_{\boldsymbol{f}}\left[||\boldsymbol{f} - \boldsymbol{f}_g||^2 + c\boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f}\right] \end{aligned} \tag{14}$$

where $\|\cdot\|$ is the $L2$-norm, $\boldsymbol{f} = (f^1, f^2, \dots, f^n) \in \mathbb{R}^n$ is the variable in solution space, and $i, j = 1, 2, \dots, n.$ c is a coefficient to determine the importance of variation of the estimated signals, and $c > 0$. According to the convex optimization, the optimal solution $\boldsymbol{f}_g^\star$ can be formulated as:

$$
\begin{aligned}
&2\left(\boldsymbol{f}_g^{\star}-\boldsymbol{f}_g\right)+2c\boldsymbol{L}\boldsymbol{f}_g^{\star}=0\\
&\Longrightarrow(\boldsymbol{I}+c\boldsymbol{L})\boldsymbol{f}_g^{\star}=\boldsymbol{f}_g\\
&\Longrightarrow(\boldsymbol{U}\boldsymbol{U}^T+c\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)\boldsymbol{f}_g^{\star}=\boldsymbol{f}_g\\
&\Longrightarrow\boldsymbol{U}(\boldsymbol{I}+c\boldsymbol{\Lambda})\boldsymbol{U}^T\boldsymbol{f}_g^{\star}=\boldsymbol{f}_g\\
&\Longrightarrow\boldsymbol{f}_g^{\star}=\boldsymbol{U}(\boldsymbol{I}+c\boldsymbol{\Lambda})^{-1}\boldsymbol{U}^T\boldsymbol{f}_g=\boldsymbol{U}(\boldsymbol{I}+c\boldsymbol{\Lambda})^{-1}\hat{\boldsymbol{f}}_g
\end{aligned}
\tag{15}
$$

where $\boldsymbol{\Lambda}=diag(\lambda_1,\ \lambda_2,\ \ldots,\lambda_n)$, and $\boldsymbol{I}$ is an identity matrix. $(\boldsymbol{I}+c\boldsymbol{\Lambda})^{-1}$ is the low-pass filter and $(\boldsymbol{I}+c\boldsymbol{\Lambda})^{-1}\hat{\boldsymbol{f}}_g$ is the enhanced frequency signal. $\boldsymbol{f}_g^{\star}=\boldsymbol{U}(\boldsymbol{I}+c\boldsymbol{\Lambda})^{-1}\hat{\boldsymbol{f}}_g$ represents the enhanced SVG expression by inverse graph Fourier transform.

Benchmarking data setup

Dataset description. Thirty-two spatial transcriptome datasets were collected from the public domain, including 30 10X Visium datasets (18 human brain data, 11 mouse brain data, and one human lymph node data) and two Slide-seq V2 datasets (mouse brain). More details can be found in **Supplementary Table 1**. Three datasets were selected as the training sets for grid-search parameter optimization in SpaGFT, including two highest read-depth datasets in Visium (HE-coronal) and Slide-seq V2 (Puck-200115-08), one signature dataset in Maynard's study[15]. The rest of the 28 datasets (excluding lymph node) were used as independent test datasets.

Data preprocessing. For all the 32 datasets, we adopt the same preprocessing steps based on *scanpy*[51] and *squidpy*[52] (version 1.2.1), including filtering genes that have expression values in less than ten spots, normalizing the raw count matrix by counts per million reads method, and implementing log-transformation to the normalized count matrix. No specific preprocessing step was performed on the spatial location data.

Benchmarking SVG collection. We collected SVG candidates from five publications[11-15], with data from either human or mouse brain subregions. (*i*) A total number of 130 layer signature genes were collected from Maynard's study[15]. These genes are potential multiple-layer markers validated in the human dorsolateral prefrontal cortex region. (*ii*) A total number of 397 cell-type-specific (CTS) genes in the adult mouse cortex were collected from the Tasic's study (2016 version)[14]. The authors performed scRNA-seq on the dissected target region, and identified 49 cell types, and constructed a cellular taxonomy of the primary visual cortex in the adult mouse. (*iii*) A total number of 182 CTS genes in mouse neocortex were collected from the Tasic's study[13]. Altogether, 133 cell types were identified from multiple cortical areas at single-cell resolution. (*iv*) A total number of 260 signature genes across different major regions of the adult mouse brain were collected from the Ortiz's study[11]. The authors' utilized spatial transcriptomics data to systematically profile subregions and delivered the subregional genes using consecutive coronal tissue sections. (*v*) A total of 86 signature genes in the cortical region shared by humans and mice were collected from the Hodge's study[12]. Collectively, a total number of 849 genes were obtained, among which 153 genes were documented by multiple papers. More details, such as gene names, targeted regions, and sources, can be found in **Supplementary Table 2**.

