



Inferring single-cell resolution spatial gene expression via fusing spot-based spatial transcriptomics, location, and histology using GCN

Shuailin Xue ^{1,†}, Fangfang Zhu^{2,†}, Jinyu Chen³, Wenwen Min ^{1,*}

¹School of Information Science and Engineering, Yunnan University, 650500 Yunnan, China

²School of Health and Nursing, Yunnan Open University, 650599 Kunming, China

³School of Mathematics, Statistics and Mechanics, Beijing University of Technology, 100124 Beijing, China

*Corresponding author. School of Information Science and Engineering, Yunnan University, East Outer Ring Road, Chenggong District, Kunming 650500, China.

E-mail: minwenwen@ynu.edu.cn

[†]Shuailin Xue and Fangfang Zhu contributed equally to this work.

Abstract

Spatial transcriptomics (ST) technology allows for the detection of cellular transcriptome information while preserving the spatial location of cells. This capability enables researchers to better understand the cellular heterogeneity, spatial organization, and functional interactions in complex biological systems. However, current technological methods are limited by low resolution, which reduces the accuracy of gene expression levels. Here, we propose scstGCN, a multimodal information fusion method based on Vision Transformer and Graph Convolutional Network that integrates histological images, spot-based ST data and spatial location information to infer super-resolution gene expression profiles at single-cell level. We evaluated the accuracy of the super-resolution gene expression profiles generated on diverse tissue ST datasets with disease and healthy by scstGCN along with their performance in identifying spatial patterns, conducting functional enrichment analysis, and tissue annotation. The results show that scstGCN can predict super-resolution gene expression accurately and aid researchers in discovering biologically meaningful differentially expressed genes and pathways. Additionally, scstGCN can segment and annotate tissues at a finer granularity, with results demonstrating strong consistency with coarse manual annotations. Our source code and all used datasets are available at <https://github.com/wenwenmin/scstGCN> and <https://zenodo.org/records/12800375>.

Keywords: spatial transcriptomics; single-cell resolution; enhancement; histology image; Graph Convolutional Network

Introduction

In biology, the spatiotemporal specificity of gene expression dictates that integrating cellular spatial location information is essential for a better understanding of the specific functions of cells within tissues [1]. Neglecting the spatial information of cells may lead to inadequate understanding of gene expression patterns and spatial distribution. However, the spatial location information of tissue cells cannot be retained in single-cell sequencing experiments due to technical limitations [2, 3]. With the innovation of spatial transcriptomics (ST) technology, it has become possible to detect the quantity of gene transcripts within tissues while retaining spatial location information [4]. Many ST techniques have been developed to reveal the spatial heterogeneity of cells and genes and their distribution patterns in tissue space [5]. Therefore, ST technologies have been used in research on the development of tissues and organs [6], neurology [7], and tumor heterogeneity [8].

At present, the main ST techniques can be divided into two categories: (1) imaging-based techniques, including *in situ* hybridization and *in situ* sequencing [9]. *In situ* sequencing-based approach is to reverse-transcribe RNA in a cell by targeting probes, followed

by sequencing using a DNA ligase-based sequencing approach, such as FISSEQ [9] and STARmap [10]. *In situ* hybridization is another imaging-based method that uses fluorescently labeled nucleic acid probes to determine the spatial location and abundance of DNA and RNA in tissues and cells, typically MerFISH [11, 12] and seqFISH [13]. (2) next-generation sequencing (NGS)-based approaches, works by fixing oligonucleotide sequences carrying spatial barcodes to a chip at a specific resolution, tissue sections were then affixed to the chip for cell lysis and RNA construction Library sequencing [14], such as ST and 10X Genomics Visium. However, each of these technologies has its drawbacks. Imaging-based technologies can provide higher resolution and sensitivity, but have lower throughput, a large number of transcriptomes can not be detected. In principle, NGS-based technologies can detect a complete tissue atlas and capture all transcriptome information, but it is limited by lower spatial resolution to more accurately study detailed gene expression patterns. 10X Visium is currently the most widely used ST technology. However, on the 10X Visium platform, the center-to-center distance between spots is 100 μm and the spots with a diameter of 55 μm may contain 5 to 30 cells. Even on the older ST platform, which has 100 μm spot diameter

Received: August 1, 2024. Revised: October 13, 2024. Accepted: November 21, 2024

© The Author(s) 2024. Published by Oxford University Press.

