Yu *et al. Journal of Translational Medicine* (2025) 23:1416
https://doi.org/10.1186/s12967-025-07404-3

Journal of Translational Medicine

RESEARCH Open Access

HiSTaR: identifying spatial domains with hierarchical spatial transcriptomics variational autoencoder

Junhua Yu[1†], Jiaqi Yuan[1†], Qianbei Yi[1], Zheng Ye[1,2*], Peng Xu[1,2*] and Wenbin Liu[1*]

Abstract

Background The development of spatial transcriptomics (ST) has enabled biologists to measure transcriptome data on an entire tissue and retain spatial information. It gives us the opportunity to fully understand the tissue microenvironment and identify spatial domains. Deep learning techniques provide an effective way to capture the latent representation of spatial transcriptomics data.

Methods In this paper, we propose a Hierarchical Spatial Transcriptomics variational autoencoder (HiSTaR) that employs multiple HiSTaR blocks to capture multi-level latent features from spots. These features were subsequently utilized for downstream analyses, including spatial domain identification and batch effect correction.

Results HiSTaR tends to perform well in identifying spatial domains across multiple datasets from diverse platforms, consistently showing superior results compared to existing methods. HiSTaR also supports trajectory analysis and differential gene expression analysis, contributing to its validation. Furthermore, HiSTaR facilitates seamless integration of multiple tissue slices, effectively addressing batch effects without external tools—a key advantage for large-scale spatial transcriptomics studies.

Conclusion HiSTaR offers an effective computational tool for spatial transcriptomics research. By capturing hierarchical latent features, it improves spatial domain identification. This framework holds potential to advance understanding of tissue heterogeneity and spatially resolved gene expression patterns.

Introduction

Spatial transcriptomics combines gene expression with spatial localization data, addressing the limitations of traditional transcriptomics [1, 2]. This integration enables a detailed analysis of tissue heterogeneity and underlying physiological mechanisms, offering valuable reference for clinical applications [3]. A central challenge in spatial transcriptomics is identifying spatial domains—regions with unique gene expression profiles that reflect functional tissue structures, such as tumor microenvironment boundaries or neuronal layers in brain tissue [4]. The advent of deep learning facilitates this process by compressing high-dimensional gene expression data from

†Junhua Yu and Jiaqi Yuan contributed equally to this work.

*Correspondence:
Zheng Ye
zheng_ye@gzhu.edu.cn
Peng Xu
gdxupeng@gzhu.edu.cn
Wenbin Liu
wbliu6910@gzhu.edu.cn
[1]Institute of Computational Science and Technology, Guangzhou University, Guangzhou 511370, China
[2]School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, China

each spot into effective latent representations, enhancing the accuracy of downstream analyses.

Current methods for extracting latent representations primarily fall into two categories: non-spatial methods and spatial methods [5]. Non-spatial methods rely solely on gene expression data. The Louvain [6, 7] optimizes modularity by iteratively grouping spots into communities based on expression profiles, aiming to maximize intra-community similarity. Similarly, the Leiden [7, 8] refines this process by introducing a refinement step to correct for poorly connected nodes, improving cluster coherence. However, both methods usually resulted in fragmented clustering that divides continuous spatial regions into multiple pseudo-spatial domains because of ignoring spatial relationships between spots [9].

Spatial clustering methods combine spatial information and gene expression data, enabling precise analysis of biological tissue spatial functional units. ST-SCSR [10] employed spatial networks through a matrix three-factor decomposition strategy, combining local, global, and meta-structural information (such as relations among spatial domains), extracting more effective latent representations. SEDR [11] employed a variational graph autoencoder [12] with GCN [13] to ultimately derive latent representations. STAGATE [14] adopted an attention mechanism [15] to adaptively learn the similarity of neighboring spots and employed an adaptive graph attention autoencoder to ultimately derive effective latent representations. STMGraph [16] employed the Dynamic Graph Attention Transformer (DGAT) [17] and a dual re-masking algorithm, ultimately deriving latent representations. DeepST [5] enhanced ST data by extracting morphological features of image organization and employed denoising autoencoder with GCN, ultimately producing latent representations.

In this paper, we propose Hierarchical ST variational autoencoder (HiSTaR), which employs hierarchical blocks to extract multi-level latent representations. Results from diverse platforms, including 10x Genomics Visium [18], Slide-seqV2 [19], Stereo-seq [20, 21], and STARmap [22], demonstrates the multi-level latent representations by HiSTaR have superior spatial domain identification compared to advanced methods like STAGATE and STMGraph. Additionally, by capturing hierarchical latent features, HiSTaR serves as an effective solution for batch correction, enabling direct alignment of multi-sample data and overcoming a major challenge in integrating heterogeneous tissue slices [23, 24]. Therefore, HiSTaR represents a potential model, delivering superior spatial domain identification performance in spatial transcriptomics analysis.

Results

Overview of HiSTaR framework

Figure 1 shows an overview of the HiSTaR framework, which includes three key components: data processing, the HiSTaR model (comprising Encoder and Decoder), and downstream analysis. During data processing, the expression matrix is randomly masked [11], and k-nearest neighbor (KNN) [25] is employed to construct an adjacency matrix from spatial transcriptomics coordinates, effectively capturing spatial relationships. The HiSTaR model employs a hierarchical architecture to extract latent representations. The encoder comprises a fully connected layer followed by two hierarchical HiSTaR blocks. Initially, the fully connected layer extracts the base feature matrix $F_0 \in \mathbb{R}^{N \times d_0}$, where N represents the number of spots and d_0 represents the dimension of the base features. The first-level HiSTaR block uses the base feature matrix and the adjacency matrix to generate the initial spatial representation $F_1 \in \mathbb{R}^{N \times d_1}$, where d_1 represents the dimension of the initial spatial representation. The second-level HiSTaR block then processes initial spatial representation F_1 and the adjacency matrix to produce the enhanced spatial representation $F_2 \in \mathbb{R}^{N \times d_2}$, where d_2 represents the dimension of the enhanced spatial representation. Then, the base feature matrix F_0, the initial spatial representation F_1, and the enhanced spatial representation F_2 are concatenated to form the final latent representation $Z \in \mathbb{R}^{N \times (d_0+d_1+d_2)}$. The HiSTaR decoder reconstructs the gene expression matrix and the adjacency matrix from the latent representation Z. Finally, The HiSTaR framework uses the final latent representation Z to perform downstream analyses, including spatial domain identification and batch-effects correction, facilitating the study of tissue heterogeneity.

HiSTaR demonstrates enhanced performance in identifying known domains in Human Dorsolateral Prefrontal Cortex Dataset

Human dorsolateral prefrontal cortex dataset (DLPFC) [26] comprise 12 slices from three distinct human brains, collected using the 10x Genomics Visium spatial transcriptome platform. Figure 2(A) illustrates a visualization of slice 151,673 from the DLPFC dataset, highlighting manually annotated cortical layers (Layer_1 to Layer_6) and white matter (WM) present in each slice of the dataset. We compare the clustering performance of HiSTaR with several advanced methods, including Leiden Louvain, ST-SCSR, SEDR, STAGATE, STMGraph, and DeepST. Figure 2(B) illustrates a visualization of the spatial domain identification results for above methods on DLPFC slice #151673, presenting the adjusted Rand index (ARI) [27] scores for each method, where HiSTaR displays smooth results with clear edges and records the