Next, the above 849 genes were manually validated on the in-situ hybridization (ISH) database deployed on the Allen Brain Atlas (https://mouse.brain-map.org/). The ISH database provided ISH

mouse brain data across 12 anatomical structures (i.e., Isocortex, Olfactory area, Hippocampal formation, Cortical subplate, Striatum, Pallidum, Thalamus, Hypothalamus, Midbrain, Pons, Medulla, and Cerebellum). We filtered the 849 genes as follows: (*i*) If a gene is showcased in multiple anatomical plane experiments (i.e., coronal plane and sagittal plane), it will be counted multiple times with different expressions in the corresponding experiments. Such that, 1,327 genes were archived (**Supplementary Table 3**). (*ii*) All 1,327 genes were first filtered by low gene expressions (cutoff is 1.0), and the *FindVariableFeatures* function ("*vst"* method) in the Seurat (v4.0.5) was used for identifying highly variable genes across twelve anatomical structures. Eventually, 458 genes were kept and considered as curated benchmarking SVGs.

SpaGFT implementation and grid search of parameter optimization

Herein, partial FMs were used, including low-frequency FMs, which reflect smooth spatial patterns, and high-frequency FMs, which can measure noises. And such a scheme reduced running time significantly. We set $K = \sqrt{n}/2$ as the default parameter for constructing the KNN graphs in SpaGFT. The number of selected low-frequency signals was set to be $\sqrt{n}/2$, and the high-frequency FMs were set to be $3\sqrt{n}$. These elements with low values in the frequency domain were filtered out in the *rank_gene_smooth* function. SVGs were determined by genes with high *GFTscore* via the *KneeLocator* function (curve='convex', direction='deceasing', and S=5) in the *kneed* package (version 0.7.0) and FDR (cutoff is less than 0.05). To obtain the optimized parameters of SpaGFT, we set a grid-search test for six parameters, including *ratio_neighbors* (1, 2) for KNN selection, *normalize_lap* (TRUE or FALSE) for Laplacian matrix normalization, *filter_peaks* (TRUE or FALSE) for noise low-frequency signal filtering, *ratio_low_freq* (0.5, 1, 1.5, 2) for the number of low-frequency signals, *ratio_high_freq* (1, 2, 3) for the number of high-frequency signals, and *S* (3, 5, 10) for the inflection point coefficient, resulting in 288 parameter combinations. Detailed implementation and tutorial can be found on SpaGFT GitHub: https://github.com/OSU-BMBL/SpaGFT.

Parameter setting of other tools

(*i*) SpatialDE (version 1.1.3) is a method for identifying and describing SVGs based on Gaussian process regression used in geostatistics. *SpatialDE* consists of four steps, establishing SpatialDE model, predicting statistical significance, selecting the model, and expressing histology automatically. We selected two key parameters, *design_formula* ('0' and '1') in the *NaiveDE.regress_out* function and *kernel_space* ("{'SE':[5.,25.,50.],'const':0}", "{'SE':[6.,16.,36.],'const':0}", "{'SE':[7.,47.,57.],'const':0}", "{'SE':[4.,34.,64.],'const':0}", "{'PER':[5.,25.,50.],'const':0}", "{'PER':[6.,16.,36.],'const':0}", "{'PER':[7.,47.,57.],'const':0}", "{'PER':[4.,34.,64.],'const':0}",and "{'linear':0,'const':0}") in the *SpatialDE.run* function for parameter tunning, resulting in 18 parameter combinations.

(*ii*) SPARK (version 1.1.1) is a statistical method for spatial count data analysis through generalized linear spatial models. Relying on statistical hypothesis testing, SPARX identifies SVGs via predefined kernels. First, raw count and spatial coordinates of spots were used to create the SPARK object via filtering low-quality spots (controlled by *min_total_counts*) or genes (controlled by *percentage*). Then the object was followed by fitting the count-based spatial model to estimate the parameters via *spark.vc* function, which is affected by the number of iterations

(*fit.maxiter*) and models (*fit.model*). Lastly, ran *spark.test* function to test multiple kernel matrices and obtain the results. We selected four key parameters, *percentage* (0.05, 0.1, 0.15), *min_total_counts* (10, 100, 500) in *CreateSPARKObject* function, *fit.maxiter* (300, 500, 700), and *fit.model* ("poisson", "gaussian") in *spark.vc* function for parameter tunning, resulting in 54 parameter combinations.

(*iii*) SPARK-X (version 1.1.1) is a non-parametric method that tests whether the expression level of the gene
displays any spatial expression pattern via a general class of covariance tests. We selected three key parameters, *percentage* (0.05, 0.1, 0.15), *min_total_counts* (10, 100, 500) in the *CreateSPARKObject* function, and *option* ("single", "mixture") in the *sparkx* function for parameter tunning, resulting in 18 parameter combinations.