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.

with 200 μm center-to-center distance, the number of cells within a spot may up to 200. Therefore, although the availability of many ST platforms, none of them provide a comprehensive solution that balances resolution and sequencing depth.

Recently, some tools have been developed to address the various shortcomings of ST technology using deep learning-based methods. For example, some spatial clustering methods have been developed to help identify spatial domains, such as AVGN [15], STAGATE [16], and STMask [17], using graph neural networks as the main framework to accurately capture spatial domain information. In addition, a few ST data imputation techniques, like SpatialScope [18] and SpaDiT [19], have been designed to improve the low sensitivity of ST data. Furthermore, spot-level gene expression prediction methods, such as BLEEP [20] and mclS-TExp [21], aim to predict spot-based spatial gene expression from histology images to eliminate the cost-prohibitive of ST field.

In ST technology studies, enhancing the resolution of spatial gene expression is a core task. Low spatial resolution and incomplete tissue coverage may hinder researchers from identifying gene expression patterns and deeply characterizing intricate tissue architectures. Ideal ST data should have single-cell resolution and cover the entire tissue surface. Enhancing the resolution of spot-based ST data to the cellular level can effectively solve the above problems and help discover more biologically significant pathways.

Several methods have been developed to enhance the resolution of spatial gene expression by imputing the gaps between spots, dividing spots into sub-spots, or mapping the entire tissue into a collection of superpixels representing single cells. For example, stEnTrans [22] establishes self-supervised tasks to predict gene expression in gaps between spots using the Transformer encoder. STAGE [23] utilizes a spatial location-supervised auto-encoder to predict high-density gene expression. BayesSpace [24] uses Bayesian modeling to estimate gene expression data at the subspot level without relying on histological images. However, none of these methods have achieved comprehensive prediction of single-cell resolution gene expression across the entire tissue. XFuse [25] integrates ST data and histology images using a deep generative model to infer super-resolution gene expression profiles. TESLA [26] generates high-resolution gene expression profiles based on Euclidean distance metric, which considers the similarity in physical locations and histological image features between superpixels and measured spots. However, they have not deeply extracted the intrinsic features of histological images and inefficiently utilized existing data. iStar [27] employs a multi-layer Transformer architecture to extract features from histological images, aiming to predict super-resolution gene expression by capturing both local patterns and global spatial relationships, but it does not consider for the strong correlations between neighboring cells clusters, failing to capture the complex relationships between adjacent superpixels. Additionally, gene expression correlates with spatial location, a factor overlooked by all methods predicting super-resolution gene expression.

Here, we developed scstGCN to robustly predict super-resolution closed to single-cell spatial gene expression with the ViT module and the multimodal-features-supervised Graph Convolutional Network (GCN) module by integrating histological image features, spatial positional information, and spot-based ST data. Compared to existing inferring super-resolution methods, scstGCN integrates histological image features and spatial location information, leveraging GCN to effectively capture complex relationships among adjacent superpixels. Comprehensive benchmarking on ST data generated on multiple

diverse disease and healthy tissues from different platforms demonstrates the exceptional super-resolution gene expression prediction capability of scstGCN. Moreover, scstGCN has exhibited its capability in enhancing spatial patterns of significant genes, discovering more biologically meaningful pathways and high-resolution annotation of tissue architecture.

