

HyperSTAR: Unveiling Tissue Structure and Tumor Microenvironment from Spatial Omics by Hypergraph Learning

Yi Liao (廖异)^{1, }, Chong Zhang (张宠)^{1, }, Zhikang Wang (王智康)^{2,3, }, Fei Qi (齐飞)^{4, },
Weitian Huang (黄伟填)^{5, }, Shangyan Cai (蔡尚妍)^{6, }, Junyu Li (李俊宇)^{5, }, Jiazhou Chen (陈佳洲)^{7, },
Robin B. Gasser^{8, }, Zhiyuan Yuan (原致远)^{9,10,*, }, Jiangning Song (宋江宁)^{2,3,*, },
Hongmin Cai (蔡宏民)^{1,5,*, }

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

²Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia

³Monash AI Institute, Monash University, Melbourne, VIC 3800, Australia

⁴School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China

⁵School of Future Technology, South China University of Technology, Guangzhou 511442, China

⁶Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519087, China

⁷School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

⁸Department of Veterinary Biosciences, Melbourne Veterinary School, Faculty of Science, The University of Melbourne, Parkville, VIC 3010, Australia

⁹Institute of Science and Technology for Brain-Inspired Intelligence, MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China

¹⁰Center for Medical Research and Innovation, Shanghai Pudong Hospital, Fudan University Pudong Medical Center, Fudan University, Shanghai 201399, China

*Corresponding authors: hmcai@scut.edu.cn (Cai H), Jiangning.Song@monash.edu (Song J), zhiyuan@fudan.edu.cn (Yuan Z).

Handling Editor: Shihua Zhang

Abstract

Spatial omics technologies have revolutionized life sciences by enabling the simultaneous acquisition of biomolecular and spatial information. Identifying spatial patterns is crucial for understanding organ development and tumor microenvironments. However, the emergence of diverse spatial omics resolutions in these technologies has made it challenging to accurately characterize spatial domains at finer resolutions. To address this, we propose HyperSTAR, a hypergraph-based method designed to precisely identify spatial domains across varying resolutions by leveraging higher-order relationships among spatially adjacent tissue programs. Specifically, a gene expression-guided hyperedge decomposition module is introduced to refine the hypergraph structure to accurately delineate spatial domain boundaries. Additionally, a hypergraph attention convolutional neural network is designed to adaptively learn the importance of each hyperedge, enhancing the model's ability to capture complex higher-order relationships within spatially neighboring multi-spots and/or single cells. HyperSTAR outperforms existing graph neural network models in tasks such as uncovering tissue substructures, inferring spatiotemporal patterns, and denoising spatially resolved gene expression. It effectively handles diverse spatial omics data types and scales seamlessly to large datasets. The method successfully reveals spatial heterogeneity in breast cancer sections, with findings validated through functional and survival analyses of independent clinical data. HyperSTAR represents a significant advancement in spatial omics analysis, representing a robust tool for exploring complex spatial patterns across varying resolutions and data types. Its ability to capture intricate higher-order relationships among spatially neighboring spots/cells makes it an invaluable tool for advancing research in life sciences, particularly in cancer and developmental biology. The toolbox is available at <https://github.com/Ringoio/HyperSTAR>.

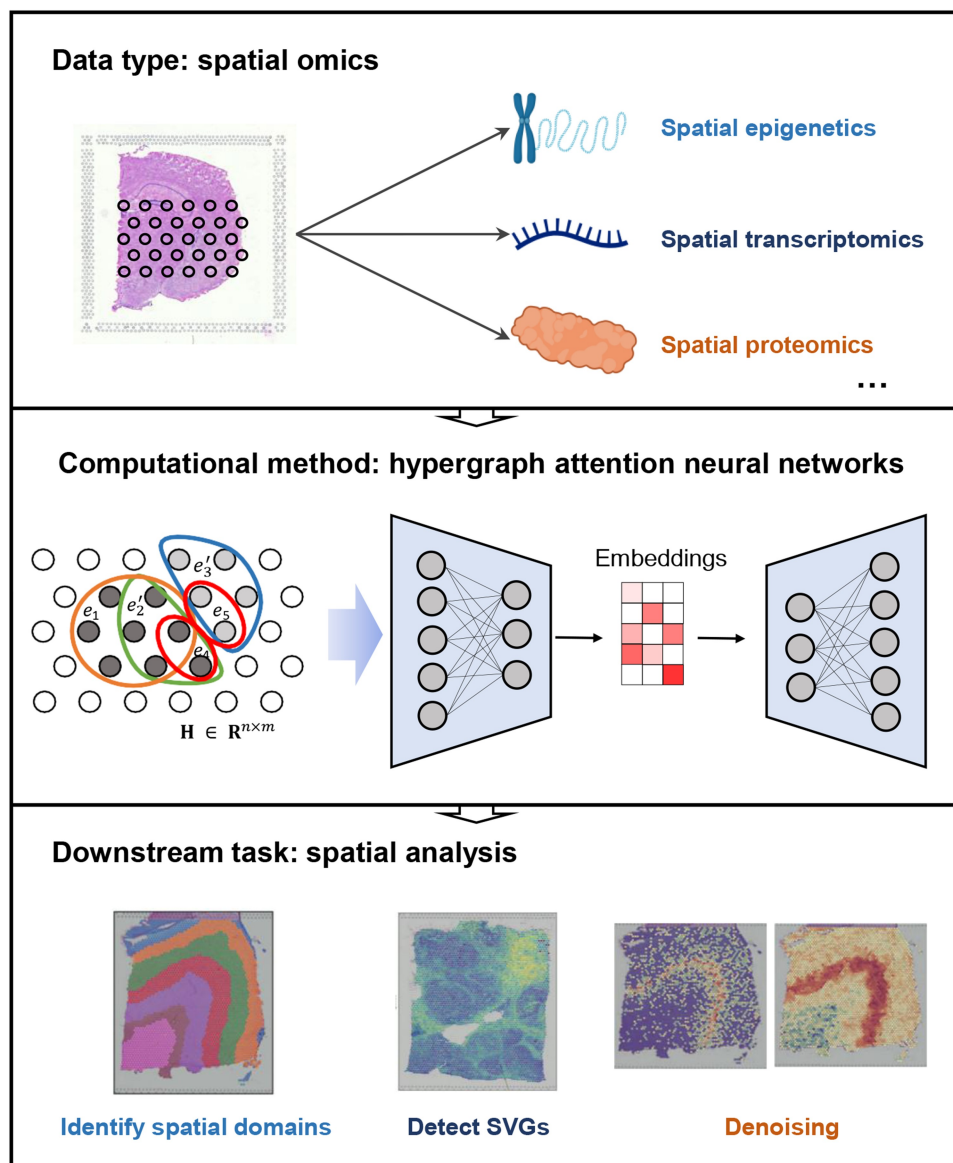
Key words: Spatial omics; Spatial domain; Resolution; Hypergraph learning; Higher-order relationship.

Received: 27 August 2024. Revised: 8 July 2025. Accepted: 7 December 2025.

© The Author(s) 2025. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Graphical abstract



Introduction

In living organisms, biological processes occur within a spatial context. The location of each cell is as crucial as its inherent characteristics in determining cell types, cell states, and functions [1]. Single-cell omics quantifies genomes, epigenomes, transcriptomes, and proteomes at the single-cell level [2,3]. Nevertheless, the process of single-cell dissociation during sample preparation results in the loss of spatial information [4]. Emerging spatially resolved transcriptomics (SRT) has facilitated the profiling of gene expression while preserving spatial location in tissues, offering valuable opportunities to explore intricate mechanisms within biological systems [5].