(*iv*) SpaGCN (version 1.2.0) is a graph convolutional network approach that integrates gene expression, spatial location, and histology in spatial transcriptomics data analysis. *SpaGCN* consisted of four steps, integrating data into a chart, setting graph convolutional layer, detecting spatial domains by clustering, and identifying SVGs in spatial domains. We selected two parameters, the value of ratio (1/3, 1/2, 2/3, and 5/6) in the *find_neighbor_cluster* function and *res* (0.8, 0.9, 1.0, 1.1, and 1.2) in the *SpaGCN.train* function for parameter tunning, resulting in 20 parameter combinations.

(*v*) MERINGUE (version 1.0) is a computational framework based on spatial autocorrelation and cross-correlation analysis. It composes of three major steps to identify SVGs. Firstly, Voronoi tessellation was utilized to partition the graph to reflect the length scale of cellular density. Secondly, the adjacency matrix is defined using geodesic distance and the partitioned graph. Finally, gene-wise autocorrelation (e.g., Moran's I) is conducted, and a permutation test is performed for significance calculation. We selected *min.read* (100, 500, 1000), *min.lib.size* (100, 500, 1000) in the *cleanCounts* function and *filterDist* (1.5, 2.5, 3.5, 7.5, 12.5, 15.5) in the *getSpatialNeighbors* function for parameter tunning, resulting in 54 parameter combinations .

Metrics used in benchmarking experiments

Denote $P = \{p_1, p_2, \dots, p_p\}$, where p is the total number of SVGs predicted by a tool in the performance comparison. The set of 458 curated benchmarking SVGs denoted as $R = \{r_1, r_2, \dots, r_t\}$, where $t = 458$. In addition, C is the complete collection of all genes in a dataset. In addition, some notions are necessary to understand the following metrics, including, $TP = |P \cap R|$, $FP = |P - P \cap R|$, $TN = |(C - P) \cap (C - R)|$ and $FN = |R - P \cap R|$, where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. The following metrics were used to test the performances of various methods. All scores were calculated using customed scripts unless specifically mentioned.

(*i*) *Jaccard index*, also named the Jaccard similarity coefficient, is used to compare the similarity and difference between limited sample sets. Define the Jaccard index between sets P and R as:

$$Jaccard = \frac{|P \cap R|}{|P \cup R|}, Jaccard \in [0,1]$$

A larger $Jaccard$ index indicates a higher similarity between the two sets.

(*ii*) *Odds ratio* of Fisher exact test is a statistical significance test used in the analysis of contingency tables, whose definition is:

$$odds\ ratio = \frac{TP/FP}{TN/FN} = \frac{TP \cdot FN}{TN \cdot FN}$$

A higher $odds\ ratio$ indicates a better prediction performance. Function *newGeneOverlap* and *testGeneOverlap* from R package *GeneOverlap* (Version 1.26.0) were used for the score calculation.

(*iii*) *Precision* (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, which is defined as:

$$precision = \frac{TP}{TP + FP}, precision \in [0,1]$$

A higher precision indicates that an algorithm returns more relevant results than irrelevant ones.

(*iv*) *Recall* (also known as sensitivity) is the fraction of relevant instances that were retrieved, which is defined as:

$$recall = \frac{TP}{TP + FN}, recall \in [0,1]$$

A higher recall indicates that an algorithm returns most of the relevant results (whether irrelevant ones are also returned).

(*v*) $F1$ score is a measure of a test's accuracy in statistical analysis of binary classification. It is calculated from the precision and recall of the test, defined as:

$$F1 = 2\frac{presion * recall}{presion + recall}, F1 \in [0,1]$$

A higher $F1$ score indicates a better prediction performance of the algorithm.

(*vi*) *Tversky index* is an asymmetric similarity measure on sets that compares a variant to a prototype, defined as

$$Tversky = \frac{TP}{TP + 0.5FN + 0.5FP}, Tversky \in [0,1]$$

A higher $Tversky$ signifies the better prediction performance of the algorithm. The *tversky.index* function from R package *tcR* (Version 2.3.2) was used for calculating the Tversky index.

(*vii*) *Moran's Index* used statistics to quantify the degree of spatial autocorrelation, defined as

$$I = \frac{n}{W}\frac{\sum_i \sum_j [w_{ij}(f_g^i - \bar{f_g})(f_g^j - \bar{f_g})]}{\sum_i (f_g^i - \bar{f_g})^2}, I \in [-1, 1]$$

where $f_g = (f_g^1, f_g^2, \dots, f_g^n)$ represents the gene expression values on n spots for gene g. w_{ij} is the spatial weight between spots i and j calculated using the 2D spatial coordinates of the spots and $W = \sum_i \sum_j w_{ij}$. For each spot, we find the top K nearest neighbors according to Euclidean

distances where $K = 6$ and $w_{ij} = 1$ if spot j is one of the nearest neighbors of spot i while $w_{ij} = 0$ otherwise. A *Moran's Index* close to 1 indicates a clear spatial pattern, a value close to 0 indicates random spatial expressions, and a value close to –1 indicates a negative correlation between two adjacent spots. We applied the *moran.test* function from the *spdep* R package to generate the score.