Materials and methods

Dataset description and preprocessing

To precisely evaluate the performance of scstGCN, we used multiple Xenium datasets including human breast cancer (HBC), human pancreas (HP), and hHeart Non-diseased (HN) generated by 10x Genomics. Xenium dataset contains subcellular-based ST data, along with spatial positional information of the subcellular locations. We preprocessed Xenium datasets to align with our experimental objectives. To obtain ground truth, we constructed a rectangle grid of superpixels and then partitioned the cell-level gene expression into this rectangle grid based on the overlap area between cells and the grids:

$$p_{mn} = \sum_{k \in \Gamma_{mn}} \frac{A_{kmn}}{A_k} c_k, \quad (1)$$

where p_{mn} and c_k represent the gene expression at superpixel P_{mn} and cell C_k , respectively; A_k and A_{kmn} represent area of cell C_k and its overlapping area with the superpixel P_{mn} , respectively; Γ_{mn} represents the set of all cells that overlap with the superpixel P_{mn} . Next, pseudo-Visium data was obtained by segmenting the gene expression of Xenium dataset into a series of spots based on the spot size and center-to-center distance of the Visium:

$$s_i = \sum_{k \in \Gamma_i} c_k, \quad (2)$$

where s_i represents the gene expression of pseudo-spot S_i , and Γ_i represents the set of all cells covered by pseudo-spot S_i . The size of each superpixel was set to $8 \times 8 \mu\text{m}^2$, which is approximately equivalent to the size of a single-cell. The spot diameter was set to 55 μm , and the center-to-center distance was set to 100 μm . The location information of the pseudo-spots are evenly covered on the entire detection tissue based on the characteristics of Visium data.

Visium HD technology is 10X Genomics' latest high-resolution spatial transcriptome high-throughput sequencing technology. It enhances the resolution to the single-cell level while maintaining the advantages of high gene throughput. Unlike Xenium data, it can detect more genes, and the sequencing unit is a square area without gaps (such as $8 \times 8 \mu\text{m}$). Therefore, processing Visium HD data only requires simulating pseudo-Visium data. We need to segment the super-resolution gene expression of Visium HD data into a series of spots based on the spot size and center-to-center distance of the Visium to obtain the pseudo-Visium data. In simple terms, the gene expression of the sequencing unit set covered by the spot is aggregated as the gene expression of the spot:

$$s_j = \sum_{(m,n) \in \Gamma_j} p_{mn}, \quad (3)$$

$$\Gamma_j = \{(m, n) \mid d(S_j, P_{mn}) \leq r\},$$

where p_{mn} and s_j represent the gene expression at sequencing unit P_{mn} and pseudo-spot S_j , respectively; $d(S_j, P_{mn})$ represent the

Table 1. Summary of ST datasets used in this study. Among them, MBHD, HBCHD, HP, HN, MBC, and MBS datasets were also collected from the 10x Genomics portal.

Datasets	Size/Radius	Sections	Genes	Platform
MBHD	single-cell	1	17 797	Visium HD
HBCHD	single-cell	1	18 085	Visium HD
HP	single-cell	1	377	Xenium
HN	single-cell	1	377	Xenium
MBC	55um	1	19 465	10X Visium
MBS	55um	1	32 285	10X Visium
HBC [28]	single-cell	3	288-313	Xenium
HER2ST [29]	100um	32	10 530	ST
DLPFC [30]	55um	12	33 538	10X Visium

center distance between pseudo-spot S_j and sequencing unit P_{mn} ; r represent the radius of the Visium data. The location information of the pseudo-spots are evenly covered on the entire detection tissue based on the characteristics of Visium data.

Additionally, we also evaluated scstGCN numerically on the human dorsolateral prefrontal cortex tissue (DLPFC) conducted from 10X Visium platform. Since the resolution of DLPFC data is at the spot-level, and the resolution of our predicted gene expression is close to the cell-level, we merged the predicted single-cell resolution gene expression into the spot-level gene expression based on the location information of all spots in the original DLPFC data. This process is basically the same as the process of generating pseudo-Visium data from Visium HD. The difference is that the location information of pseudo-spot S_j comes directly from the original data, spot S_j^{true} . Flow of preprocessing for all single-cell resolution datasets can be inferred in [Supplementary Figure S1](#).

For all Xenium datasets, we predicted all genes in each dataset (313 genes in HBC_S1R1 and HBC_S1R2, 288 genes in HBC_S2, 377 genes in HP and HN). For all spot-based ST datasets and Visium HD datasets, we selected the top 1000 highly variable genes as experimental samples. Other details of description for all datasets can be found in [Table 1](#) and [Supplementary Table S1](#).