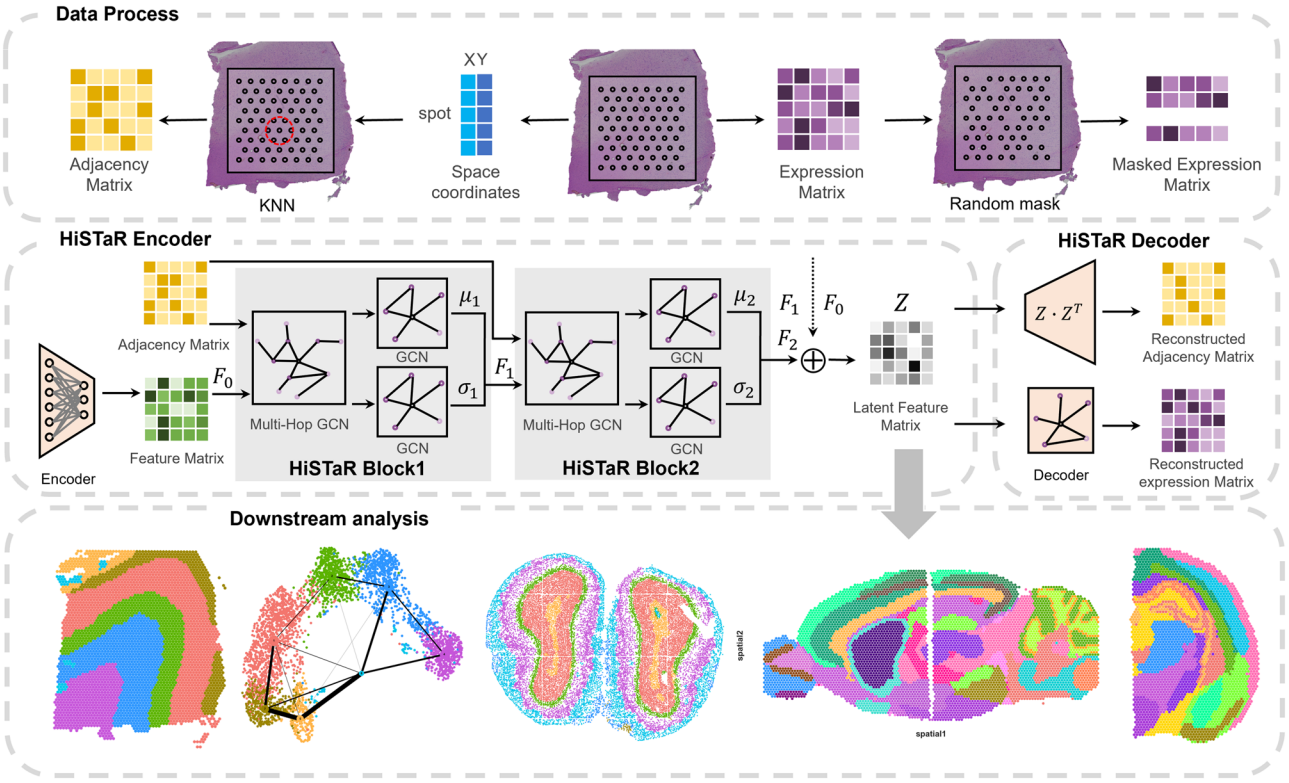

Fig. 1 Overview of HiSTaR framework. Framework comprises three components: data process, HiSTaR (encoder and decoder), and downstream analysis. In data process, the gene expression matrix is randomly masked, and the adjacency matrix is constructed. The encoder includes a fully connected network and two-level HiSTaR blocks to capture multilevel latent representations and the decoder reconstructs the expression matrix and the adjacency matrix. Latent representations can be used to perform downstream analyses, including spatial domain identification and batch-effects correction

highest ARI score of 0.7182, comparing favorably with the other methods.

Figure 2(C) shows a boxplot of the adjusted Rand index (ARI) [27], normalized mutual information (NMI) [28], and Fowlkes-Mallows score (FMS) [29] scores for the clustering results of the tested models across 12 DLPFC dataset slices. HiSTaR achieving a median ARI of 0.65 and a peak ARI of 0.76 across 12 DLPFC dataset slices, with median NMI and FMS of 0.73 and 0.76 respectively, demonstrating effective clustering performance. In comparison, non-spatial methods like Leiden and Louvain, which rely solely on gene expression data, produce lower median ARI scores of approximately 0.3, often resulting in fragmented clusters due to their neglect of spatial relationships. Spatial clustering methods, such as STAGATE and STMGraph, use graph attention networks (GAT and DGAT, respectively) to incorporate spatial dependencies, yielding median ARI scores of around 0.54 and median FMS scores of 0.63 and 0.66, respectively. The ARI, NMI, and FMS scores for the 12 slices of DLPFC, derived from the above methods, are presented in Supplementary Fig. S1 and Tables S1–S3.

Additionally, Fig. 2(D) shows the UMAP [30] of PAGA trajectory [31] for STAGATE, STMGraph, and HiSTaR on the DLPFC slice #151673. Their trajectories are 1→3→WM→5→2→6→4, 1→6→2→3→4→5→WM and 6→5→WM→1→2→3→4 respectively. Obviously, STMGraph and HiSTaR could capture the layer orders somewhat better aligning with the expected order, while STAGATE completely fails in this view.

HiSTaR effectively identifies tumor domains in human breast cancer dataset

Human breast cancer dataset (HBC) [32] from the 10x Genomics Visium spatial transcriptome platform includes 3798 spots and 36,601 genes derived from a single tissue slice. Figure 3(A) illustrates a visualization of a breast cancer tissue slice with manually annotated regions, including ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), invasive ductal carcinoma (IDC), healthy tissue (Healthy), and tumor periphery (Tumor_edge), where a black box highlights two specific cancer regions (IDC_3 in the lower left and DCIS/LCIS_2 at the right edge) for evaluate spatial domain identification methods.

Figure 3(B) illustrates a visualization of the spatial domain identification for HiSTaR, STAGATE, and STMGraph on the HBC dataset, featuring two fine-grained regions, IDC_3 and DCIS/LCIS_2, marked by two black boxes. Beyond IDC_3 and DCIS/LCIS_2, HiSTaR excels

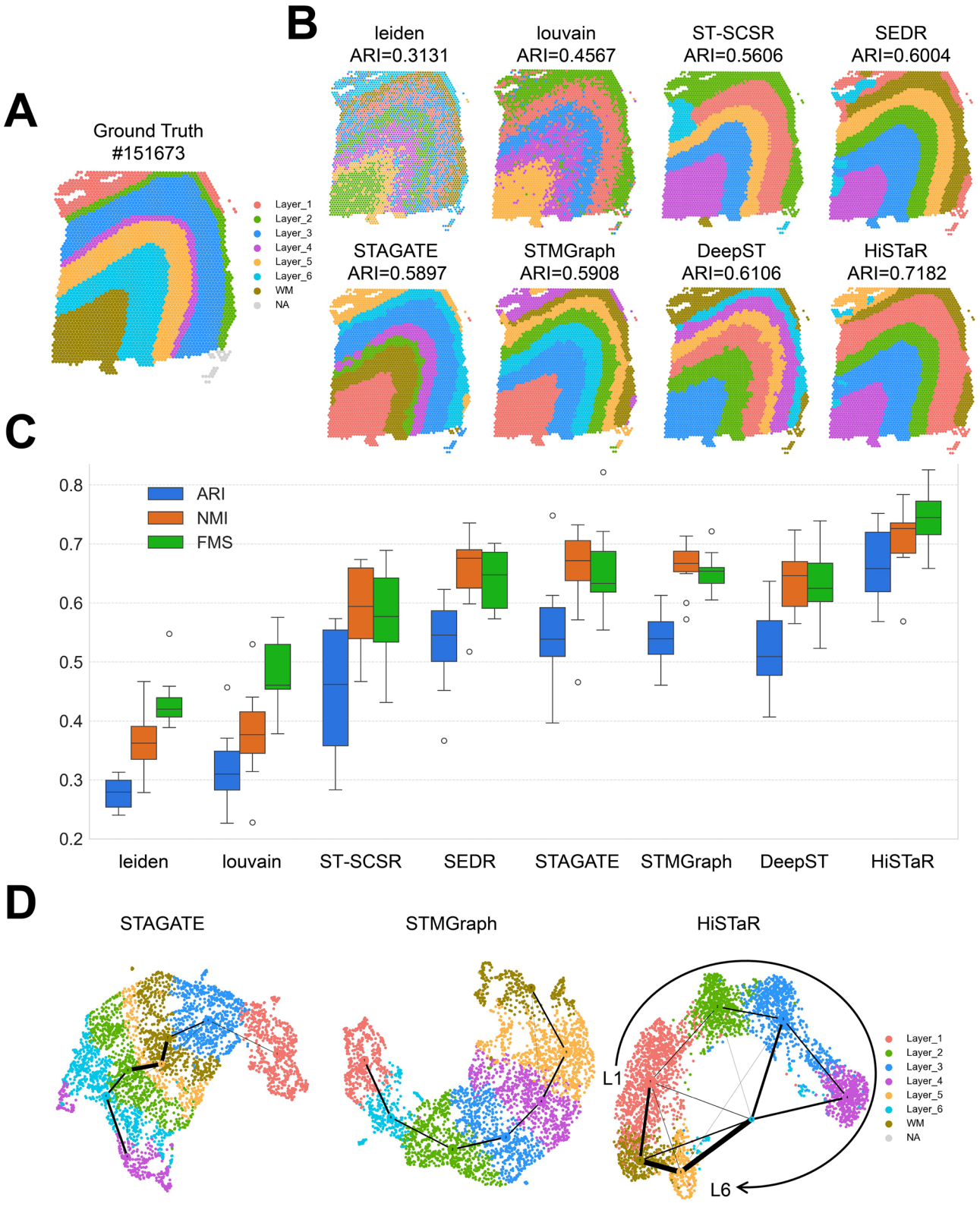

Fig. 2 HiSTaR is evaluated in human DLPFC dataset. (**a**) Ground truth of spatial domains in slice #151673 from the human DLPFC dataset. (**b**) Spatial domain identification results on slice #151673 from the human DLPFC dataset, generated by Leiden, Louvain, ST-SCSR, sedr, STAGATE, STMGraph, DeepST, and HiSTaR. (**c**) Quantitative comparison of spatial domain identification across all 12 slices of the human DLPFC dataset, evaluated using ARI, NMI, and FMS metrics. (**d**) Umap plots and PAGA visualizations of the latent space representations learned by STAGATE, STMGraph, and HiSTaR on slice #151673 from the human DLPFC dataset