(*viii*) *Geary's C* is a metric measuring spatial autocorrelation, defined as

$$C = \frac{n}{2W} \cdot \frac{\sum_i \sum_j [w_{ij}(f_g^i - f_g^j)]^2}{\sum_i (f_g^i - \bar{f}_g)^2}, C \in [0, 2]$$

where a small C indicates strong spatial autocorrelation, and all notations used here are the same as the notations when defining Moran's Index. Generally, to convert it to range -1 to 1, the following formula is adopted

$$C^* = 1 - C$$

Here, the meaning of the value C^* is similar to Moran's Index mentioned above. We used *geary.test* function from R package *spdep* to generate the score C, and then obtained C^* using a customized script.

Analysis on HE-coronal sample

SVG prediction. The spot number of mouse brain (HE coronal sample) is 2,702, so the first 50 low-frequency signals were used for UMAP dimension reduction and visualization. To demonstrate the advantage of low-frequency signals in terms of SVG representation, PCA was also used for producing low-dimension representation. The transposed and normalized expression matrix was decomposed via using the *sc.tl.pca* function from the *scanpy* package (version 1.9.1). The top 50 principal components (PC) were used for UMAP dimension reduction and visualization. The function *sc.tl.umap* was applied to further conduct dimension reduction for the top 104 low-frequency signals and the top 50 PCs in two-dimensional latent space, respectively.

TM and TM-associated SVG identification. We applied SpaGFT on the HE-coronal mouse data (**Figs. 2b-2e**) to identify TMs and sub-TMs using default parameters. The *Louvain* clustering algorithm (*neighbors* = 15 and *resolution* = 1) was applied on the top 104 low-frequency signals of SVGs ($104 = 2\sqrt{n}$, $n = 2702$ spots), followed by the pseudo-expression calculation. To demonstrate the biological functions of identified TMs, pathway enrichment analysis was conducted using the Enrichr package[53,54] based on the hypergeometric test for SVGs within individual TMs. Three databases were selected, (*i*) ChEA (2016 version) for transcription factor enrichment analysis, (*ii*) BioPlanet (2019 version) for functional pathway enrichment analysis, and (*iii*) GO Biological Process (2021 version). To further investigate sub-TMs, the SVGs in one TM were re-clustered via Louvain clustering with *resolution*=0.5, leading to the calculation of the pseudo-expression and binary-expression for sub-TMs.

Low-expression gene signal enhancement. Specifically, in HE-coronal mouse brain data analysis, we selected the 260 ($= 5\sqrt{n}, n = 2702$), 780 ($= 15\sqrt{n}, n = 2702$), and 1,300 ($= 25\sqrt{n}, n = 2702$) low-frequency signals in the frequency domain and performed the inverse graph Fourier transform

with $c = 0.0001$ to smooth spatial patterns.

Cell2location deconvolution for generating TM-cell type matrix. To generate the TM-cell type matrix, defined in **Fig. 3e**, we first followed the online tutorial of cell2location (https://cell2location.readthedocs.io/en/latest/notebooks/cell2location_tutorial.html) and calculated the cell proportion of each of the 59 cell types for the HE-coronal data across all spots (**Supplementary Table 9**). Then, pseudo-expression values across all spots for one sub-TM were computed using the method from the **TM identification and characterization section**. Then, an element of the TM-cell type matrix was calculated by computing the Pearson correlation coefficient between the proportion of a cell type and the pseudo-expression of a sub-TM across all the spots. Lastly, the TM-cell type matrix was obtained by calculating all elements as described above, with rows representing TMs and coloumns representing cell types.

Analysis of the lymph node sample

TM identification and interpretation. SVGs were identified on the human lymph node data (Visium) with default setting of SpaGFT, and TMs and TM-associated SVGs were determined as described above. Binary TMs were determined using 0.85 percentile as cutoff. To demonstrate the relations between cell composition and TMs, cell2location[25] was implemented to deconvolute spot and resolve fine-grained cell types in spatial transcriptomic data. Cell2location was used to generate the spot-cell type proportion matrix as described above, resulting in cell proportion of 34 cell types (**Supplementary Table 11**). A TM-cell type matrix was calculated using 34 lymph node cell types via the same method as previously described (the **Method** section of *Cell2location deconvolution for generating TM-cell type matrix*). Then, the TM-cell type matrix was generated and visualized on a heatmap, and three major TMs in the lymph node were annotated, i.e., the T cell zone, GC, and B follicle.