Overview of scstGCN

scstGCN is specifically designed to infer super-resolution gene expression by integrating histological image features, spatial position information, and spot-based ST data. We first extract multimodal feature map, and then use the GCN module to predict single-cell resolution gene expression from multimodal feature map by a weakly supervised framework ([Fig. 1](#)).

To obtain histological feature map, scstGCN first divides the histological image into a series of subimages, and then extracts histological image features using the ViT module. The positional feature map is constructed based on the two-dimensional coordinates of the superpixels, while the Red-Green-Blue (RGB) feature map is obtained through simple downsampling. Next, multimodal feature map is obtained by stacking the histological feature map, positional feature map, and RGB feature map ([Fig. 1](#)). Generally, gene expression is largely influenced by complex interactions between adjacent cells. To model this, scstGCN adaptively represents the intricate correlations between neighboring superpixels using the GCN modules, thereby effectively establishing the relationship between multimodal feature map and super-resolution gene expression ([Fig. 1](#)). We adopt a weakly supervised learning framework to train scstGCN because the model outputs are at the superpixel-level while the training data are at the spot-level.

Specifically, we model the sum of gene expression of the superpixels covered by spots and the gene expression of spots.

Extracting multimodal feature map based ViT module

A substantial body of study has confirmed the strong correlation between histology images and gene expression patterns [31, 32]. ViT has demonstrated remarkable success in image analysis. Therefore, we aim to leverage ViT to extract comprehensive features from histology images. Due to varying pixel sizes in histology images from different datasets, we need to resize the histology images, defaulting to scaling the pixel size to $0.5 \mu\text{m}$. This process ensures that 16×16 patches approximately correspond to the size of a single cell. When the scaled image is not divisible by 224, padding operation is required for the image so that its height and width are both divisible by 224.

Next, we partition the entire histological image into subimages of size 224×224 to serve as inputs to the ViT module to extract histology image features in tiles. Let M and N denote the height and width of the histology image, then it can be expressed as $X = [X_{mn}]_{m=1, n=1}^{M/224, N/224}$, where $X \in \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^3$ and each subimage $X_{mn} \in \mathbb{R}^{224 \times 224 \times 3}$.

Following the traditional principles of ViT [33], we reshape each subimage into a series of 16×16 patches, and then flatten the image into sequence data for compatibility with the Transformer architecture:

$$X'_{mn} = X_{mn} * W + b, \quad (4)$$

$$\hat{X}_{mn} = \text{Flatten}(X'_{mn}), \quad (5)$$

where X'_{mn} and \hat{X}_{mn} have shapes $(16 \times 16 \times 3, 224/16, 224/16)$ and $(196, 768)$; W and b represent the filter parameters and biases of the convolutional structures.

Next, we briefly describe the basic components of the Transformer block, including Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP).

In the MSA module, the inputs $U \in \mathbb{R}^{n \times d}$ are linearly transformed into three parts, namely $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{n \times d_k}$, and $V \in \mathbb{R}^{n \times d_v}$ where n is the sequence length, d , d_k , and d_v are the dimensions of inputs, keys (queries) and values, respectively. Linear transformation can be described as follows:

$$[Q_i, K_i, V_i] = UW_i + b_i, i = 1, 2, \dots, h, \quad (6)$$

where h denotes the numbers of self-attention operations in MSA, W_i and b_i represent parameters and biases of the linear transformation. The scaled dot-product attention is applied on Q , K , V and then concat all head as the output:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i), \quad (8)$$

$$\text{Mutihead}(U) = \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \times W_{\text{msa}}, \quad (9)$$

where $\frac{QK^T}{\sqrt{d_k}}$ represents the similarity between Query and W_{msa} is the weight matrix for multiple heads.