Widely used SRT can be broadly categorized into two groups: (1) imaging-based approaches, and (2) next-generation sequencing (NGS)-based approaches [6–10]. These diverse technologies simultaneously capture both gene expression and spatial location,

while differing in gene throughput, sensitivity, resolution, and field of view. Imaging-based methods provide subcellular resolution with high sensitivity but limited gene throughput [11–17], while NGS-based technologies generate diverse spatial resolutions, ranging from multicellular to subcellular levels, for the captured spots with whole-transcriptome coverage but lower sensitivity [18–20].

Several computational methods have been developed to identify spatial domains in SRT data. Previous studies utilized traditional methods (e.g., K-means [21,22] and Louvain [23]), which did not take into account the spatial information of SRT data, leading to a discontinuous division of regions. Recent methods have introduced spatial information to improve spatial domain identification. For example, BayesSpace [24] employed a Bayesian statistical approach, using the information from spatial neighborhoods for spatial clustering. Deep learning methods

have also been developed to integrate spatial information and gene expression of SRT data. StLearn [25] defined morphological distances based on features extracted from histology images and utilized these distances, along with spatial neighbors, to smooth gene expression. SpaGCN [26] also used histology images by employing graph convolutional networks to integrate gene expression, spatial location, and image data. SpaceFlow [27] utilized spatially regularized deep graph networks to integrate both expression similarity and spatial information. STAGATE [28] applied a graph attention auto-encoder to learn low-dimensional latent embeddings of SRT profiles. Although these methods have shown significant improvements on some datasets, they primarily focus on the cell-cell pairwise relationships, neglecting the higher-order structures inherent in multicellular programs. Furthermore, they fail to account for the delineation of domain boundaries, a crucial aspect for improving spatial domain identification, as they predominantly focus on clustering. This limitation becomes more pronounced when confronted with higher spatial resolutions, making it challenging to elucidate the tissue structure comprehensively. Moreover, these approaches encounter difficulties in extending their capabilities to accommodate other omics.

To tackle these issues, we propose HyperSTAR, the first hypergraph learning-based spatial clustering method that explicitly learns the higher-order tissue structures. HyperSTAR can effectively leverage complex relationships among spatially adjacent multicellular programs. A gene expression-guided hyperedge decomposition module and a hypergraph attention convolutional layer are employed to precisely delineate boundaries of spatial domains. Extensive experiments show that HyperSTAR can reveal tissue structure and tumor spatial heterogeneity across varying spatial resolutions, demonstrating effectiveness in diverse downstream analyses, such as spatiotemporal trajectory inference, gene expression denoising, and spatially variable gene (SVG) detection.

Method

Data description

We applied HyperSTAR to a diverse range of spatial omics datasets generated by various technologies to demonstrate its effectiveness. These datasets span multiple platforms, including SRT technologies such as 10x Visium, Stereo-seq, ourboros single-molecule fluorescence *in situ* hybridization (osmFISH), multiplexed error-robust fluorescence *in situ* hybridization (MERFISH), and 10x Xenium, as well as CODEX for spatial proteomics. These platforms produce data with varying resolutions, ranging from spot-level to single-cell resolution. For example, 10x Visium datasets have a resolution of 55 μm (1–10 cells per spot), while Stereo-seq provides spot resolutions with bin sizes of 20, 50, and 100 μm , as well as single-cell resolution.

Specifically, we used the human dorsolateral prefrontal cortex (DLPFC) dataset [29], generated by 10x Visium, as the benchmark for spatial domain identification. This dataset includes 12 human DLPFC sections from 3 individuals, with manual annotations provided by the original authors. The 10x Visium mouse brain coronal dataset comprises 2702 spots and 21,949 genes as raw features. The Stereo-seq mouse brain olfactory bulb

dataset, with single-cell resolution, includes 19,109 spots and 27,106 genes as raw features. The breast cancer section dataset, generated by 10x Visium, has a median of 20,762 unique molecular identifiers (UMIs) per spot and a median of 6026 genes per spot. These datasets cover diverse species, organs, and resolutions, showing the versatility of HyperSTAR.

Data preprocessing

To effectively analyze the raw data, we employed a series of preprocessing steps. First, we removed spots located outside the primary tissue region to focus on biologically relevant areas. Next, we performed feature selection to identify the top highly variable genes (HVGs) (typically 3000 genes). After selecting the HVGs, we applied a log transformation to the gene expression values. The resulting normalized expression values of HVGs were then used as input node representations in the hypergraph learning module. This preprocessing ensures that the hypergraph learning module can effectively capture the intricate relationships among spots and identify biologically meaningful spatial domains.

Hypergraph construction

After preprocessing to obtain spot representations, we constructed the hypergraph structure in two steps:

Step 1. Initial hyperedge construction: For each spot, we generated a hyperedge using the k -nearest neighbors (kNN) approach based on spatial information of the spots. This step captures local spatial relationships among spots, ensuring that spatially adjacent spots are connected in the hypergraph.

Step 2. Hyperedge refinement: To address the scenarios where spatially adjacent spots belong to different biological domains, we refined the hyperedge set using pre-clustering results based on gene expression. This refinement aligns the hypergraph structure with underlying biological domains, thereby improving the accuracy of spatial domain identification. Specifically, if the spots within a hyperedge belong to the same cluster, the hyperedge remains unchanged, indicating that spatially adjacent spots share similar gene expression patterns and likely belong to the same biological domain. If the spots within a hyperedge belong to different clusters, we consider two scenarios: (1) Single outlier spot: If only one spot belongs to a different cluster, it is removed from the hyperedge to maintain consistency. This ensures that the hyperedge only contains spots that are both spatially adjacent and share similar gene expression patterns; (2) Multiple outlier spots: If multiple spots belong to different clusters, the hyperedge is decomposed into two new hyperedges, each containing spots from the same cluster. This effectively separates the spots that are spatially adjacent but have distinct gene expression patterns. By refining the hyperedge set based on the pre-clustering results, we better capture the spatial and biological relationships among spots, leading to a more accurate representation of spatial domains in the hypergraph structure.

Hypergraph learning

Once the hypergraph is constructed, we input the spot features and hypergraph structure into the hypergraph learning module, which consists of two parts: the encoder and the decoder. The

encoder generates low-dimensional latent embeddings of spots, used for spatial domain identification and downstream analyses. The decoder produces reconstructed features of spots, utilized for denoising.

Given a hypergraph $\mathbf{G} = (\mathbf{X}, \mathbf{H})$, where $\mathbf{X} \in \mathbf{R}^{n \times d}$ represents the feature matrix of n spots and $\mathbf{H} \in \mathbf{R}^{n \times m}$ represents the incidence matrix from vertexes to hyperedges defined as [30,31]:

$$h_{ij} = \begin{cases} 1, & \text{if } x_i \in N_j^e \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where N_j^e denotes the set of vertexes connected to the j -th hyperedge.

Encoder

The encoder takes the normalized gene expression $\mathbf{X}_0$ as input and generates the spot embeddings by collectively aggregating the information from its neighbors in the same hyperedge. To adaptively learn the weights of the hyperedges, we use three hypergraph attention layers.