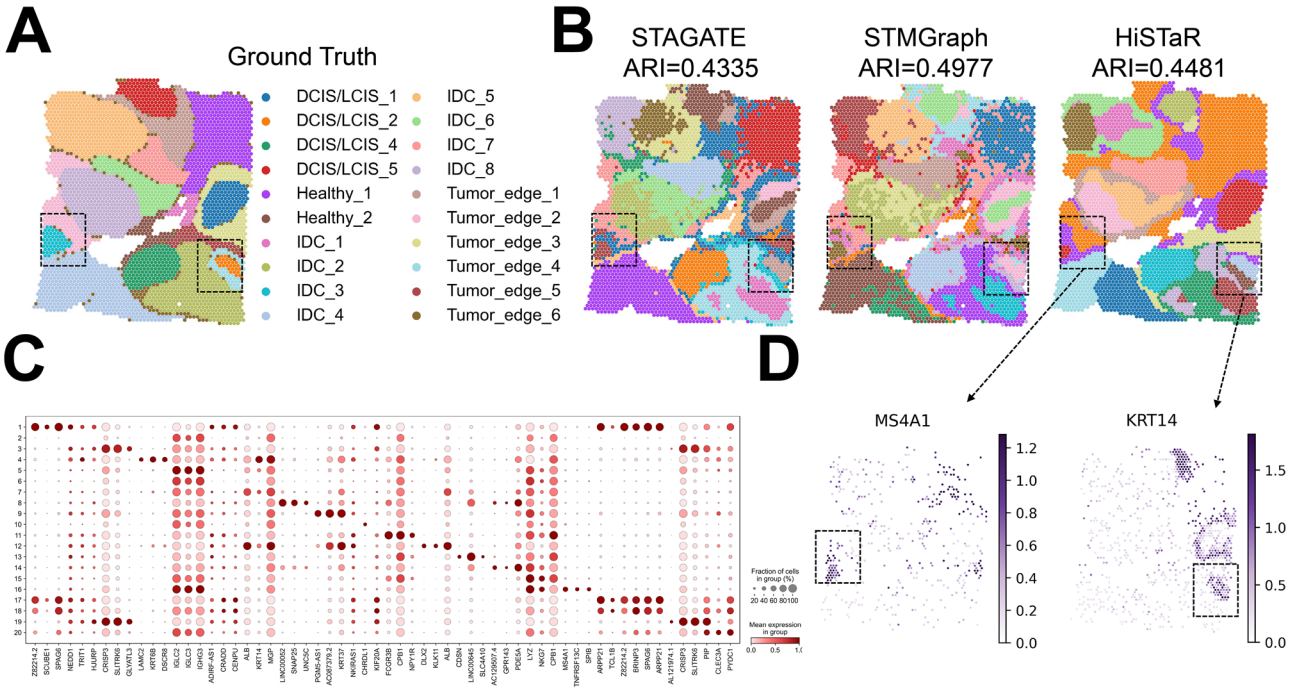

Fig. 3 HiSTaR is evaluated on human breast cancer dataset. (**a**) Ground truth annotations of spatial domains in the human breast cancer dataset. (**b**) Spatial domain identification results on the human breast cancer dataset, generated by STAGATE, STMGraph, and HiSTaR. (**c**) Dot plot of the top three differentially expressed genes (DEGs) in each spatial domain. Circle color indicates gene expression level, while circle size reflects the proportion of spots expressing the gene. (**d**) Representative spatial expression patterns of MS4A1 (left) and KRT14 (right)

in separating healthy tissue from tumor margins, with an ARI of 0.4481 compared to STAGATEs 0.4335 and STMGraph's 0.4977. This improved spatial coherence, when compared to STAGATE, is evident in the smoother transitions and clearer boundaries observed in the spatial visualizations, providing a more biologically intuitive separation of these critical domains.

To investigate the molecular basis of these spatial domains, we perform a differential expression genes (DEGs) analysis [33] to identify marker genes associated with each region [34]. Figure 3(C) shows Dot plot of the top three DEGs for each spatial domain, while Fig. 3(D) shows two key marker genes: MS4A1 (left) and KRT14 (right). MS4A1 expression in IDC_3, consistent with its role in enhancing immune cell activity (e.g., CD8+ T cells) and potentially inhibiting tumor progression [35]. Similarly, KRT14 in the DCIS/LCIS_2 domain, which indicate that DCIS retains more basal epithelial cell characteristics. This aligns with KRT14's high expression in DCIS in breast cancer tissue samples [36].

HiSTaR effectively identifies spatial domains across diverse platforms

Figure 4(A) shows a brain atlas from the Allen Mouse Brain Atlas [37] as a reference, while Fig. 4(B) shows an H&E-stained image from a mouse brain dataset obtained via the 10x Genomics Visium platform, together offering support for the subsequent spatial domain identification analysis. Figure 4(C) illustrates a visualization of HiSTaR's delineation of key brain regions, including the cornu ammonis and dentate gyrus, with boundaries aligning with reference annotations, compared to STAGATE and STMGraph, which show less consistent alignment. The UMAP plot presents the clustering performance of HiSTaR, STAGATE, and STMGraph, conveying greater cohesion within domains and clearer separation between domains for HiSTaR.

Figure 4(D) shows a DAPI-stained image as a reference for the mouse olfactory bulb (MOB) dataset obtained from the Stereo-seq platform. Figure 4(E) illustrates a visualization of the spatial domain identification results for HiSTaR, STAGATE, and STMGraph, featuring three regions (the rostral migratory stream (RMS), granule cell layer (GCL), and external plexiform layer (EPL)) marked by a black box. In the zoomed-in view, HiSTaR displays the clearest edges, while STAGATE and STMGraph appear cluttered, suggesting HiSTaR may offer better-defined boundaries that align with biological references. The UMAP plot displays the spot distributions of HiSTaR, STAGATE, and STMGraph, revealing HiSTaR's spot distribution as more gathered with greater cohesion within domains and clearer separation between domains. The silhouette coefficient (SC) [38], which assesses cluster cohesion and separation, registers values of 0.5609 (HiSTaR), 0.5313 (STAGATE), and 0.5212 (STMGraph),

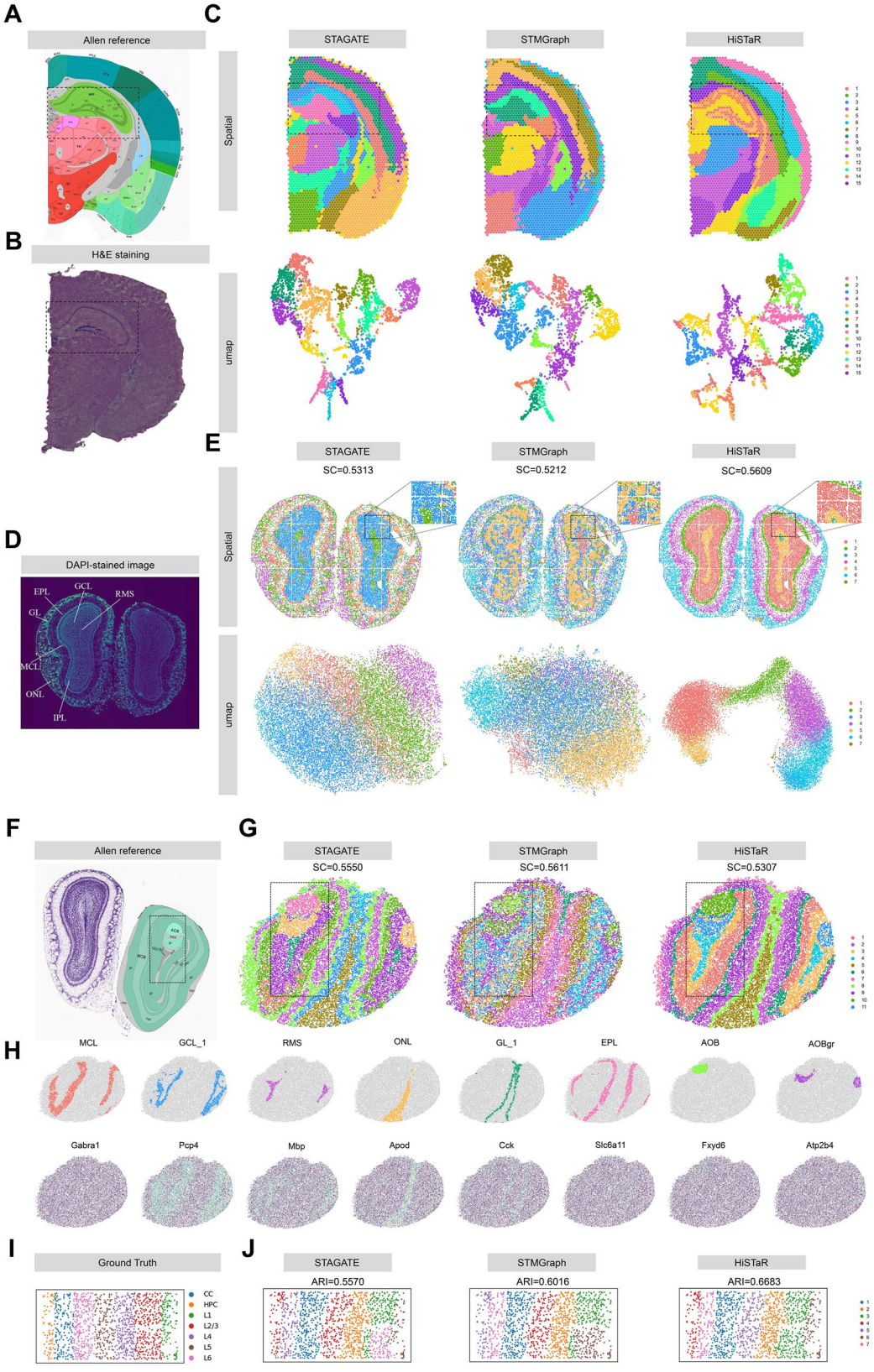

Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 HiSTaR is evaluated on multiple datasets from diverse platforms. (**a**) Reference atlas from the Allen Mouse Brain atlas. (**b**) H&E-stained image of the mouse brain tissue section. (**c**) Spatial domain identification results and UMAP plots of the mouse brain dataset generated by STAGATE, STMGraph, and HiSTaR, based on the 10x genomics visium platform. (**d**) DAPI-stained image of the mouse olfactory bulb. (**e**) Spatial domain identification results and UMAP plots of the mouse olfactory bulb dataset generated by STAGATE, STMGraph, and HiSTaR, based on the stereo-seq platform. (**f**) Reference atlas from the Allen Mouse olfactory bulb atlas. (**g**) Spatial domain identification results of the mouse olfactory bulb dataset generated by STAGATE, STMGraph, and HiSTaR, based on the silde-seqV2 platform. (**h**) Spatial domains identified by HiSTaR (upper) and the corresponding expression patterns of a specific biomarker genes (lower). (**i**) Ground truth of mouse visual cortex dataset. (**j**) Spatial domain identified by STAGATE, STMGraph, and HiSTaR, based on the STARmap platform

suggesting HiSTaR exhibit more defined spatial domain boundaries.