Visualization of GC, T cell zone, and B follicles in the Barycentric coordinate system. Spot-cell proportion matrix was used to select and merge signature cell types of GC, T cell zone, and B follicles for generating a merged spot-cell type proportion matrix (an N-by-3 matrix, N is equal to the number of spots). For GC, B_Cycling, B_GC_DZ, B_GC_LZ, B_GC_prePB, FDC, and T_CD4_TfH_GC were selected as signature cell types. For T cell zone, T_CD4, T_CD4_TfH, T_TfR, T_Treg, T_CD4_naive, and T_CD8_naive were selected as signature cell types. For B follicle, B_mem, B_naive, and B_preGC were regarded as signature cell types. The merged spot-cell type proportion matrix was calculated by summing up the proportion of signature cell types for GC, T cell zone, and B follicle, respectively. Finally, GC, T-cell zone, and B follicle assigned spots (spot assignment in **Supplementary Table 12**) were selected from the merged spot-cell type proportion matrix for visualization. The subset spots from the merged matrix were projected on an equilateral triangle via Barycentric coordinate project methods[9]. The projected spots were colored by TM assignment results.

Identification of TM clusters among seven samples. The SVGs and TMs of HE-coronal had been identified from previous Methods section (Analysis on HE-coronal sample), and the SVGs and TMs of othe other six samples (SA1, SA2, SP1, SP2, GSM5519054, and IF-FFPE) were identified using SpaGFT with the default parameters. Then, SVGs identified from the seven

samples were concatenated into an SVG-TM matrix (with 3,690 SVGs and 67 TMs), where values in the matrix were marked as 1 (existence) and 0 (not existence). The SVG-TM matrix was fit into PCA for dimension reduction and Louvain algorithm for TM clustering, resulting in 14 TM clusters. Among those 14 TM clusters, three TM clusters contains conserved TMs from at least six samples. To investigate the cell type composition of three TM clusters, 59 mouse brain cell types were used for generating seven TM-cell type matrices as previously described (the **Method** section of *Cell2location deconvolution for generating TM-cell type matrix*). the cell type-TM matrix was binarized using a cutoff of 0.1, with the correlation larger than 0.1 as a colored element (**Fig. 5c**),. To obtain the overlapping SVGs across identified TMs, Fisher's exact test was performed, and the *p*-values were adjusted using the Benjamini-Hochberg method. The overlapped SVGs between any two modules were calculated, and the odds ratio and adjusted *p*-values were shown on the heatmap (**Fig. 5d**).

Computational environment and running time

All experiments were performed on our lab server set up at the Ohio Supercomputing Center. The server has a 2.6GHz AMD EPYC 7H12 processor, 64 cores, and 1 TB RAM. We tested the computing time of SpaGFT and other tools on three datasets, (*i*) HE-coronal mouse brain datasets with 2,702 spots. (*ii*) the 151673 Visium human brain datasets with 3,639 spots, and (*iii*) the Puck-200115-08 slide-seq v2 datasets with 53,208 spots. For the first dataset, SpaGFT, SPARK, SPARK-X, MERINGUE, SpatialDE, and SpaGCN used 25 seconds, 6 hours, 52 seconds, 3 hours, 1.5 hours, and 17 minutes. For the second dataset, SpaGFT, SPARK, SPARK-X, MERINGUE, SpatialDE, and SpaGCN spent 21 seconds, 6 hours, 50 seconds, 3 hours, 72 minutes, and 25 minutes. For the third dataset, only SpaGFT (15 minutes) and SPARK-X (56 seconds) successfully completed the SVG identification, while the rest of the tools spent over 48 hours or failed.

Data Availability

The 11 datasets from 10x Visium (ten mouse brain datasets and one human lymph node sample)[29] can be accessed from https://www.10xgenomics.com/products/spatial-gene-expression. GSM5519054_Visium_MouseBrain dataset is available from the GEO database with an accession number GSM5519054[42]. Regarding the human brain dataset[15], twelve samples can be accessed via endpoint “jhpce#HumanPilot10x” on Globus data transfer platform at http://research.libd.org/globus/. The other six human brain datasets (2-3-AD_Visium_HumanBrain, 2-8-AD_Visium_HumanBrain, T4857-AD_Visium_HumanBrain, 2-5_Visium_HumanBrain, 18-64_Visium_HumanBrain, and 1-1_Visium_HumanBrain) are available in a BioRxiv study[55]. The two Slide-seq V2 datasets[2] are available as accession number SCP815 in the Single Cell Portal via the link https://singlecell.broadinstitute.org/single_cell.

Code Availability

SpaGFT is a python package for modeling and analyzing spatial transcriptomics data. The SpaGFT source code and the analysis scripts for generating results and figures in this paper are available at https://github.com/OSU-BMBL/SpaGFT.