Hypergraph convolutional layer

Message propagation on a hypergraph is more complex than that on a graph, involving two stages: attentive hyperedge aggregation and attentive vertex aggregation [32]. In the attentive hyperedge aggregation, the connected vertexes in the same hyperedge are aggregated to generate the hyperedge embeddings. As for the t -th layer, the hyperedge features $\mathbf{E}_{t-1}$ can be calculated as:

$$\mathbf{E}_{t-1} = \mathbf{H}^T \mathbf{X}_{t-1} \quad (2)$$

where $\mathbf{X}_{t-1}$ represents the output vertex features of the previous layer. For each pair of hyperedges (e_i, e_j) where $e_j \in N_i^e$, we calculate the coefficient a_{ij} between hyperedge i and hyperedge j by introducing the attention mechanism [33]:

$$a_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{\theta}_a [\mathbf{W}_a \vec{e}_i \| \mathbf{W}_a \vec{e}_j]))}{\sum_{e_k \in N_i^e} \exp(\text{LeakyReLU}(\vec{\theta}_a [\mathbf{W}_a \vec{e}_i \| \mathbf{W}_a \vec{e}_k]))} \quad (3)$$

where $\mathbf{W} \in \mathbf{R}^{d' \times d}$ is a shared linear transformation, parametrized by a weight matrix. $\vec{\theta}_a \in \mathbf{R}^{2d'}$ represents a shared attentional mechanism, parametrized by a vector. N_i^e denotes the set of hyperedges connected to the i -th vertex. The hyperedge features aggregated by attentive hyperedge aggregation can be calculated as follows:

$$\mathbf{E}_t = \mathbf{A} \mathbf{W} \mathbf{H}^T \mathbf{X}_{t-1} \quad (4)$$

The attentive vertex aggregation module aggregates the connected hyperedge information to generate the vertex embeddings. First, the vertex features can be calculated by

$$\mathbf{X}_t^e = \mathbf{H} \mathbf{E}_t \quad (5)$$

We calculate the coefficient b_{ij} between hyperedge i and vertex j based on the attention mechanism:

$$b_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{\theta}_b [\vec{x}_i^e \| \vec{x}_j^e]))}{\sum_{x_k \in N_i^e} \exp(\text{LeakyReLU}(\vec{\theta}_b [\vec{x}_i^e \| \vec{x}_k^e]))} \quad (6)$$

where $\vec{\theta}_b \in \mathbf{R}^{2d'}$ represents a shared attentional mechanism, parametrized by a vector, while N_i^e denotes the set of vertexes connected to the i -th hyperedge. Next, the output vertex features $\mathbf{X}_e$ can be calculated by

$$\mathbf{X}_e = \sigma(\mathbf{B} \mathbf{H} \mathbf{A} \mathbf{W} \mathbf{H}^T \mathbf{X}_{t-1}) \quad (7)$$

where σ is the activation function.

Decoder

The decoder takes the low-dimensional latent embeddings of spots as input and reconstructs the spot features in the original input dimensions of the encoder. We also use hypergraph attention networks as input to build a symmetrical decoder for the encoder. After applying three layers of hypergraph attention networks, the output of the last layer is considered as the reconstructed spot features. The output of the t -th layer in the decoder can be obtained by:

$$\hat{\mathbf{X}}_t = \sigma(\hat{\mathbf{X}}_{t-1} \mathbf{H} \mathbf{W}_{k-t+1}^T \mathbf{A}_{k-t+1}^T \mathbf{H}^T \mathbf{B}_{k-t+1}^T) \quad (8)$$

where $\hat{\mathbf{X}}_0$ is the latent embeddings learned by the encoder, while k represents the number of layers of the encoder or decoder in our model. We set $k = 3$ in this study.

Loss function

Hypergraph-structured data include node features and the graph structure, both of which should be encoded by high-quality node representations. We minimize the reconstruction loss of node features as follows:

$$\sum_{i=1}^N \|\mathbf{X}_{in} - \mathbf{X}_{out}\|_2 \quad (9)$$

where $\mathbf{X}_{in}$ is the input normalized gene expression $\mathbf{X}_0$ of the encoder, and $\mathbf{X}_{out} = \hat{\mathbf{X}}_k$ represents the final output of the decoder.

Downstream analysis

Clustering

To identify spatial domains based on the embeddings obtained by HyperSTAR, we employed different strategies depending on the availability of prior information about the number of labels or domains. When the number of labels is known, we utilized the mclust clustering algorithm [34]. For data without prior information about the number of domains, we used the Louvain algorithm implemented in the Scanpy package. By employing these two different clustering strategies based on the availability of prior information, we can flexibly adapt the spatial domain identification process to various datasets and research questions.

Spatial trajectory inference

To visualize spatial trajectories, we employed two complementary approaches: the Uniform Manifold Approximation and Projection (UMAP) plot and the partition-based graph abstraction (PAGA) graph [35]. The UMAP plot is used to visualize the embeddings generated by HyperSTAR. The PAGA graph is another powerful tool for visualizing spatial trajectories. PAGA is a graph-based method that combines clustering with trajectory inference, allowing for the identification of both discrete cell states and continuous transitions between them. In the context of spatial omics data, the PAGA graph can reveal the relationships between spatial domains and potential trajectories connecting them. We generated the PAGA graph using the `scanpy.pl.paga` function from the Scanpy package, which takes the HyperSTAR embeddings as input and produces a visually interpretable graph representation of the spatial trajectories. By combining the UMAP plot and the PAGA graph, we can gain complementary and valuable insights into the spatial organization and potential trajectories present in the data, thereby facilitating the interpretation of the underlying biological processes.

SVG detection

We performed the Wilcoxon test implemented in the Scanpy package to identify SVGs for spatial domains with a Benjamini-Hochberg-corrected false discovery rate (FDR) threshold of 1%.

Denoising

The denoised gene expression is generated by the decoder, and the visualization demonstrates that the denoised features are more effective and representative for distinguishing different domains.

Results

Overview of HyperSTAR

SRT data typically contain a gene expression matrix and a spatial coordinate matrix, as illustrated in [Figure 1A](#). To comprehensively capture the intricate associations among spatially adjacent multiple spots, we develop a hypergraph-based method, as illustrated in [Figure 1B and C](#). The construction of this hypergraph can be primarily divided into two key steps: (1) acquiring spot representations and (2) establishing the hypergraph structure. Spot representations are acquired from preprocessed gene expression data. The hypergraph structure is attained through two stages: (1) constructing a naive hyperedge set among spatially neighboring multi-spots, and (2) employing a gene expression-guided hyperedge decomposition module to refine these naive hyperedges. Finally, a hypergraph structure that integrates spatial information and gene expression is obtained ([Figure 1B](#)). The initially built naive hyperedge set contains complex high-order relationships among spatially neighboring multi-spots. However, it is important to note that adjacent spots in space do not always belong to the same domain. For instance, although spots near a domain boundary are spatially adjacent, they might actually belong to different domains. To tackle this issue, a gene expression-guided hyperedge decomposition module is employed to enhance the effects

of gene expression. After constructing the hypergraph, we utilize hypergraph attention convolutional neural networks to learn low-dimensional latent embeddings of the SRT data. Subsequently, these latent embeddings are used to visualize data, identify spatial domains, infer trajectories, detect SVGs, and denoise spatial gene expression ([Figure 1C](#)). Using the aforementioned strategy to integrate spatial information and gene expression in SRT data, HyperSTAR improves the performance of spatial omics analysis.

HyperSTAR achieves state-of-the-art performance on the benchmark dataset

To evaluate the performance of HyperSTAR, we first used the widely recognized SpatialLIBD dataset, a 10x Visium dataset of 12 human DLPFC sections [29]. This dataset includes manually annotated DLPFC layers (Layer 1 to Layer 6) and white matter (WM), which are defined based on morphological features and marker genes. Using these annotations as the ground truth, we employed multiple established metrics — Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Moran's I, and Geary's C — to quantitatively assess the accuracy and robustness of spatial domain identification ([Figure 2A, Figure S1](#)).