Figure 4(F) shows a MOB atlas from the Allen Mouse Brain Atlas as a reference for the mouse olfactory bulb (MOB) dataset from the Slide-seqV2 platform. Figure 4(G) illustrates a visualization of the spatial domain identification for HiSTaR, STAGATE, and STMGraph, featuring three regions (the accessory olfactory bulb (AOB), accessory olfactory bulb granular layer (AOBgr), and rostral migratory stream (RMS)) marked by a black box. In this view, HiSTaR achieves a silhouette score of 0.5307, compared to 0.5550 for STAGATE and 0.5611 for STMGraph, indicating competitive intra-domain cohesion and inter-domain separation despite a slight numerical difference. HiSTaR displays the clearest edges, while STAGATE and STMGraph appear cluttered, suggesting a potential difference in boundary definition. Figure 4(H) shows the spatial visualization, witch upper panel displaying HiSTaR's identified spatial domains and the lower panel presenting the high-expression marker genes for these domains, such as Mbp for the rostral migratory stream (RMS) and Fxyd6 for the accessory olfactory bulb (AOB), with the identification suggesting alignment with these gene expressions. Higher resolution figure in Supplementary Material Fig. S2.

Figure 4(I) shows the Ground Truth for the mouse visual cortex dataset from the STARmap platform as a reference. Figure 4(J) illustrates a visualization of the spatial domain identification for HiSTaR, STAGATE, and STMGraph. Within this visualization, HiSTaR exhibits distinct layering for L1, represented by green spots, and L5, represented by brown spots, while STAGATE shows overlap across three different regions, and STMGraph's identification reveals L1 encroaching on the L5 area. The adjusted Rand index (ARI) scores are 0.5570 for STAGATE, 0.6016 for STMGraph, and 0.6683 for HiSTaR, reflecting their respective accuracies.

In this section, we employ ST data from 10x Genomics Visium, Stereo-seq, Slide-seqV2, and STARmap platforms to evaluate HiSTaR for spatial domain identification. Results indicating that, compared to STAGATE and STMGraph, HiSTaR tends to show improved performance through smoother boundaries and more consistent identification. This suggests that HiSTaR holds potential for generalizability across multi-platform data in spatial transcriptomics research.

HiSTaR effectively corrects batch-effects across multiple slices

We assess HiSTaR's potential to correct batch-effects using three datasets, obtained from the 10x Genomics Visium platform. The first includes two vertical sections (Sect. 1 & Sect. 2) of mouse breast cancer tissue, capturing tumor heterogeneity. The second includes two vertical sections (Sect. 1 & Sect. 2) of mouse brain tissue, representing complex neural structures. The third includes of two horizontal sections (Anterior & Posterior) of mouse brain tissue.

For the vertical sections, we first examine mouse breast cancer dataset, with aligned images shown in Fig. 5(A). Figure 5(B) shows UMAP plots visualizing clustering patterns between Sect. 1 & Sect. 2, highlighting batch-effects through differences in cluster distributions. HiSTaR, unlike methods relying on PASTE [39], directly corrects these batch-effects. Figure 5(C) shows a boxplot displaying the integrated local inverse Simpson index (iLISI) [40] scores for HiSTaR, STAGATE, and STMGraph, where HiSTaR and STAGATE exhibit median scores of approximately 1.4, suggesting a trend of similar performance, while both lead STMGraph with its median score of around 1.3. Fig. 5(D) illustrates a visualization of the integration performance for HiSTaR, STAGATE, and STMGraph, featuring both UMAP plots and spatial domain identification results. The UMAP visualization presents the distribution of spots for the two slices, aiming for uniform mixing of blue spots representing Sect. 1 and orange spots representing Sect. 2. In the UMAPs for STAGATE and STMGraph, gaps persist between the blue spots of Sect. 1 and the orange spots of Sect. 2, whereas HiSTaR's UMAP suggests greater uniformity in blending these two sections. Additionally, the UMAP colored by cluster assignments indicates that HiSTaR and STAGATE exhibits better separation between different categories compared to STMGraph. Next, it is the spatial domain identification results of above methods, where HiSTaR suggests clearer domain boundaries and better alignment between the two sections, while STAGATE and STMGraph tend to exhibit scattered spots and poorer alignment between their identified regions. Similarly, for the mouse brain coronal vertical sections, Fig. 5(E) shows aligned images and Fig. 5(F) shows UMAP plots that together illustrate the presence of batch-effects across mouse brain vertical sections. Figure 5(G) shows

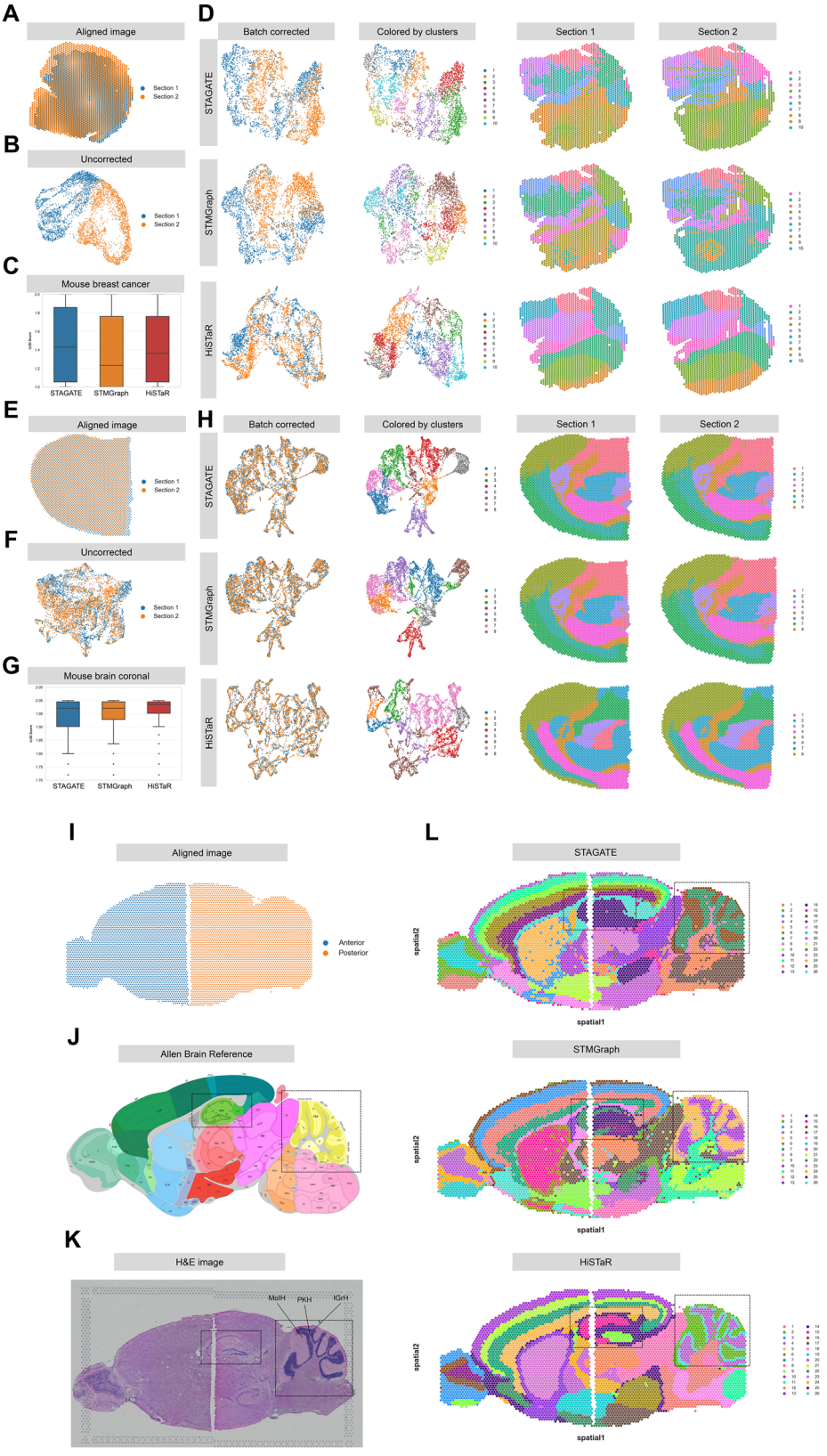

Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 HiSTaR corrects batch-effects on multiple slices. (**a**) Vertical alignment of spatial Sect. 1 and Sect. 2 from the mouse breast cancer dataset. (**b**) UMAP plot of Sect. 1 and Sect. 2. (**c**) Boxplot of iLISI scores for STAGATE, STMGraph, and HiSTaR. (**d**) Clustering results (left) and spatial domain identification results (right) after batch-effects correction by STAGATE, STMGraph, and HiSTaR. (**e**) Vertical alignment of spatial Sect. 1 and Sect. 2 from the mouse brain dataset. (**f**) UMAP plot of Sect. 1 and Sect. 2. (**g**) Boxplot of iLISI scores for STAGATE, STMGraph, and HiSTaR. (**h**) Clustering results (left) and spatial domain identification results (right) after batch-effects correction by STAGATE, STMGraph, and HiSTaR. (**i**) Alignment of anterior and posterior sections from the mouse brain dataset. (**j**) Reference annotations from the Allen mouse brain Atlas. (**k**) H&E-stained images of aligned anterior and posterior, with PkH, MolH, and IGrH. (**l**) Spatial domain after batch-effects correction by STAGATE, STMGraph, and HiSTaR

a boxplot presenting the iLISI scores for HiSTaR, STAGATE, and STMGraph, where HiSTaR records the highest median score of 1.97, indicating its potential to address batch-effects. Figure 5(H) illustrates a visualization of the integration performance for these methods, featuring UMAP plots and spatial domain identification results; the UMAP plots reveal the mixing of spots post-integration, with HiSTaR showing the most uniform blending across the two sections, while the spatial domain identification suggests HiSTaR's edges appear the clearest, compared to the less defined boundaries of STAGATE and STMGraph.