References

50 Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. in 2011 31st international conference on distributed computing systems workshops. 166-171 (IEEE).
51 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 19, 15, doi:10.1186/s13059-017-1382-0 (2018).
52 Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. Nature Methods 19, 171-178, doi:10.1038/s41592-021-01358-2 (2022).
53 Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128, doi:10.1186/1471-2105-14-128 (2013).
54 Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90-97, doi:10.1093/nar/gkw377 (2016).
55 Chen, S. et al. Spatially resolved transcriptomics reveals unique gene signatures associated with human temporal cortical architecture and Alzheimer's pathology. bioRxiv, 2021.2007.2007.451554, doi:10.1101/2021.07.07.451554 (2021).

Acknowledgments
This work was supported by awards R01-GM131399, R21HG012482, and U54AG075931 from the National Institutes of Health. The work was also supported by the award NSF1945971 from the National Science Foundation. In addition, we thank M. McNutt, Q. Guo, S. Sun, X. Xiong for the help of data collection and gene validation on ISH. This work was also supported by the Pelotonia Institute of Immuno-Oncology (PIIO). We thank the anatomical carton generated from the BioRender website.

Author contributions
Conceptualization: Q.M.; methodology: J.L., Y.C., B.L., Q.M.; software coding: J.L. and Y.C.; data collection and investigation: Y.C.; data analysis and visualization: Y.C., A.M., and J.L.; case study design and interpretation: Z.L., Y.C., and A.M.; software testing and tutorial: J.L. and Y.C.; graphic demonstration: Y.C. and A. M.; manuscript writing, review, and editing: all the authors.

Supplementary Tables
Supplementary Table 1 | Data information. The table includes information on 32 spatial transcriptome datasets from the public domain. The first column shows the data ID in the original paper or data source; the second column shows the use of the data (i.e., for grid-search optimization, independent test, or case study); the third column shows the sequencing platform; the fourth to the sixth columns show the sample information, including species, conditions, and tissue sources; the rest of the columns shows the statistical information of each data, including the number of spots, the number of genes, the number of total reads, the mean read per spot, the standard deviation of the number of reads per spot, the mean number of genes per spot, and the standard deviation of genes per spots.

Supplementary Table 2 | 849 SVG candidates collected from the public domain. The table collects 849 unique cell-type- or layer-specific markers from five different kinds of literature. The first column records the mouse gene symbol. The second column records the paper source. The third column records the experiment object in each gene, where "M," "H," and "M&H" represent

mouse, human, and both. The fourth column records the human gene symbol. The fifth column records the original source in the paper for each gene, either figures or supplementary files.

Supplementary Table 3 | 458 curated benchmarking SVGs validated by the Allen Brain Atlas. The first six columns correspond to general information on gene identifiers, including gene symbol (mouse), gene symbol (human), UniqueID, probe name, plane, and the experiment ID in the ISH database. The ISH intensity on 12 brain regions was recorded from column G to Column R, respectively, including Isocortex, Olfactory area (OLF), Hippocampal formation (HPF), Cortical subplate (CTXsp), Striatum (STR), Pallidum (PAL), Thalamus (TH), Hypothalamus (HY), Midbrain (MB), Pons (P), Medulla (MY), and Cerebellum (CB). All the records were downloaded from the ISH database. Column S records the mean ISH intensity of 12 mouse brain regions. Column T records variance calculated based on the *FindVariableFeatures* function in the Seurat package. Column U records whether the gene is considered as a curated benchmarking SVG in this paper.

Supplementary Table 4 | Grid-search of parameter combination for SVG prediction. The table records the details of the performance comparison in terms of the grid-search of parameter optimization. The first four columns correspond to sample ID, tested software, sequence technology, and parameter combinations. The rest of the columns records eight evaluation matrices, including the Jaccard index, Tversky index, the odds ratio of Fisher's exact test, precision, recall, F1 score, Moran's I, and Geary's C. If an element in this table is "NA," the software shows an error or ran out of time (running time was greater than 48 hours) during testing.

Supplementary Table 5 | Running time of SpaGFT and other tools on the three grid-search test data. The table records the running time and memory cost of SpaGFT, SPARK, SPARK-X, MERINGUE, SpatialDE, and SpaGCN on the HE-coronal, 151673, and Puck-200115-08 datasets. All tools and experiments were carried out in the same computing environment introduced in Methods. Columns A and B show tool names and sample names; Column C and D records the running time with the unit as second (S) and $\log_{10}$(S), respectively. Column E is memory cost with the unit as a megabyte. For any experiments spent over 24 hours, we labeled them as "NA".