We compared HyperSTAR with six commonly used packages for identifying spatial domains, including Scanpy [23] (originally designed for single-cell omics), two methods that integrate histology images (stLearn and SpaGCN), two deep learning methods based on graphs (SpaceFlow and STAGATE), and one Bayesian statistical method (BayesSpace). HyperSTAR effectively identified the expected cortical layer structures, demonstrating significant improvement compared to other methods ([Figure 2A](#)). For example, in sections 151673 and 151507 (ground truth shown in [Figure 2B](#)), HyperSTAR clearly delineated the layer boundaries and achieved the highest clustering accuracy (ARI = 0.57 and ARI = 0.60, respectively) ([Figure 2C and D](#)). The results of spatial visualization for all sections are provided in [Figures S2–S16](#). We further confirmed the inferred trajectory using a trajectory inference algorithm named PAGA. The PAGA graphs illustrate that embeddings belonging to different clusters obtained by Scanpy, stLearn, and SpaceFlow are intermingled. As for STAGATE, it successfully separates WM, Layer 6, and Layer 5; however, Layer 1, Layer 2, Layer 3, and Layer 4 remain mixed. In contrast, our HyperSTAR embeddings effectively separate these cortical layers, providing a clear depiction of tissue topology ([Figure 2E](#)).

We further conducted ablation experiments to illustrate the effectiveness of our strategies. As shown in [Figure 2F](#), we evaluated the performance of four graph-based networks: graph convolution network (GCN) [36], graph attention network (GAT) [33], hypergraph convolution network (HGNN) [30], and our HyperSTAR, which combines hypergraph convolution operation and a graph attention mechanism. The results demonstrate that HyperSTAR achieves higher ARI scores than other methods on almost all sections of the DLPFC dataset ([Figure 2F](#)). As shown in [Figure 2G](#), UMAP plots of embeddings obtained from the four networks demonstrate the superiority of HyperSTAR. For GCN, GAT, and HGNN, there is some overlap between cortical layers, indicating that some spots are mixed. In contrast, HyperSTAR demonstrates exceptional performance by clearly distinguishing all cortical layers without any overlap or mixing. To ensure the robustness of our findings, we included multiple running results

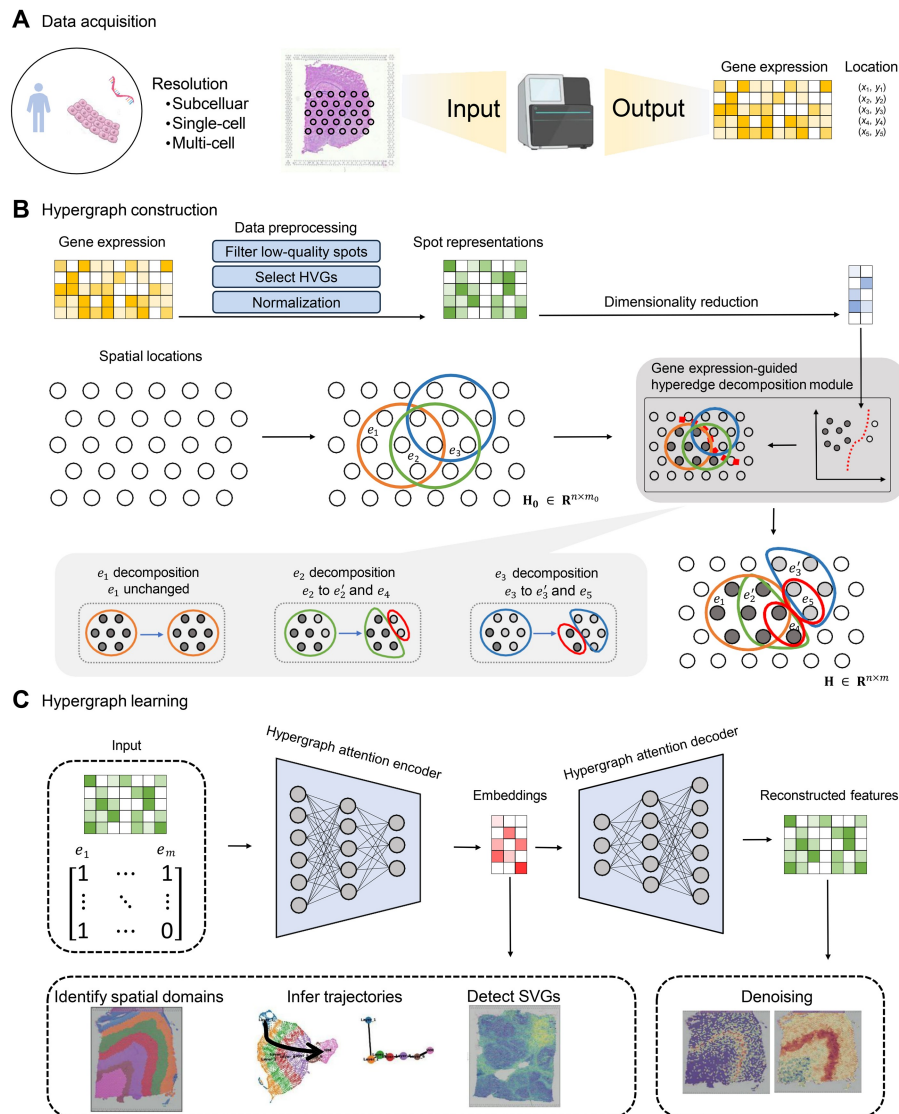


Figure 1 Overview of HyperSTAR

A. SRT data vary in resolution depending on the sequencing technology. Typically, SRT data consist of gene expression profiles and their corresponding spatial coordinates. B. A hypergraph is constructed to model relationships among multi-spots. Spot representations are derived directly from gene expression data. The hypergraph structure is built in two stages: (1) for each spot, a hyperedge is formed by connecting its spatially adjacent multi-spots, and (2) hyperedge decomposition is performed to refine the hyperedges based on gene expression clustering results. C. The refined hyperedge set and spot representations are input into hypergraph attention convolutional neural networks to learn latent embeddings of spots for downstream analyses. SRT, spatially resolved transcriptomics; HVG, highly variable gene; SVG, spatially variable gene.

in Figure S17. Additionally, the denoising results and the detection of SVGs are presented in Figures S18 and S19, respectively.

HyperSTAR reveals a substructure at finer resolutions in the mouse olfactory bulb

To assess HyperSTAR's performance across different resolutions, we selected a mouse olfactory bulb (MOB) section from Stereo-seq, with finer spatial resolution at the single-cell level. Spatial domains on this section were initially identified using three methods: Scanpy, a widely used single-cell method; STAGATE, a graph-based deep learning method; and our proposed method, HyperSTAR. Using mouse anatomy collected from the Allen Brain Atlas as reference (Figure 3A), Scanpy yielded discontinuous results, while STAGATE showed improved

performance, albeit with some mixing between layers, such as the fusion of domain 3 and domain 4 (Figure 3B). In contrast, HyperSTAR identified a clear and continuous layer structure in the MOB.