For the horizontal sections of the mouse brain, we analyze the Anterior and Posterior sections, Fig. 5(I)–(K) shows an aligned visual, reference annotations from the Allen Mouse Brain Atlas, and an H&E-stained visualization, respectively, supporting as references the experimental analysis, where the annotations and H&E image feature a black box in the middle highlighting the mouse hippocampus, including the "Dentate Gyrus" and "Cornu Ammonis" domains, and a black box on the right indicating the mouse cerebellum with complex cortical structures, including the Purkinje cell layer (PkH), molecular layer (MolH), and internal granular layer (IGrH). Figure 5(L) illustrates a visualization of the spatial domain identification results for STAGATE, STMGraph, and HiSTaR, featuring a black box in the middle highlighting the hippocampal region where HiSTaR indicates potential for effectively identifying the "Dentate Gyrus" and "Cornu Ammonis" structures, while STAGATE and STMGraph tend to treat this area as a whole region, undifferentiated domain. Additionally, a black box on the right highlights the cerebellar region, where HiSTaR suggests clearer edges of PKH MolH and IGrH, and minimal scattered spots in identifying complex cortical structures, contrasting with STAGATE and STMGraph, which display numerous scattered spots and less effective identification.

Evaluating HiSTaR's Hierarchical Structure via Ablation Experiments

HiSTaR relies on two key components: a multi-hop graph convolutional network (multi-hop GCN) and a cross-level similarity loss (L_{sim}). The multi-hop GCN integrates features from neighboring spots in tissue sections, capturing spatial relationships across multiple scales. The cross-level similarity loss ensures consistency between initial and enhanced spatial representations, stabilizing the model's ability to identify spatial domains. We assess variants with one, two, or three HiSTaR Blocks (N_{blocks}) and test configurations with or without multi-hop GCN andL_{sim}.

To validate the effectiveness, Table 1 shows ARI, NMI, and FMS of HiSTaR variants on the Human DLPFC slice #151673. The feature extraction process varies with the number of HiSTaR Blocks: a single block produces features (F_0, F_1), two blocks yield (F_0, F_1, F_2), and three blocks generate (F_0, F_1, F_2, F_3), reflecting the hierarchical representation captured. For instance, a single HiSTaR Block with multi-hop GCN, but without L_{sim} (due to its requirement for at least two blocks), achieves ARI score of 0.6359, indicating moderate performance. Variants with two blocks but lacking the L_{sim} struggle to generate effective latent representations, as its absence leads to inconsistencies between spatial representations (F_1, F_2), resulting in errors when clustering with the mclust package [41]. This suggests that the missing L_{sim} may hinder the model's ability to maintain coherent spatial features during the process. In contrast, the complete two-block HiSTaR model, incorporating both multi-hop GCN and L_{sim}, achieves the highest ARI score of 0.7182, suggesting optimal performance. For the three-block configuration, similar to two-block situation, the absence of L_{sim} prevents effective latent representation. On the other hand, using only L_{sim} without multi-hop GCN limits long-range dependencies, which also reduces performance. When both components are used together, the three-block HiSTaR variant performs well. However, having three spatial representations may introduce redundant information, which could slightly lower its performance compared to the two-block model. These results indicate that the combination of multi-hop GCN and L_{sim} in a two-block configuration enhances HiSTaR's ability to accurately identify spatial domains, offering an effective tool for spatial transcriptomics research.

HiSTaR demonstrates competitive computational efficiency in Spatial Domain Identification

We evaluate the computational efficiency of HiSTaR, STAGATE, STMGraph, and DeepST using the DLPFC Slice #151673 as an example, with experimental hardware comprising an AMD Ryzen 95900HX 16-core processor, an NVIDIA GeForce RTX 3060, and 16GB of memory. Table 2 presents the total number of epochs (Epoch_num), average training time per epoch (Epoch_time), and

Table 1 ARI, NMI, and FMS scores for HiSTaR variants on human DLPFC slice #151673

N_{blocks}	Multi-hop GCN	L_{sim}	ARI	NMI	FMS
1(F_0, F_1)	✓	-	0.6359	0.7106	0.7018
2(F_0, F_1, F_2)	✓	✖	Failed	Failed	Failed
2(F_0, F_1, F_2)	✖	✓	0.6518	0.7210	0.7163
2(F_0, F_1, F_2)	✓	✓	0.7182	0.7535	0.7736
3(F_0, F_1, F_2, F_3)	✓	✖	Failed	Failed	Failed
3(F_0, F_1, F_2, F_3)	✖	✓	0.6601	0.7298	0.7221
3(F_0, F_1, F_2, F_3)	✓	✓	0.6931	0.7483	0.7702

-: Cannot be applied L_{sim}. Failed: Failure to perform clustering using mclust

Table 2 Computational efficiency for STMGraph, STAGATE, DeepST, and HiSTaR on DLPFC slice #151673

	Epoch_num	Epoch_time (s)	Total_time (s)
STMGraph	1000	0.468	96.56
STAGATE	1000	0.299	44.59
DeepST	1800	0.039	273.88
HiSTaR	400	0.048	27.29

total training time (Total_time) for these methods, offering insights into their computational demands on this setup.

HiSTaR demonstrates optimal performance in terms of epoch numbers, requiring only 400 epochs to achieve a good fit, thanks to its streamlined and efficient structure. HiSTaR also excels in average training time per epoch, with an Epoch_time of 0.048 seconds, reported epochs reflect convergence points, with HiSTaR achieving optimal performance at 400 epochs, while others required more due to slower stabilization. contributing to its overall efficiency. This leads to HiSTaR's shortest total training time of 27.29 seconds, contrasting with 44.59s (STAGATE), 96.56s (STMGraph), and 273.88s (DeepST), suggesting its effectiveness across these metrics, with further performance indicators and comparisons detailed in Supplementary Materials Table S4.

Conclusion

Spatial transcriptomics represents a transformative approach in biological research, enabling the study of gene expression within the spatial context of tissues. A fundamental aspect of this field is spatial domain identification, which plays a critical role in uncovering the intricate organization of cells and their interactions within tissue architecture. However, a significant challenge persists in effectively capturing long-range dependencies among spots. Traditional methods often struggle to fully represent the complex spatial patterns inherent in spatial transcriptomics data.

To address this limitation, we introduce the Hierarchical ST variational autoencoder (HiSTaR), a potential method designed to enhance spatial domain identification. HiSTaR supports spatial domain identification by integrating multi-hop GCN and a hierarchical structure through a series of HiSTaR Blocks alongside a variational autoencoder. Multi-hop GCN connects spots beyond immediate neighbors to include long-range dependencies while the hierarchical structure blends local and global features across scales, supported by cross-level similarity loss that helps maintain semantic consistency and enrich the representation of spatial domains. HiSTaR's hierarchical feature extraction provides a window into its decision-making by aligning F_0 with raw gene expression profiles that reflect local cellular states, F_1 with spatially coordinated gene activities indicative of tissue microenvironments such as tumor boundaries or neuronal layers, and $F_2$with broader biological patterns that reveal functional tissue zones, such as immune cell infiltration gradients in cancer tissues. The variational autoencoder complements this design by providing a probabilistic framework that addresses data uncertainty and variability, thereby strengthening the comprehensive nature of latent representations.

HiSTaR demonstrates its potential in spatial domain identification across diverse datasets, including the human dorsolateral prefrontal cortex (DLPFC), where it suggests effective delineation of domains, and the Human Breast Cancer (HBC) dataset, where it reveals fine-grained tumor structures with greater visual consistency compared to STAGATE and STMGraph. Across multi-platform mouse brain datasets, HiSTaR recognizes key regions with boundaries that align reasonably well with biological references, supported by a cohesive clustering pattern, reflecting its ability to capture spatial relationships. Additionally, the PAGA trajectory analysis ofHiSTaR(6→5→WM→1→2→3→4)intheDLPFC#151673 outperforms STAGATE, though it deviates from the expected linear cortical layering (L1 through L6 to WM). This difference likely arises from transcriptional noise and biological variability in the data, as well as HiSTaR's hierarchical assumptions, which may emphasize developmental patterns like inside-out cortical neurogenesis over strict spatial progression, highlighting a need for future refinements with anatomical priors. Furthermore, HiSTaR shows promise in addressing batch-effects across multiple slices, offering a useful approach to integrating data while preserving biological distinctions, which supports its application in comprehensive studies. More downstream analysis in supplementary materials.

The practical value of HiSTaR stems from its computational efficiency, enable its use with high-resolution single-cell data, such as from Stereo-seq, and large-scale tissue sections. This efficient resource use extends HiSTaR's applicability across various platforms, including 10x Genomics Visium, Slide-seqV2, and STARmap,

catering to a range of research needs. As a result, HiSTaR provides a versatile tool for analyzing complex spatial transcriptomic datasets, facilitating insights into tissue architecture and cellular diversity.