Supplementary Table 6 | SVG prediction performance on 28 independent test datasets using default parameters. The table records the details of the performance comparison in terms of the independent test. The first column indicates the dataset ID, corresponding to the Dataset ID in Supplementary Table 1. The second column shows eight evaluation matrices, including the Jaccard index, Tversky index, the odds ratio of Fisher's exact test, precision, recall, F1 score, Moran's I, and Geary's C. The other columns are the software. If an element in this table is "NA," the software shows an error or runs out of time (running time was greater than 48 hours) during testing.

Supplementary Table 7 | Summary of top 100 genes identified by SpaGFT, and the fix benchmarking tools. The table records the unique and consistent SVGs of the top 100 SVGs identified by six tools for mouse brain data (HE-coronal). The first column is the gene name. Columns B, C, D, E, F, and G are software names. The values in Columns B to G indicate whether the gene is identified by this tool. If the value is equal to 1, it means the gene is the output of the

top 100 SVGs in this software, and vice versa. Column H is the sum of values from Columns B to G, indicating the consistency of identified genes (the higher value, the higher consistency). When the value in Column H is "1," it means that this gene is uniquely identified by this one of the tools from Columns B to G.

Supplementary Table 8 | SVG results in the HE coronal data. The table records all SVGs predicted from SpaGFT on the HE-coronal data. Column A is the SVG name; Column B is the number of spots having this SVG expressed; Column C is the corresponding *GFTscore*; Column D is the ranking of GFTscore. Columns E and F are the *p*-value and *q*-value of SVG, respectively; Columns H and I are the TM labels and sub-TM labels, respectively. SVGs are arranged based on the SVG_rank from high to low.

Supplementary Table 9 | Deconvolution results for HE-coronal sample. The table shows the proportions of 59 cell types calculated by cell2location. The first column is the spot ID of the mouse sample. The rest of the columns are the cell proportions in 59 cell types, respectively.

Supplementary Table 10 | SVG results in the lymph node data. The table records all SVGs predicted from SpaGFT on the lymph node data. Column A is the SVG name; Column B is the number of spots having this SVG expressed; Column C is the corresponding *GFTscore*; Column D is the ranking of GFTscore. Columns E and F are the *p*-value and *q*-value of SVG, respectively; Columns H and I are the TM labels and sub-TM labels, respectively. SVGs are arranged in the decreasing order of the SVG_rank score.

Supplementary Table 11 | Cell2location cell deconvolution results for Human lymph node. The table shows the proportions of 34 cell types calculated by cell2location. The first column is the spot ID of the human lymph node sample. The rest of the columns are the cell proportions in 34 cell types, respectively.

Supplementary Table 12 | TM assignment to each spot from lymph node data in terms of GC, T cell zone, and B follicle. The table demonstrates GC, T cell zone, B follicle, and their interactive region assignment label. The first column is the spot ID. The second column is the assignment label, where “0” is no assignment; “T.zone” is the spot assigned as T cell zone; “B.follicle” is the spot assigned as B follicle; “GC” is the spot assigned as germinal center; “T.zone-B.follicle” is the spot assigned as the interactive region between T cell zone and B follicle; “GC-T.zone” is the spot assigned as the interactive region between GC and T cell zone; “GC-B.follicle” is the spot assigned as the interactive region between GC and B follicle; “GC-T.zone-B.follicle” is the spot assigned as interactive region among GC, T zone, and B follicle.

Supplementary Table 13 | SVG results in the seven mouse brain data. The table records all SVGs predicted from SpaGFT in the seven mouse brain data. Column A is the SVG name; Column B is the number of spots having this SVG expressed; Column C is the corresponding *GFTscore*; Column D is the ranking of *GFTscore*. Columns E and F are the *p*-value and *q*-value of SVG, respectively; Columns H is the TM labels; Column I indicates the sample names.

Supplementary Table 14 | TM-associated SVG and TM assignment to each spot from seven samples in terms of TM cluster 1, TM cluster 2, and TM cluster 3. The table shows 3690 SVG, TMs, and two labels for TMs. The first row is the clustering results of the Louvain algorithm. The second row is the TM clusters label assignment, including TM cluster 1, TM cluster 2, and TM cluster 3. The rest rows are SVGs (3690 SVGs). Columns indicate samples and their TMs.

Supplementary Table 15 | The overlapped SVGs across seven mouse brain samples in terms of TM cluster 1, TM cluster 2, and TM cluster 3. The table records overlapped SVGs among TMs in tissue motif 1 and tissue motif 2. Column A indicates TM cluster label and overlapped names. Column B indicates tissue motif ID. The other columns indicate sample ID. The value from column C to column I represents the number of spots. If an element in this table is "NA," the no overlapping spot between two TM clusters.