To further analyze the results obtained by HyperSTAR, we ranked genes for characterizing groups, and the top genes identified for each group are shown in Figure 3E. The gene expression of *Calb2*, *Pcp4*, and *Ptgds*, as shown in Figure 3D, represents the three top genes identified by HyperSTAR for characterizing domains 7, 4, and 2, respectively. The distribution of domains 7, 4, and 2 identified by HyperSTAR is displayed in Figure 3C, corresponding to the outer plexiform layer (OPL), granular layer (GR), and glomerular layer (GL) in the reference shown in Figure 3A. The spatial visualization and corresponding *in situ* hybridization (ISH) images of marker genes

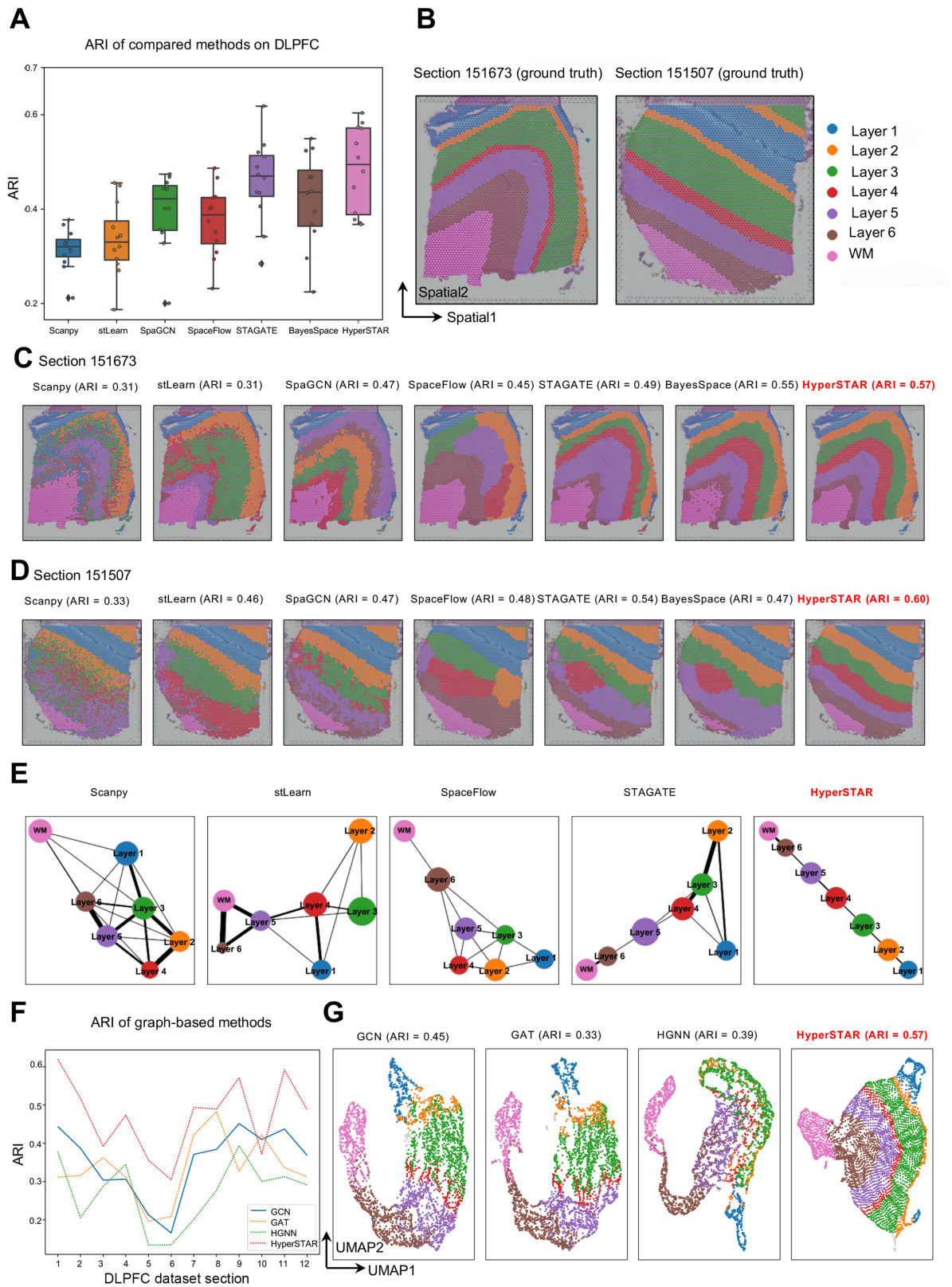


Figure 2 HyperSTAR achieves state-of-the-art performance on the DLPFC benchmark dataset

A. Box plots of clustering accuracy across all 12 sections of the DLPFC dataset, measured by ARI scores for 7 methods. **B.** Ground-truth segmentation of cortical layers (from Layer 1 to Layer 6) and WM in sections 151673 and 151507. **C.** Spatial visualization of clustering results of HyperSTAR in section 151673 compared to six state-of-the-art methods. **D.** Spatial visualization of clustering results of HyperSTAR in section 151507 compared to six state-of-the-art methods. **E.** PAGA graphs generated by Scanpy, stLearn, SpaceFlow, STAGATE, and HyperSTAR. **F.** Line plot of clustering accuracy across all 12 sections of the dataset, measured by ARI scores for four different graph-based networks (GCN, GAT, HGNN, and HyperSTAR). **G.** UMAP plots of section 151673 obtained by the four graph-based networks. DLPFC, dorsolateral prefrontal cortex; WM, white matter; UMAP, Uniform Manifold Approximation and Projection; ARI, Adjusted Rand Index; PAGA, partition-based graph abstraction; GCN, graph convolution network; GAT, graph attention network; HGNN, hypergraph convolution network.

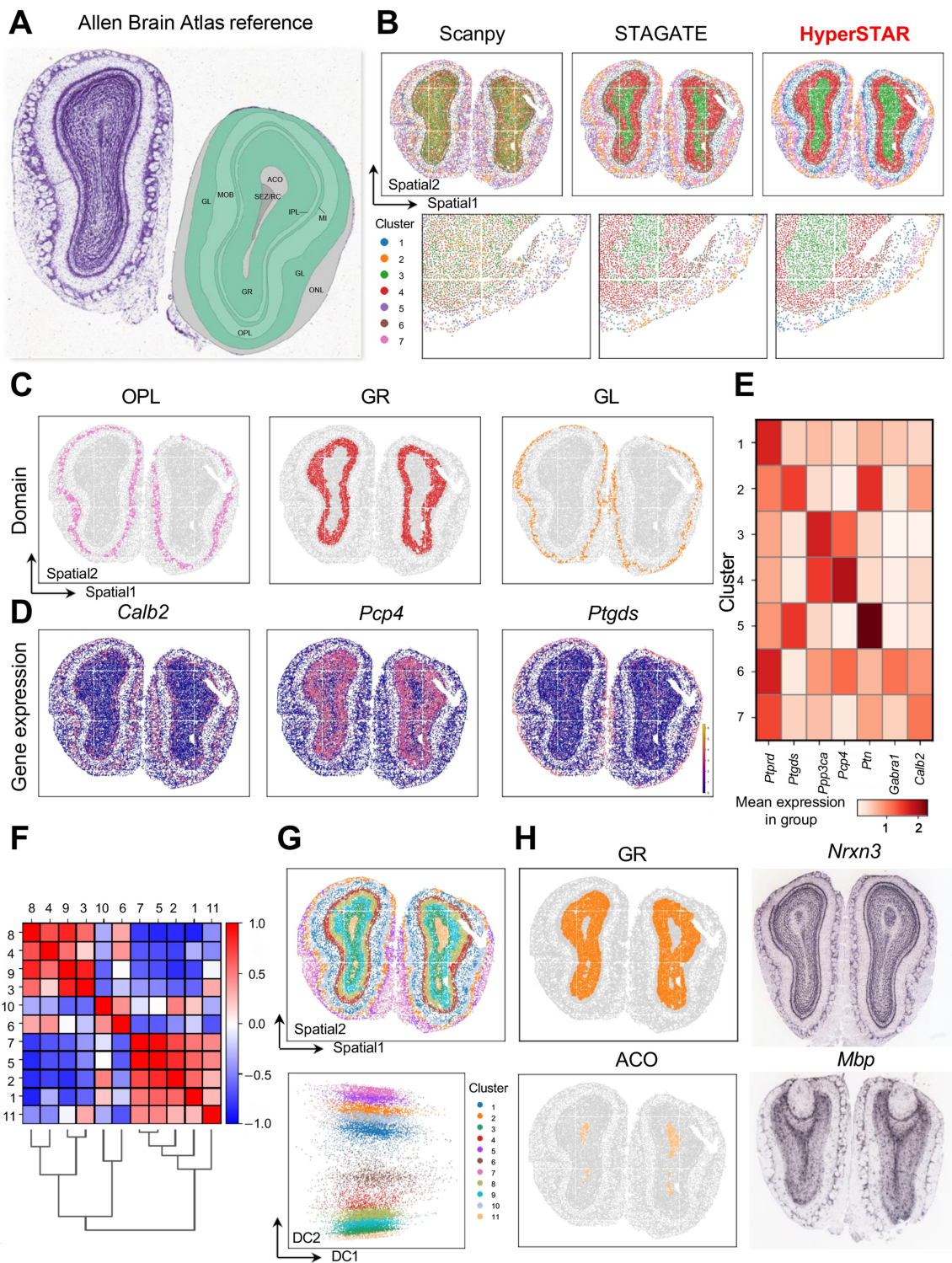


Figure 3 HyperSTAR reveals the ACO substructure at finer resolutions in the Stereo-seq MOB