In the future, HiSTaR can make some improvements in the automatic optimization of hyperparameters, especially in the cross-level similarity loss. At present, the performance of HiSTaR is very sensitive to the setting of cross-level similarity loss, which will affect the stability of the model. These hyperparameters were tuned using grid search, with details provided in supplementary materials Table S5. Developing an adaptive parameter learning framework might simplify the process of selecting optimal hyperparameters for different datasets. Beyond hyperparameter sensitivity, HiSTaR may face constraints with discontinuous tissue geometries, such as fragmented or irregularly shaped regions, where KNN-based adjacency could introduce artifacts; future work could incorporate adaptive graphs. Additionally, future versions of HiSTaR could benefit from integrating multimodal data, such as gene co-expression networks or morphological information from H&E images [42], to enrich model interpretability and support fine-grained identification. Building on adaptive hyperparameter learning and multimodal integration, hierarchical models like HiSTaR hold promise for future spatial omics data, such as spatial proteomics (e.g., integrating protein localization with transcriptomics) and spatial metabolomics (e.g., mapping metabolite distributions in disease contexts). This could enable comprehensive multi-omics views of tissue microenvironments, accelerating discoveries in precision medicine and systems biology. Potential improvements include creating an adaptive framework for hyperparameter learning and incorporating multimodal data to expand HiSTaR's utility in complex spatial transcriptomics scenarios. Continued validation across a wide range of datasets remains essential to ensure its effectiveness and adaptability in diverse research contexts.

Materials and methods

Data description

HiSTaR is adaptable to spatial transcriptomics data from diverse sequencing platforms, including 10x Genomics Visium, Stereo-seq, Slide-seqV2, and STARmap. Specifically, the human dorsolateral prefrontal cortex (DLPFC) dataset and the human breast cancer dataset are derived from the 10x Genomics Visium platform. The DLPFC dataset comprises 12 slices from three human donors, with a sequencing resolution of approximately 55 micrometers, and each slice includes pre-annotated regions of DLPFC and white matter (WM). The human breast cancer dataset contains 3798 spots. Additionally, Stereo-seq, Slide-seqV2, and STARmap are high-resolution single-cell sequencing platforms. The mouse olfactory bulb (MOB) datasets from Stereo-seq and Slide-seqV2 offer spatial resolutions of approximately 14 and 10 micrometers, respectively, with 191,091 and 201,391 spots, respectively. The mouse visual cortex dataset, generated by STARmap, includes 1207 spots. For batch-effects correction experiments, we utilized the mouse breast cancer dataset, divided into two sections with a total of 3818 spots, and the mouse brain dataset, also split into two sections (anterior and posterior) with a total of 6114 spots. Furthermore, we incorporated publicly available annotation map images downloaded from the Allen Brain Atlas website to support the analysis.

Data preprocessing

After obtaining the spatial coordinate information and gene expression information of spatial transcriptomics (ST) data, HiSTaR first performs logarithmic transformation to normalize the original gene expression data, and then selects highly variable genes. Finally, it uses PCA to reduce the dimension of the highly variable genes selected earlier to the input dimension of the model. The above operations are implemented using Scanpy [33] and Scikit-learn packages [43].

Graph construction for spatial transcriptomics data

The biggest difference between spatial transcriptomics and traditional transcriptomics is that spatial transcriptomics utilizes the spatial relationship between spots, which allows us to better identify specific tissues. We obtain the spatial coordinates of each spot in the spatial transcriptomics data. To obtain a network that can represent spatial relationships, we calculate the Euclidean distance between spots and select K spots as neighbors around the target spot, using KNN to construct an adjacency matrix. This adjacency matrix is represented by A, and if $A_{ij} = A_{ji} = 1$, then $Spot_i$ and $Spot_j$ are neighbors, otherwise 0. In order to save CPU Memory usage and speed up calculation efficiency, the adjacency matrix A is stored in a sparse format.

HiSTaR Block

Effective utilization of spatial information in spatial transcriptomics (ST) data helps enhance the identification of spatial domains. Within the HiSTaR framework, the HiSTaR Block, a key component of the HiSTaR Encoder, supports the integration of expression matrices with spatial data, facilitating a hierarchy of local-to-global features and latent representation learning through a collaborative structure of multi-hop graph convolution [44] and variational inference. This module consists of a multi-hop graph convolution network [44] (Multi-Hop GCN) and variational parameter generation unit, and its core design is as follows:

Multi-scale neighborhood feature extraction: Due to the differentiation laws of biological tissues and structures, the biological expression forms of spots are usually the same within a certain tissue or a certain spatial domain. Therefore, it is particularly important to make good use of these spots with the same biological expression. Traditional methods may use GCN or GAT as tools to integrate spatial information and expression information, but their "receptive fields" usually only stay in the vicinity of the target spot and ignore the information of more distant spots, which may lead to useful information being ignored, thus losing the model's ability to identify spatial domains.

Therefore, we use Multi-Hop GCN as a feature extraction tool in the HiSTaR Block, aiming to break through the local perception limitations of traditional GCN by aggregating information from multi-hop neighborhoods of nodes, thus more comprehensively modeling the global topological relationships and long-range dependencies in the graph, and capturing the complex spatial interactions between spots and the multi-scale features of gene expression patterns. Multi-Hop GCN extends the graph convolution operation to the k-hop neighborhood. Its core idea is: by iteratively or explicitly aggregating features within the k-hop neighborhood, nodes can perceive a wider range of spatial or functional correlations. To save Memory usage and avoid the amplification of noise in the k-hop adjacency matrix $\tilde{A}^k\,(k=1,2,\ldots,k)$ caused by the calculation of the adjacency matrix powers, resulting in numerical instability, we use an implicit Multi-Hop GCN to capture spatial patterns. Assuming that the input data of our HiSTaR Block is a gene expression matrix $X \in \mathbb{R}^{N\times d}$, there are N spots, each spot has d-dimensional features, and the adjacency matrix $A \in \mathbb{R}^{N\times N}$, if $Spot_i$ is connected to $Spot_j$ then $A_{ij}=A_{ji}=1$, otherwise 0. HiSTaR Block will first normalize the adjacency matrix:

$$\overset{=}{A}\, D^{-1/2}AD^{-1/2}$$

Where D is the degree matrix, $D_{ii}=\sum_j A_{ij}$. Then we will perform the first graph convolution operation, which is a single-hop graph convolution at this time. The single-hop graph convolution operation updates the features by aggregating the first-order neighbors (directly connected nodes) of the node:

$$F^{(1)}=\sigma\left(\overset{X}{A}\,W^{(1)}\right)$$

Where $W^{(1)}\in\mathbb{R}^{d\times h}$ is the trainable weight matrix, and h is the dimension of the hidden layer. $\sigma(\cdot)$ is the activation function (such as $ReLU$). Then, a multi-hop operation will be performed, where L layers of GCN are stacked, and the output of each layer is used as the input of the next layer to achieve implicit multi-hop information aggregation. The output of the L layer is:

$$F^{(l)}=\sigma\left(\overset{F^{(l-1)}}{A}\,W^{(l)}\right), l=1,2,\ldots,L$$

Where $F^{(0)}=X$ is the initial gene expression matrix.

In the HiSTaR Block, the $L=2$ layer design of the implicit multi-hop GCN: The first layer GCN aggregates the gene expression of adjacent cells (1-hop), identifying local co-expression modules. The second layer GCN captures the long-distance associations across tissue regions (2-hop), modeling cell groups that are functionally similar but spatially dispersed. Unlike stacking standard GCN layers, which propagate messages sequentially and may dilute distant signals, HiSTaR's implicit multi-hop GCN employs a SoftMax-weighted aggregation that directly computes attention over extended neighborhoods via a powered adjacency matrix ($\tilde{A}^k$ for k hops), enabling efficient capture of long-range dependencies in a single layer. To highlight the importance of one-hop neighbors to the target spot, we set θ as the weighted fusion weight of different hop adjacency matrices, use residual connection plus the original input matrix, and apply an activation function. The formula for the 2-hop graph convolution network used by HiSTaR Block is as follows:

$$F_{\mathrm{multi-hop}}=\sigma\left(\sum_{k=1}^{K}\mathrm{softmax}(\theta)_k\cdot\tilde{A}^k\,(XW)+X\right)$$

Where $F_{\mathrm{multi-hop}}\in\mathbb{R}^{N\times d_1}$ is the feature output of the multi-hop graph convolution in the HiSTaR Block, and $\tilde{A}^k$ represents the normalized weight of the $k-hop$ path from node i to j. The implicit multi-hop graph convolution network of the HiSTaR Block is superior to traditional graph convolution networks in terms of computational efficiency, hierarchical learning, and anti-overfitting, and is particularly suitable for multi-scale correlation modeling tasks such as spatial transcriptomics.