Supplementary Figures

Supplementary Fig. 1 | FM identification and visualization. a, Workflow of FM identification. Spot graph is constructed by KNN, where K is equal to the number of spot n. The degree and adjacency matrix are generated, then the Laplacian matrix can be calculated by subtracting the degree matrix and adjacency matrix. Through decomposing the Laplacian matrix, eigenvalue and eigenvector are obtained, where eigenvectors are the FMs. **b**, Visualizations of FM patterns in the different frequency domains of the Visium 151673 dataset, where LFM means low-frequency FM and HFM means high-frequency FM. **c**, Impact of the number of FMs in identifying SVGs. Different numbers of FMs were selected, i.e., $0.5\sqrt{n}$, $\sqrt{n}$, $2\sqrt{n}$, $3\sqrt{n}$, $4\sqrt{n}$, $5\sqrt{n}$, $8\sqrt{n}$, and $10\sqrt{n}$. Under each selection, the top 1,000 genes with high *GFTscore* are kept for pair-wise comparison, where the number in the heatmap block indicates the number of overlapped genes. The results showed high consistency of SVG results even if we selected different numbers of FMs, which demonstrates the robustness of SpaGFT.

Supplementary Fig. 2 | Comparison of evaluation matrices (Morans' I and Gearys'C). a, Moran's I and Geary's C score on the grid-search testing for the HE-coronal sample. The boxplot indicates the Moran's I and Geary's C score distribution for six tools' grid-search results, respectively. The Black line in the box indicates the median value. **b**, Moran's I and Geary's C score on 28 independent datasets using optimized parameters of SpaGFT and default parameters in the five benchmarking tools. The black line in the box indicates the median value.

Supplementary Fig. 3 | ISH evidence of four SVGs uniquely identified by SpaGFT. The ISH database webpage shows four major information, including experiment information (top left), ISH high-resolution image (right), 3D expression (middle left), and ISH intensity of 12 mouse brain regions (bottom). In addition, we used a dashed line to circle out ISH high-intensity regions on ISH high-resolution image. **a**, The screenshot of gene Calb2, which is in the coronal plane, shows a highly consistent expression pattern of HE-coronal spatial data. **b**, The screenshot of gene Hcrt. Due to the lack of coronal plane data, we use sagittal instead of the coronal plane. Interestingly, the ISH intensity is not high in the 12 regions on the barplot (bottom), but we can clearly observe enriched intensity in the hypothalamus region. **c-h**, The screenshot of gene Gda, Zfhx3, Gal,

Cacnb3, Asb4, and Mpped1, which is also in the coronal plane, shows a highly consistent expression pattern of HE-coronal spatial data, respectively.

Supplementary Fig. 4 | The ID card of TM 2 for the HE-coronal data. TM 2 includes 256 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 5 | The ID card of TM 3 for the HE-coronal data. TM 3 includes 251 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 6 | The ID card of TM 4 for the HE-coronal data. TM 4 includes 227 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 7 | The ID card of TM 5 for the HE-coronal data. TM 5 includes 192 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 8 | The ID card of TM 6 for the HE-coronal data. TM 6 includes 159 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 9 | The ID card of TM 7 for the HE-coronal data. TM 7 includes 96 SVGs, and all components on the ID card are the same as TM 1 in Fig. 1c, including a spatial map, a TM map, the frequency signal histogram, the spatial map of the top four SVGs, and the top five functional enrichment results in three databases using Enrichr.

Supplementary Fig. 10 | Brain region atlas. The figure shows the mouse brain's six structures obtained from Allen Brain Atlas, including Field CA1 (**a**), Hippocampal region (**b**), Hypothalamus (**c**), Cortical subplate (**d**), Thalamus (**e**), and Fiber tracts (**f**). The purple color highlights the corresponding brain regions.

Supplementary Fig. 11 | The sub-TMs of TM 1–7 in the HE-coronal data. The figure shows the sub-TMs (in TMs 1, 2, 4, 5, 6, and 7) by reclustering SVGs in each TM (from left to right and top to bottom), similar to Fig. 3e. Each sub-TM has a group of unique SVGs, showing different spatial expression patterns among each other.

Supplementary Fig. 12 | Cell type distribution of other TMs. The figures show TMs 1-7 and TM6 Sub-TM 4 cell type component and distribution generated from cell2location. The left box represents TM pseudo-expression and its binary form. The right box represents cell-type compositions.

Supplementary Fig. 13 | Pathway and other gene signatures enriched within GC, T cell zone, and B follicle region. The figure shows pseudo-expression TM, binary TM, TM enriched functional pathway (left), TM associated SVG (upright), and TM correlated cell types (downright) for GC, T cell zone, and B follicle.

Supplementary Fig. 14 | The intact heatmap of TM intersections across three TM clusters. The heatmap shows gene overlapping of 22 TMs derived from three TM clusters. The color indicates the log-odds ratio of the Fisher exact test. *p*-value (Benjamini-Hochberg adjusted) between two samples is showcased on the heatmap. Three anatomical structures (cerebrum, hypothalamus, and white matter) were derived from Allen Brain Altas, and targeted regions are indicated by the purple color.