A. MOB anatomical structure from the Allen Brain Atlas. **B.** Spatial visualization of results obtained by Scanpy, STAGATE, and HyperSTAR. **C.** The OPL, GL, and GR layers identified by HyperSTAR. **D.** Spatial visualization of three marker genes expressed in OPL (*Calb2*), GL (*Pcp4*), and GR (*Ptgsd*), respectively. **E.** Heatmap displaying top genes characterizing seven groups identified by HyperSTAR. **F.** Correlation matrix of HyperSTAR results, showing similarity within clusters and dissimilarity between clusters. **G.** Top: spatial visualization of results from HyperSTAR. Bottom: diffusion map revealing the ACO substructure within GR. **H.** Top: GR identified by HyperSTAR and ISH image of its marker gene *Nrnx3* from the Allen Brain Atlas; Bottom: ACO identified by HyperSTAR and ISH image of its marker gene *Mbp* from the Allen Brain Atlas. MOB, mouse olfactory bulb; SEZ/RC, subependymal zone; ACO, anterior commissure; GR, granular layer; IPL, inner plexiform layer; MI, mitral layer; OPL, outer plexiform layer; GL, glomerular layer; ONL, olfactory nerve layer; DC, dorsal cochlear; ISH, *in situ* hybridization.

(distinguishing different layers) from the Allen Brain Atlas are also provided in [Figures S20 and S21](#). These results collectively demonstrate HyperSTAR's ability to reveal continuous tissue structure and identify effective marker genes.

We adjusted the parameters for fine-grained clustering, and the result is shown in [Figure 3G](#). HyperSTAR reveals a substructure, anterior commissure (ACO), within GR, together with its top mark gene *Mbp* ([Figure 3H](#)). Additionally, we identified *Nrxn3* as the top mark gene for GR. The ISH images illustrate the effectiveness of the results obtained by HyperSTAR. The heatmap of the correlation matrix of clustering also demonstrates the effectiveness of HyperSTAR, showing similarity within clusters and dissimilarity between clusters ([Figure 3F](#)).

We further assessed HyperSTAR's cross-resolution capability on the aforementioned MOB section at various spatial resolutions. Across bin 20, bin 50, bin 100, and single-cell levels ([Figures S22 and S23](#)), our HyperSTAR results consistently exhibited a continuous structure aligned with the reference shown in [Figure 3A](#). Extending our evaluation to an adult axolotl brain slice further confirmed HyperSTAR's efficacy, revealing a continuous structure corresponding to the reference ([Figure S24](#)).

HyperSTAR unveils spatial intra-tumor heterogeneity in the breast cancer section and its association with survival

To investigate the capability of HyperSTAR for tumor analysis, we applied it to a breast cancer section obtained from 10x Visium. This section was manually annotated by a pathologist into four distinct regions: ductal carcinoma *in situ*/lobular carcinoma *in situ* (DCIS/LCIS), invasive ductal carcinoma (IDC), healthy tissue, and the tumor edge ([Figure 4A](#)) [37]. Remarkably, HyperSTAR accurately identified these regions, demonstrating high consistency with the manual annotations ([Figure 4B](#), [Figure S25](#)).

We next interrogated the diagnostic power of HyperSTAR's results. We ranked and plotted the top 5 genes for characterizing these 4 domains ([Figure 4C](#)). *KRT8* and *KRT18* were highly expressed in the IDC region. *KRT8* is an independent prognostic indicator for poor survival in lung adenocarcinoma [38]. *KRT18* often acts together with *KRT8* in interleukin-6 (IL-6)-mediated barrier protection. This patient has been clinically diagnosed as Grade III. Walker et al. claimed that *KRT8/18* expression differentiates distinct subtypes of Grade III IDC of the breast cancer [39], which is consistent with our results. We further compared transcriptional differences between IDC and the remaining regions by performing differential expression analysis followed by Gene Ontology enrichment analysis. The significantly upregulated genes in IDC were enriched in mitotic processes, indicating abnormal proliferation of cancer cells ([Figure 4D](#)). This suggests that the IDC region is probably infiltrated with cancer cells and highly active. We further investigated the correlation of these significantly upregulated genes with survival using an independent clinical dataset from The Cancer Genome Atlas (TCGA) breast cancer cohort. Ten signature genes were selected to stratify patients. The result showed that low expression of these ten signature genes correlated with shorter overall survival ([Figure 4E](#)). Hence, these results indicate a possible developmental trajectory from cancer stem cells to malignant cells.

Furthermore, some studies have indicated that *KRT8* is a pan-cancer indicator [40]. In light of this, we conducted pan-cancer survival analysis using the significantly upregulated genes identified in IDC. The outcomes further revealed a correlation between these significantly upregulated genes and pan-cancer development ([Figure S26](#)).

We further investigated the subtypes within the IDC region. According to our HyperSTAR result, the IDC region is divided into four subtypes ([Figure 4F](#)). The heatmap, generated by the top 2 genes of each subtype ([Figure 4G](#)), highlights the unique expression profiles within each subtype. Notably, for domain 4, *SLITRK6* and *CRISP3* exhibited high expression levels, with *SLITRK6* emerging as a potential marker for this specific subtype ([Figure 4H](#)). The accompanying box plot showed that *SLITRK6* was significantly upregulated in IDC_4 compared to IDC, tumor, and non-tumor regions ([Figure 4I](#)). In summary, HyperSTAR not only enables the identification of intra-tumor heterogeneity but also facilitates the exploration of distinct subtypes within the intra-tumoral environment.

HyperSTAR is generalizable

To comprehensively validate the versatility of HyperSTAR across various technologies and different omics, we applied it to multiple platforms, including 10x Genomics (10x Visium and 10x Xenium), MERFISH [41], NanoString (CosMx), and osmFISH ([Figure 5](#), [Figure S27](#)). Our evaluation was further extended to different omics layers, encompassing both transcriptomics and proteomics, demonstrating the robust adaptability and broad applicability of HyperSTAR.

HyperSTAR reveals the cortex distribution of the somatosensory cortex ([Figure 5A](#)) in a dataset produced by osmFISH, a technique based on multiplexed fluorescence *in situ* hybridization [42]. osmFISH enables the detection of hundreds of genes at single-cell resolution, providing valuable insights into the spatial organization of complex tissues. By applying HyperSTAR to this dataset, we identified distinct layers and regions within the somatosensory cortex, showcasing HyperSTAR's ability to uncover biologically meaningful structures. To further demonstrate the versatility of HyperSTAR, we applied it to the STARmap mouse brain data [43]. The resulting correlation heatmap of the clustering results showed high correlations within clusters but low correlations between clusters ([Figure 5B](#)). This indicates that HyperSTAR successfully captures the underlying structure of the STARmap data, grouping domains with similar spatial gene expression profiles while separating domains from different layers or regions. HyperSTAR explicitly revealed the distribution of brain layers ([Figure 5C and D](#)), highlighting its effectiveness in identifying spatial patterns across different experimental platforms.