Variational parameter generation unit is responsible for mapping the features extracted by multi-hop GCN into the probability distribution parameters of latent variables (mean μ and log variance $\log\sigma^2$). Its design goal is to use the preliminary features $F_{\mathrm{multi-hop}}$ generated by multi-hop graph convolution to enhance the model's robustness to data heterogeneity and noise while retaining the multi-scale features of spatial transcriptome data. Generating the mean μ and log variance $\log\sigma^2$ of latent variables separately through the independent graph

convolutional network (GCN) branch, and the formula is as follows:

$$\mu = \sigma\left(\overset{F_{\text{multi-hop}}}{A} W_{\mu}\right)$$

$$\log\sigma^2 = \sigma\left(\overset{F_{\text{multi-hop}}}{A} W_{\sigma}\right)$$

Where W_{μ} and W_{σ} are independent trainable parameters, and $\sigma(\cdot)$ is the activation function. The latent variable F is sampled using the reparameterization trick:

$$F = \mu + \in \odot \exp\left(\frac{\log\sigma^2}{2}\right), \in \sim N(0, I)$$

In order for $F^{(l)}$ to be close to the Gaussian distribution preset by VAE in the latent space, we use hierarchical KL divergence to calculate the KL divergence of the mean and variance generated in the middle of multi-hop graph convolution for hierarchical latent variables, ($L = 2$ in HiSTaR), the formula is as follows:

$$\mathcal{L}_{\text{GCL}} = \sum_{l=1}^{L} \frac{-1}{2N} \sum_{i=1}^{N} \left(1 + \log\sigma_{l,i}^2 - \mu_{l,i}^2 - \sigma_{l,i}^2\right)$$

Where $\mu_l \in \mathbb{R}^{N \times d_l}$ is the mean of the latent variables of the l layer. $\log\sigma_l^2 \in \mathbb{R}^{N \times d_l}$ is the logarithmic variance of the latent variables of the l layer, and l is the number of layers of hierarchical latent variables. HiSTaR Block enhances the model's ability to model spatial heterogeneity and technical noise by generating means and variances separately.

HiSTaR structure

In order to allow the model to recover complete information from partial observation data, avoid overfitting to local noise or redundant features, and make the model rely more on the topological relationship of the graph adjacency matrix for inference, thereby strengthening the modeling of graph structure dependence. Before ST data enters the HiSTaR Encoder for encoding, we will randomly mask the expression matrix. Specifically, the gene expression matrix $X \in \mathbb{R}^{N \times d}$ after preprocessing has N spots, and each spot has d-dimensional features. We randomly replace some of the d-dimensional features of some spots with d-dimensional learnable vectors according to a certain proportion. Therefore, the input expression matrix $X \in \mathbb{R}^{N \times d}$ is redefined as $X' \in \mathbb{R}^{N \times d}$.

HiSTaR Encoder consists of a traditional encoder and two HiSTaR Blocks in series. The encoder is composed of two fully connected layers, and the input data is the expression matrix $X' \in \mathbb{R}^{N \times d}$ after masking. The two fully connected layers in the encoder will gradually reduce the feature dimension. The output of the first fully connected layer is the feature $f_1 \in \mathbb{R}^{N \times d_1}$, and then $f_1 \in \mathbb{R}^{N \times d_1}$ is used as the input of the second fully connected layer. The second fully connected layer reduces the dimension of the expression matrix and outputs a base feature $F_0 \in \mathbb{R}^{N \times d_0} (d_0 < d_1)$, the formula is as follows:

$$f_1 = \sigma\left(BN\left(X' W_1\right)\right) \left(W_1 \in \mathbb{R}^{d \times d_1}\right)$$

$$F_0 = \sigma\left(BN\left(f_1 W_2\right)\right) \left(W_2 \in \mathbb{R}^{d_1 \times d_0}\right)$$

Where $\sigma(\cdot)$ is the activation function, and $BN(\cdot)$ is Batch Normalization.

Then we input the adjacency matrix $A \in \mathbb{R}^{N \times N}$ and the base feature $F_0 \in \mathbb{R}^{N \times d_0}$ into HiSTaR Block1, and output the initial spatial representation $F_1 \in \mathbb{R}^{N \times d_1}$. We input the initial spatial representation $F_1 \in \mathbb{R}^{N \times d_1}$ and the original adjacency matrix $A \in \mathbb{R}^{N \times N}$ into HiSTaR Block2, extract the enhanced spatial representation $F_2 \in \mathbb{R}^{N \times d_2}$. The generation processes of F_1 and F_2 do not interfere with each other, and the parameters are independent, in order to achieve the purpose of decoupling between different levels of HiSTaR Blocks. However, the deep latent variable F_2 usually encodes more abstract features, while the shallow variable F_1 retains more local information. Without constraints, it may lead to semantic fragmentation between levels and impair the ability to identify spatial domains. Here, we innovatively proposed cross-level similarity to constrain the latent variables of two HiSTaR Blocks to maintain geometric consistency. We calculate the cosine similarity between F_1 and F_2 and make F_1 and F_2 locally similar. The similarity loss is defined as:

$$\mathcal{L}_{\text{sim}} = -\frac{1}{N} \sum_{i=1}^{N} \cos\left(F_1, F_2\right)$$

Where $\cos(\cdot, \cdot)$ represents the cosine similarity [45]:

$$\cos(a, b) = \frac{a \cdot b}{a_2 b_2}$$

This way, we can both maintain the specificity of potential representations F_1 and F_2 and ensure their consistency in the latent space, improving the identification ability in the spatial domain. Then, we concatenate $F_0 \in \mathbb{R}^{N \times d_0}$, $F_1 \in \mathbb{R}^{N \times d_1}$ and $F_2 \in \mathbb{R}^{N \times d_2}$ to form the final low-dimensional latent representation $Z \in \mathbb{R}^{N \times (d_0 + d_1 + d_2)}$.

HiSTaR Decoder consists of an encoder and an inner product decoder. The decoder uses a traditional GCN

that can capture more spatial information to decode the latent representation $Z \in \mathbb{R}^{N\times(d_0+d_1+d_2)}$and generate the reconstruction expression matrix $X_{recon} \in \mathbb{R}^{N\times d}$. The objective function of HiSTaR maximizes the similarity between the input gene and the reconstructed expression, which is measured using the mean square error (MSE) loss function [46], as follows:

$$\mathcal{L}_{recon} = \frac{1}{N}\sum_{i=1}^{N} X_i - X_{\text{recon},i}{}_2^2$$

Where $X \in \mathbb{R}^{N\times d}$ is the input feature matrix, $X_{\text{recon}} \in \mathbb{R}^{N\times d}$ is the feature matrix reconstructed by the decoder, and $\cdot_2$ is the $L2$ norm.

HiSTaR uses the Inner Product Decoder to calculate the similarity between the latent representations of nodes to reconstruct the adjacency matrix $\widehat{A} \in \mathbb{R}^{N\times N}$, the formula is as follows:

$$\widehat{A}_{ij} = \sigma\left({Z_i}^{\mathrm{T}} Z_j\right)$$

Where $\sigma(\cdot)$ is the activation function, and Z_i and $Z_j$is the latent representations of $Spot_i$ and $Spot_j$. HiSTaR use the binary cross-entropy loss (BCE) [47] to calculate the reconstruction loss of the adjacency matrix, the formula is as follows:

$$\mathcal{L}_{adj} = \mathcal{BCE}\left(\widehat{A}, A\right)$$

The total loss function L of HiSTaR is as follows:

$$\begin{aligned}\mathcal{L} =& \lambda_{sim}\times\mathcal{L}_{sim} + \lambda_{recon}\times\mathcal{L}_{recon} \\ &+ \lambda_{DEC}\times\mathcal{L}_{D\mathcal{E}C} + \lambda_{GCN}\times\mathcal{L}_{GCN} + \mathcal{L}_{adj}\end{aligned}$$

Where λ_{sim}, λ_{recon}, $\lambda_{DEC}$and λ_{GCN} are hyperparameters that balance the contributions of the respective loss terms. In actual experiments, we set $\lambda_{recon} = 10$, $\lambda_{DEC} = 1$, $\lambda_{GCN} = 0.1$. Because of the model's sensitivity to hyperparameter λ_{sim}, we shall discuss the settings for λ_{sim} in the conclusion and detail its settings across different datasets in the supplementary materials Table S6.

Clustering and visualization

After extracting potential features with HiSTaR, HiSTaR uses mclust in the R package to assign clustering labels to the potential representation of each spot. Gene expression in spatial transcriptomics data shows non-uniform distribution in space and may contain irregular cluster morphologies (e.g., ellipsoidal, band-like distribution). The covariance flexibility of mclust can model such structures. Therefore, many spatial transcriptomics deep learning methods use mclust as a clustering tool. In HiSTaR, a dataset of determined categories is used in conjunction with mclust as a clustering tool. HiSTaR employs an unsupervised deep embedded clustering (DEC) [48] method to iteratively group the spots into different clusters. And we use UMAP as a tool for clustering visualization. UMAP is a nonlinear dimensionality reduction method that can effectively capture nonlinear relationships (such as manifold structures) in high-dimensional data, while linear methods (such as PCA) can only handle linear relationships. For high-dimensional complex data such as gene expression, single-cell data, and image features, UMAP can more accurately reflect the inherent structure of the data.

Analysis of differentially expressed genes (DEGs)

For spatial transcriptome data, we used the Wilcoxon rank-sum test [49] for differential analysis in the spatial domain. We calculated gene expression differences in each spatial domain using Scanpy, and set a multiple hypothesis testing correction threshold $FDR < 1\times10-5$ to screen for significant genes. We required the absolute value of the logarithmic fold change in gene expression $|log2FC| \geq 1$, and when the number of candidate genes under the strict threshold was less than the preset number, we relaxed the significance condition to $FDR < 5\%$, while maintaining the expression change threshold unchanged.