There are two linear transformations separated by an activation function in the MLP module for feature transformation and

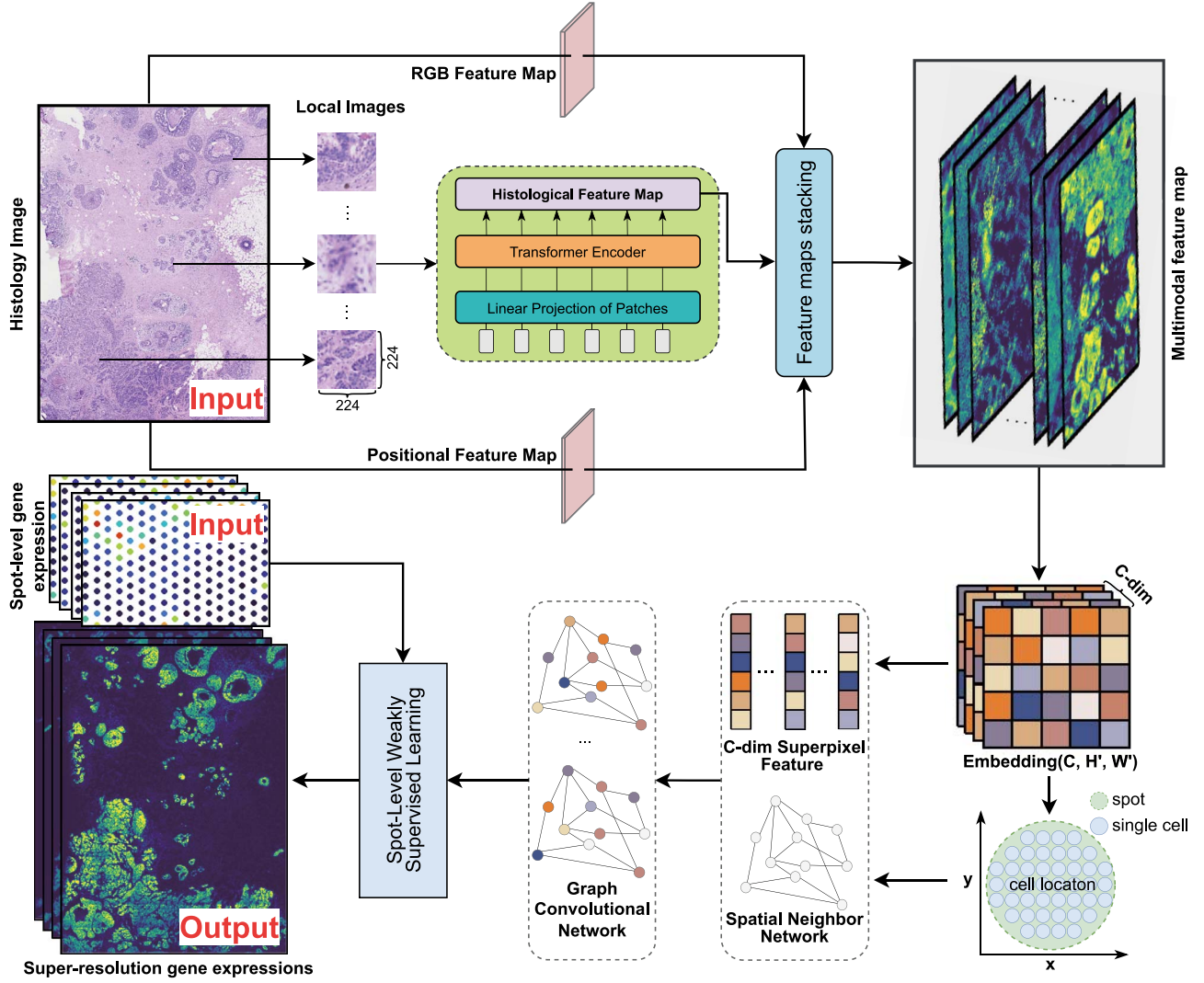


Figure 1. The architecture summary of scstGCN. First, the histological image is divided into sub-images of size 224 to fit the input of the ViT module, which can extract histological feature maps by converting image data into sequence data. Second, the RGB feature map can be obtained from the histological image through downsampling and then the location feature map is calculated based on the two-dimensional spatial coordinates of the superpixels. These features are stacked to obtain the multimodal feature map. Next, The GCN module is utilized to further capture the complex relationships between adjacent cells. Finally, based on a weakly supervised GCN framework, the original gene expression data are employed as pseudo-labels to predict super-resolution gene expression.

non-linearity:

$$\text{MLP}(X) = \sigma(XW_1 + b_1)W_2 + b_2, \quad (10)$$

where W_1 , b_1 , W_2 , and b_2 are the weight and bias term of the first and second fully-connected layer, respectively; σ represents the activation function such as GELU [34].