The hypergraph framework is a highly scalable and versatile approach that can be applied to jointly cluster data from multiple sources. To evaluate its capability, we selected MOB SRT data from two different sequencing platforms: SlideSeqV2 and Stereo-seq. These datasets exhibited differences in quality and distribution ([Figure S28](#)). To integrate data from these platforms, we extended the hypergraph construction process beyond single slices. In addition to creating hyperedges within each slice, we established inter-slice relationships by



Figure 4. HyperSTAR identifies marker genes in breast cancer subtypes and their association with survival. **A.** Manual annotation of a breast cancer section from 10x Visium by pathology experts. **B.** Spatial result and UMAP plot generated by HyperSTAR for this section. **C.** Dot plot of marker genes identified by HyperSTAR. **D.** Gene Ontology enrichment analysis of significantly upregulated genes in IDC compared to the rest of the section. **E.** Overall survival curves for patients with 10 signature genes for IDC, analyzed using TCGA breast cancer gene expression data via GEPIA2. **F.** Subtypes identified by HyperSTAR in the breast cancer section. **G.** Heatmap showing the expression of marker genes in IDC subtypes. **H.** Expression levels of the top 5 marker genes of the IDC_SLITRK6 subtype compared to the rest of the IDC subtypes. **I.** Expression levels of *SLITRK6* in IDC_SLITRK6, IDC, tumor, and non-tumor regions. Significant difference was determined by Wilcoxon rank-sum test (IDC_SLITRK6 vs. IDC: **, $P < 0.01$; Tumor vs. Non-tumor: The Cancer Genome Atlas).

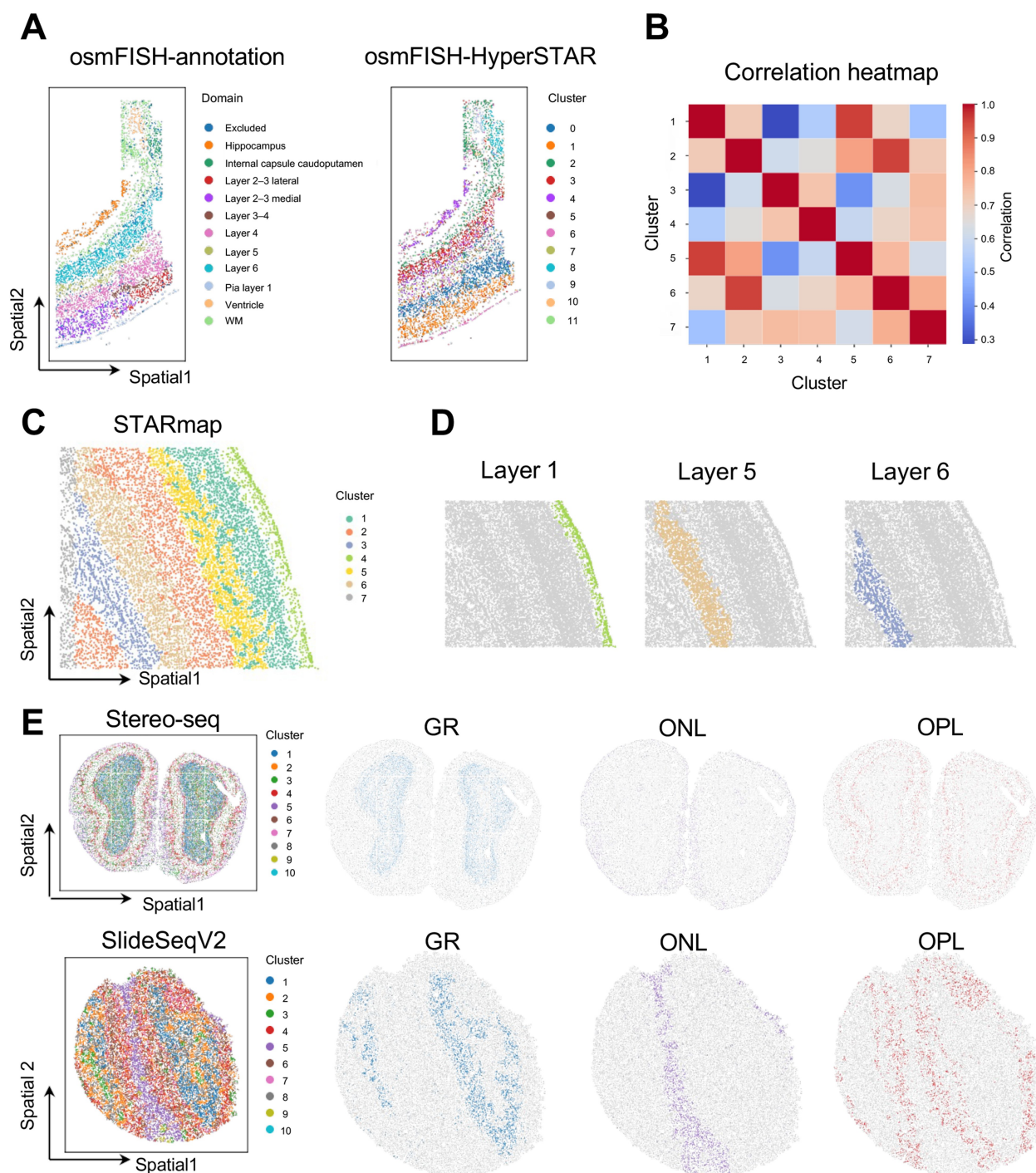


Figure 5 HyperSTAR is applicable to various spatial omics datasets and demonstrates joint clustering capabilities

A. Manual annotation and HyperSTAR-based identification of the somatosensory cortex from osmFISH data. **B.** Correlation heatmap of clustering results obtained by HyperSTAR on STARmap data. **C.** Spatial results of HyperSTAR on STARmap data from a mouse brain section. **D.** Spatial distribution of different layers revealed by HyperSTAR on STARmap data. **E.** Joint clustering of Stereo-seq and SlideSeqV2 data by HyperSTAR. The top panel shows results for Stereo-seq data: clustering results (left) and spatial distribution of different layers identified by HyperSTAR (right). The bottom panel shows results for SlideSeqV2 data: clustering results (left) and spatial distribution of different layers identified by HyperSTAR (right). osmFISH, ouroboros single-molecule fluorescence *in situ* hybridization.

constructing hyperedges that connect spots across slices based on their normalized log-transformed gene expression profiles. This approach allows for the identification of shared spatial patterns across different datasets.

To address potential batch effects arising from different sequencing platforms, we utilized Harmony for batch effect correction. The joint clustering results generated by HyperSTAR (Figure 5E, Figure S29) successfully delineated distinct layers

within the MOB, aligning with its known anatomical structure. By leveraging the complementary information provided by SlideSeqV2 and Stereo-seq, HyperSTAR effectively identified shared spatial patterns across datasets, offering a more comprehensive understanding of tissue organization.

Discussion

The advent of spatial omics technologies has revolutionized our understanding of biological processes by enabling the simultaneous acquisition of biomolecular and spatial information. Accurate identification of spatial domains is crucial for deciphering spatiotemporal life processes. In this study, we introduce HyperSTAR, a powerful and versatile method for spatial omics analysis that excels in accurately identifying spatial domains, even in high-resolution spatial data. By incorporating a gene expression-guided hyperedge decomposition module and a hypergraph attention convolutional neural network, HyperSTAR demonstrates exceptional precision in recognizing spatial domains across varying resolutions. The advancements hold significant implications for exploring organ development and tumor microenvironments.