Benchmarking

After HiSTaR clustering is completed and the clustering labels of each spot are generated, we will use the clustering labels generated by HiSTaR and Ground Truth (GT) to calculate three metrics, namely Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Score (FMS). If there is no GT, we use the silhouette coefficient (SC) as the evaluation metric of each model. The evaluation formula of each metric is as follows:

$$\text{ARI}(Y,Z) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{I}(y_i=y_j)\,\mathbb{I}(z_i=z_j) - \frac{1}{\binom{N}{2}}\left[\sum_{i=1}^{N}\binom{c_i}{2}\sum_{i=1}^{N}\binom{d_i}{2}\right]}{\frac{1}{2}\left[\sum_{i=1}^{N}\binom{c_i}{2}+\sum_{i=1}^{N}\binom{d_i}{2}\right] - \frac{1}{\binom{N}{2}}\left[\sum_{i=1}^{N}\binom{c_i}{2}\sum_{i=1}^{N}\binom{d_i}{2}\right]}$$

Where $Y = \{y_1,\ldots,y_N\}$ is the GT set, $Z = \{z_1,\ldots,z_N\}$ is the model prediction label set, $c_i = \sum_{j=1}^{N}\mathbb{I}(y_j = y_i)$ is the number of spots in GT that are in the same cluster as

spot i, $d_i = \sum_{j=1}^{N} \mathbb{I}(z_j = z_i)$ is the number of spots in the model prediction label set that are in the same cluster as spot i, $\mathbb{I}(\cdot)$is the metric function (takes 1 if the condition is true, otherwise it takes 0), and $\binom{n}{2} = \frac{n(n-1)}{2}$ is the operation rule of combination numbers.

$$\mathrm{NMI}(Y,Z) = \frac{2 \cdot \sum_{k=1}^{K_Y} \sum_{l=1}^{K_Z} \frac{n_{kl}}{N} \log\left(\frac{n_{kl} \cdot N}{n_k \cdot m_l}\right)}{-\sum_{k=1}^{K_Y} \frac{n_k}{N} \log\frac{n_k}{N} - \sum_{l=1}^{K_Z} \frac{m_l}{N} \log\frac{m_l}{N}}$$

Where $n_{kl} = \sum_{i=1}^{N} \mathbb{I}(y_i = k)\mathbb{I}(z_i = l)$ is the number of spots in the intersection of the true cluster k and the predicted cluster l, $n_k = \sum_{i=1}^{N} \mathbb{I}(y_i = k)$ is the number of spots in the true cluster k, and $m_l = \sum_{i=1}^{N} \mathbb{I}(z_i = l)$ is the number of spots in the predicted cluster$l$.

$$\mathrm{FMS}(Y,Z) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{I}(y_i = y_j)\mathbb{I}(z_i = z_j)}{\sqrt{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{I}(z_i = z_j)\right) \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{I}(y_i = y_j)\right)}}$$

Where $y_i \in \{1, \ldots, K_Y\}$ is the GT of spot i, $z_i \in \{1, \ldots, K_Z\}$ is the model prediction label of spot i.

$$\mathrm{SC}(Z) = \frac{1}{N} \sum_{i=1}^{N} \frac{\min\limits_{l \neq z_i} \left(\frac{1}{m_l} \sum_{j=1}^{N} (z_j = l) d(\mathrm{h}_i, \mathrm{h}_j) \right) - \frac{1}{m_{z_i} - 1} \sum_{j=1}^{N} (z_j = z_i) d(\mathrm{h}_i, \mathrm{h}_j)}{\max\left(\frac{1}{m_{z_i} - 1} \sum_{j=1}^{N} (z_j = z_i) d(\mathrm{h}_i, \mathrm{h}_j), \min\limits_{l \neq z_i} \left(\frac{1}{m_l} \sum_{j=1}^{N} (z_j = l) d(\mathrm{h}_i, \mathrm{h}_j) \right) \right)}$$

Where $\mathrm{h}_i \in \mathbb{R}^d$is the feature vector of spot i, $m_l = \sum_{i=1}^{N} \mathbb{I}(z_i = l)$ is the number of spots predicted to be in cluster l, $d(\mathrm{x}_i, \mathrm{x}_j)$ is the Euclidean distance between spot i and spot j.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-07404-3.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
J.Y. and J.Y. contributed equally. J.Y. conceived, designed and performed computational modeling, validations and benchmarks. J.Y analyzed data and validated the results. Z.Y. conceived and supervised the project. P.X. validated and analyzed the results. W.L. and J.Y. drafted the manuscript and revised the manuscript.

Funding
This work was supported in part by the National Natural Science Foundation of China (No. 62573143 and 62072128), the Natural Science Foundation of Guangdong Province of China (No. 2023A1515011401), the National key R and D Program of China (Grant 2019YFA0706338402).

Data availability
The dataset is published in a public repository (https://zenodo.org/records/15599070). Tutorial of HiSTaR can be viewed here (https://histar-tutorials.readthedocs.io/en/latest/index.html). Supplementary material is provided.

Code availability
The scripts used in this study are developed by the Python programming language and have been deposited in a public Github repository (https://github.com/Anglejuebi/HiSTaR).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 21 August 2025 / Accepted: 29 October 2025
Published online: 23 December 2025

References

1. Asp M, Bergenstråhle J, Lundeberg J. Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays. 2020;42:1900221.
2. Palla G, Fischer DS, Regev A, Theis FJ. Spatial components of molecular tissue biology. Nat Biotechnol. 2022;40:308–18.
3. Ståhl PL, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.
4. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596:211–20.
5. Xu C, et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. Nucleic Acids Res. 2022;50:e131–131.
6. De Meo P, Ferrara E, Fiumara G, Provetti A. Generalized louvain method for community detection in large networks. IEEE; 2011. In: 2011 11th international conference on intelligent systems design and applications.
7. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory. 2008;P10008.
8. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:1–12.
9. Anuar SHH, et al. Comparison between Louvain and Leiden algorithm for network structure: a review. IOP Publishing; 2021. In: J Phys Conf Ser.
10. Zhang M, Zhang W, Ma X. ST-SCSR: identifying spatial domains in spatial transcriptomics data via structure correlation and self-representation. Briefings In Bioinf. 2024;25:bbae437.
11. Xu H, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Med. 2024;16:12.
12. Pinheiro Cinelli L, Araújo Marins M, Barros da Silva EA, Lima Netto S. Variational autoencoder. In: Variational methods for machine learning with applications to deep networks. Springer; 2021.
13. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016.

14. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun. 2022;13:1739.
15. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903. 2017.
16. Lin L, et al. Stmgraph: spatial-context-aware of transcriptomes via a dual-remasked dynamic graph attention model. Briefings In Bioinf. 2025;26:bbae685.
17. Brody S, Alon U, Yahav E. How attentive are graph attention networks? arXiv preprint arXiv:210514491. 2021.
18. Ji AL, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. Cell. 2020;182:497–514. e422.
19. Rodriques SG, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363:1463–67.
20. Chen A, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. Cell. 2022;185:1777–92. e1721.
21. Chen A, et al. Large field of view-spatially resolved transcriptomics at nanoscale resolution. BioRxiv. 2021.
22. Wang X, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361:eaat5691.
23. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22:627–44.
24. Zhang W, Li S, Wang Y, Zhang W. Eemsnet: Eagle-eye multi-scale supervised network for cardiac segmentation. Biomed Signal Process and Control. 2024;96:106638.
25. Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4:1883.
26. Maynard KR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.
27. Steinley D. Properties of the hubert-arable adjusted rand index. Psychological Methods. 2004;9:386.
28. Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans on Neural Networks. 2009;20:189–201.
29. Nemec A, Brinkhurst R. The Fowlkes–mallows statistic and the comparison of two independently determined dendrograms. Can J Fish and Aquat Sci. 1988;45:971–75.
30. Becht E, et al. Dimensionality reduction for visualizing single-cell data using umap. Nat Biotechnol. 2019;37:38–44.
31. Wolf FA, et al. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019;20:1–9.
32. Buache E, et al. Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice. Oncogene. 2011;30:3261–73.
33. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
34. Yuan J, Yu J, Yi Q, Ye Z, Xu P, Liu W. SpateCV: cross-modality alignment regularization of cell types improves spatial gene imputation for spatial transcriptomics. J Transl Med. 2025;23:1188.
35. Mudd TW Jr, Lu C, Klement JD, Liu K. MS4A1 expression and function in T cells in the colorectal cancer tumor microenvironment. Cellular Immunol. 2021;360:104260.
36. Rebbeck CA, et al. Gene expression signatures of individual ductal carcinoma in situ lesions identify processes and biomarkers associated with progression towards invasive ductal carcinoma. Nat Commun. 2022;13:3399.
37. Lein ES, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.
38. Dinh DT, Fujinami T, Huynh VN. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In: Knowledge and systems sciences: Proceedings of the 20th International Symposium, KSS 2019; 2019 Nov 29–Dec 1; Da Nang, Vietnam. Cham: Springer; 2019.
39. Zeira R, Land M, Strzalkowski A, Raphael BJ. Alignment and integration of spatial transcriptomics data. Nat Methods. 2022;19:567–75.
40. Korsunsky I, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
41. Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R J. 2016;8:289.
42. Zhang W, Ding X, Liu Y, Qiao B. Slide deep reinforcement learning networks: application for left ventricle segmentation. Pattern Recognit. 2023;141:109667.
43. Pedregosa F, et al. Scikit-learn: machine learning in Python. The J Mach Learn Res. 2011;12:2825–30.
44. Xue H, Sun X-K, Sun W-X. Multi-hop hierarchical graph neural networks. IEEE; 2020. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp).
45. Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. Inf Sci. 2015;307:39–52.
46. Hodson TO. Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Devel. 2022;1–10.
47. Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. Int J Adv Trends Comput Sci Eng. 2020;9.
48. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. PMLR; 2016. In: International conference on machine learning.
49. Cuzick J. A Wilcoxon-type test for trend. Stat Med. 1985;4:87–90.

Publisher's Note