Benchmarking and comparisons with existing advanced graph neural network models underscore HyperSTAR's superior performance in uncovering tissue substructures, inferring spatiotemporal patterns, and denoising spatially resolved gene expression. Its application to spatial omics data from diverse platforms — including 10x Genomics, NanoString, CODEX, and osmFISH — highlights its adaptability and robustness across different technologies. Furthermore, the validation of HyperSTAR's findings through functional and survival analyses of independent clinical data, particularly in the context of breast cancer, reinforces its utility in deciphering spatial heterogeneity and its potential for clinical relevance.

A key strength of HyperSTAR lies in its ability to capture intricate higher-order relationships among spatially neighboring multi-spots. However, the current version of HyperSTAR is primarily focused on identifying spatial domains and has yet to fully exploit the potential of hypergraphs. In future work, we aim to extend HyperSTAR to integrate multi-omics data from spatial omics experiments generated by different platform technologies. Additionally, incorporating histology images could provide valuable insights for addressing clinically relevant questions and further enhancing the method's applicability.

In conclusion, HyperSTAR addressed challenges associated with diverse spatial resolutions and enabling the exploration of complex biological systems. Its ability to capture intricate higher-order relationships within spatially neighboring multi-spots makes it an invaluable tool for researchers studying organ development, tumor microenvironments, and other spatially resolved biological processes. This study establishes HyperSTAR as a versatile and effective computational approach with broad applicability across diverse spatial omics datasets and large-scale analyses. Future directions include refining and expanding HyperSTAR's capabilities, integrating multi-omics data and histology images, and exploring its application in additional biological contexts.

Code availability

The HyperSTAR method is implemented in Python. It is available on GitHub (<https://github.com/Ringoio/HyperSTAR>). The code has also been submitted to BioCode at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) (BioCode: BT007916), which is publicly accessible at <https://ngdc.cncb.ac.cn/biocode/tools/BT007916>.

CRedit author statement

Yi Liao: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Chong Zhang:** Investigation, Data curation, Formal analysis, Writing – review & editing. **Zhikang Wang:** Conceptualization, Methodology, Visualization, Writing – review & editing. **Fei Qi:** Conceptualization, Methodology, Writing – review & editing. **Weitian Huang:** Conceptualization, Methodology, Writing – review & editing. **Shangyan Cai:** Visualization, Writing – review & editing. **Jinyu Li:** Visualization, Writing – review & editing. **Jiazhou Chen:** Writing – review & editing. **Robin B. Gasser:** Writing – review & editing. **Zhiyuan Yuan:** Supervision, Writing – review & editing. **Jiangning Song:** Supervision, Project administration, Funding acquisition, Writing – review & editing. **Hongmin Cai:** Supervision, Project administration, Funding acquisition, Writing – review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzaf128>).

Acknowledgments

This work was supported by the National Science and Technology Major Project, China (Grant No. 2022ZD0117700), the National Natural Science Foundation of China (Grant No. 62325204), the Australian National Health and Medical Research Council (NHMRC; Grant No. APP 2020646), the Australian Research Council Linkage project grant (Grant No. LP220200614), and the Major and Seed Inter-Disciplinary Research Projects awarded by Monash University, Australia.

ORCID

0009-0007-4577-8501 (Yi Liao)
 0009-0006-3719-4592 (Chong Zhang)
 0000-0001-9587-1965 (Zhikang Wang)
 0000-0001-7128-7764 (Fei Qi)
 0000-0001-8848-4801 (Weitian Huang)
 0009-0000-2444-6718 (Shangyan Cai)
 0000-0002-2472-8958 (Jinyu Li)

0000-0001-7171-9547 (Jiazhou Chen)
 0000-0002-4423-1690 (Robin B. Gasser)
 0000-0002-9367-4236 (Zhiyuan Yuan)
 0000-0001-8031-9086 (Jiangning Song)
 0000-0002-2747-7234 (Hongmin Cai)

References

- [1] Bressan D, Battistoni G, Hannon GJ. The dawn of spatial omics. *Science* 2023;381:eabq4964.
- [2] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6:377–82.
- [3] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573–87.e29.
- [4] Efremova M, Teichmann SA. Computational methods for single-cell omics across modalities. *Nat Methods* 2020;17:14–7.
- [5] Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;596:211–20.
- [6] Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;19:534–46.
- [7] Lewis SM, Asselin-Labat ML, Nguyen Q, Berthelet J, Tan X, Wimmer VC, et al. Spatial omics and multiplexed imaging to explore cancer biology. *Nat Methods* 2021;18:997–1012.
- [8] Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. *Genome Med* 2022;14:68.
- [9] Zhuang X. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nat Methods* 2021;18:18–22.
- [10] Larsson L, Frisén J, Lundeberg J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat Methods* 2021;18:15–8.
- [11] Jin L, Lloyd RV. *In situ* hybridization: methods and applications. *J Clin Lab Anal* 1997;11:2–9.
- [12] Jensen E. Technical review: *in situ* hybridization. *Anat Rec (Hoboken)* 2014; 297:1349–53.
- [13] Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell *in situ* RNA profiling by sequential hybridization. *Nat Methods* 2014;11:360–1.
- [14] Eng CHL, Lawson M, Zhu Q, Dries R, Koulina N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 2019;568:235–9.
- [15] Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;348:aaa6090.
- [16] Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, et al. *In situ* sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* 2013;10:857–60.
- [17] Wang X, Allen WE, Wright MA, Sylvestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;361:eaat5691.
- [18] Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet* 2010; 11:31–46.
- [19] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–51.
- [20] Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011;470:198–203.
- [21] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell* 2002;24:881–92.
- [22] Krishna K, Murty MN. Genetic K-means algorithm. *IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics)* 1999;29:433–9.
- [23] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15.
- [24] Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;39:1375–84.
- [25] Pham D, Tan X, Balderson B, Xu J, Grice LF, Yoon S, et al. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat Commun* 2023;14:7739.
- [26] Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021; 18:1342–51.
- [27] Ren H, Walker BL, Cang Z, Nie Q. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat Commun* 2022;13:4076.
- [28] Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 2022;13: 1739.
- [29] Maynard KR, Collado-Torres L, Weber LM, Uyttingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;24:425–36.
- [30] Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 2019;33:3558–65.
- [31] Gao Y, Feng Y, Ji S, Ji R. HGNN⁺: general hypergraph neural networks. *IEEE Trans Pattern Anal Mach Intell* 2023; 45:3181–99.
- [32] Chen C, Cheng Z, Li Z, Wang M. Hypergraph attention networks. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) 2020:1560–5.
- [33] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. *The 6th International Conference on Learning Representations* 2018.
- [34] Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* 2016;8:289–317.
- [35] Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;20:59.
- [36] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *The 5th International Conference on Learning Representations (Poster)* 2017.
- [37] Xu C, Jin X, Wei S, Wang P, Luo M, Xu Z, et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res* 2022;50:e131.

- [38] Xie L, Dang Y, Guo J, Sun X, Xie T, Zhang L, et al. High *KRT8* expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes (Basel)* 2019;10:36.
- [39] Walker LC, Harris GC, Holloway AJ, McKenzie GW, Wells JE, Robinson BA, et al. Cytokeratin *KRT8/18* expression differentiates distinct subtypes of grade 3 invasive ductal carcinoma of the breast. *Cancer Genet Cytogenet* 2007;178:94–103.
- [40] Scott MKD, Limaye M, Schaffert S, West R, Ozawa MG, Chu P, et al. A multi-scale integrated analysis identifies *KRT8* as a pan-cancer early biomarker. *Pac Symp Biocomput* 2021; 26:297–308.
- [41] Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* 2021; 598: 137–43.
- [42] Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* 2018; 15:932–5.
- [43] Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018; 361:eaat5691.