We use the function $f(\cdot)$ to represent the above process of feature extraction using the ViT module. ViT maps each 16×16 patch within each subimage X_{mn} into a feature vector of length 1024, and then the histological image features are obtained by applying the inverse operation of Flatten:

$$Y'_{mn} = f(X_{mn}), \quad (11)$$

$$Y_{mn} = \text{Flatten}^{-1}(Y'_{mn}), \quad (12)$$

where $Y'_{mn} \in \mathbb{R}^{196 \times 1024}$ is 2D sequential data, and $Y_{mn} \in \mathbb{R}^{1024 \times 14 \times 14}$ is 3D image data. After extracting features from all subimages, the

entire histological feature map $Y = [Y_{mn}]_{m=1, n=1}^{M/16, N/16} \in \mathbb{R}^{1024 \times M/16 \times N/16}$ are obtained.

Additionally, certain regions within the tissue may exhibit similar histological features while differing in gene expression levels. By incorporating spatial location features, we aim to enhance the robustness and accuracy of the model. We calculate the positional feature map $P \in \mathbb{R}^{2 \times M/16 \times N/16}$ based on the two-dimensional spatial location information of the superpixels:

$$P(k, i, j) = \begin{cases} \alpha \frac{i}{M/16}, & k = 0 \\ \alpha \frac{j}{N/16}, & k = 1 \end{cases}, \quad (13)$$

where i and j represent the two-dimensional coordinates of superpixel, and the hyperparameter α is set to be 1 in the experiments.

Finally, we obtain the RGB feature map $T \in \mathbb{R}^{3 \times M/16 \times N/16}$ by resizing the entire histological image to the desired size $M/16 \times N/16$ using average pooling. From this point, we have obtained histological feature map Y , positional feature map P , and RGB

feature map T . By stacking these feature maps along the channel dimension, we obtain the multimodal feature map:

$$H = \text{concat}(Y, P, T) \in \mathbb{R}^C \times \mathbb{R}^{M/16} \times \mathbb{R}^{N/16}, \quad (14)$$

where H can be represented as $H = [h_{pq}]_{p=1,q=1}^{M/16,N/16}$, with each superpixel $h_{pq} \in \mathbb{R}^C$ is the feature vector at pixel (p, q) . $C = 1024 + 2 + 3$, which is the sum of the channels of the feature maps Y, P , and T .

Additionally, due to the benefits of transfer learning in ViT architecture, we pretrain ViT module through self-supervised learning. This step only requires histological images, making many publicly available histopathology datasets suitable for model pretraining. In our experiments, we utilized the general pathology self-supervised model named UNI [35], which uses DINOv2 [36] to pretrain on over 100 million images from diagnostic H&E-stained WSIs spanning 20 major tissue types.

Inferring super-resolution gene expression using GCN module

Our purpose in extracting multimodal feature map is to use it for inferring super-resolution gene expression profiles. In tissue structures, the gene expression exhibits a high degree of correlation with adjacent cells, far exceeding that with regions of tissue that are physically distant from each other. GCN module is well-suited for establishing complex communication relationships between neighboring cells. Therefore, we capture complex relational information between adjacent superpixels using an undirected graph $G(V, E)$ when predicting gene expression at the single-cell level. Each vertex V symbolizes the superpixel, characterized by a multimodal feature vector h_{pq} . And the edge E represents the connections between two vertices. Due to the large number of superpixels, we consider the set of superpixels spanned by each spot as a graph to reduce computational complexity. Let S be the number of spots in the original ST data, then training data $H_{\text{train}} = \{H_{\text{train}}^1, H_{\text{train}}^2, \dots, H_{\text{train}}^S\}$, where H_{train}^i is the set of superpixels spanned by spot i of size $D \times D$ with C channels. Here, D denotes the number of superpixels covered by the spot's diameter. Next, we define each superpixel as a node, and select the four closest nodes based on physical distance as its neighbors to generate the adjacency matrix A . After two layers of GCN, which fuse information between adjacent cell-level, a 512-dimensional feature vector at each superpixel is obtained. The above process can be described as follows:

$$H_{\text{train}}^{j(0)} = \text{Flatten}(H_{\text{train}}^j), \quad (15)$$

$$H_{\text{train}}^{j(1)} = \text{ReLU}(AH_{\text{train}}^{j(0)}W^0), \quad (16)$$

$$H_{\text{train}}^{j(2)} = \text{ReLU}(AH_{\text{train}}^{j(1)}W^1), \quad (17)$$

where $H_{\text{train}}^{j(0)}$ has shape (D^2, C) , $H_{\text{train}}^{j(1)}$ and $H_{\text{train}}^{j(2)}$ have the same shape $(D^2, 512)$, with D^2 represents the number of nodes for graph.

Next, $H_{\text{train}}^{j(2)}$ are input into the output layer containing a linear layer with 512 input nodes and K output nodes, where K represents the number of genes. The outputs are activated by an exponential linear unit (ELU) [37] to ensure that the predicted super-resolution

gene expressions are non-negative. The process of transferring from feature vectors to gene expressions is as follows:

$$Z_{\text{train}}^j = \text{ELU}(H_{\text{train}}^{j(2)}W^2 + b). \quad (18)$$

To train GCN module and output layer, due to training data lacks high-resolution gene expression profiles as labels, we employ a weak supervision framework. Specifically, we model the gene expression of each spot as the sum of the gene expressions of the superpixels covered by that spot. Regions not covered by any spot are excluded from the training process. Let H_{train}^j be the set of superpixels spanned by spot j , $F(\cdot)$ represents the network composed of the GCN module and the output layer, Z_{train}^j be the predicted super-resolution gene expressions corresponding to H_{train}^j , $\text{Filter}(\cdot)$ represents the operation of filtering out superpixels not contained within round area of spots, g_j be the observed gene expressions at spot j . Then the weak supervision framework can be described as follows:

$$Z_{\text{train}}^j = F(H_{\text{train}}^j), \quad (19)$$

$$\mathcal{M}(Z_{\text{train}}^j) = \text{Sum}(\text{Filter}(Z_{\text{train}}^j), 0), \quad (20)$$

$$\text{Loss}(\Theta) = \sum_{j=1}^S (g_j - \mathcal{M}(Z_{\text{train}}^j))^2, \quad (21)$$

where g_j and $\mathcal{M}(Z_{\text{train}}^j)$ are both spot-level gene expression vectors of length K .

We use ADAM optimizer to minimize the loss via mini-batch gradient descent. After completing the training phase, we neatly divide the entire multimodal feature map into a series of graphs as predicting data $H_{\text{test}} = [H_{\text{test}}^{ij}]_{i=1,j=1}^{M'/D,N'/D}$, where $M' = M/16$ and $N' = N/16$. And then we can infer super-resolution gene expression use the trained network $F(\cdot)$:

$$Y_{\text{test}}^{ij} = \text{Flatten}^{-1}(F(H_{\text{test}}^{ij})), \quad (22)$$

$$Y = \text{Mask}\left(\left[Y_{\text{test}}^{ij}\right]_{i=1,j=1}^{M'/D,N'/D}, X\right), \quad (23)$$

where the shape of Y is (M', N', K) , with each Y_{test}^{ij} has shape (D, D, K) . Function $\text{Mask}(\cdot)$ first detects the contour of the captured tissue regions in the histological image X , then masks out values outside that detected contour in super-resolution gene expression data Y . The k -th channel of Y with shape (M', N') represents the super-resolution gene expression profile predicted by scstGCN for the gene k .

Results

The details on the baseline methods, implementation details, and evaluation metrics can be found in [Supplementary Note 1, 2, and 3](#).

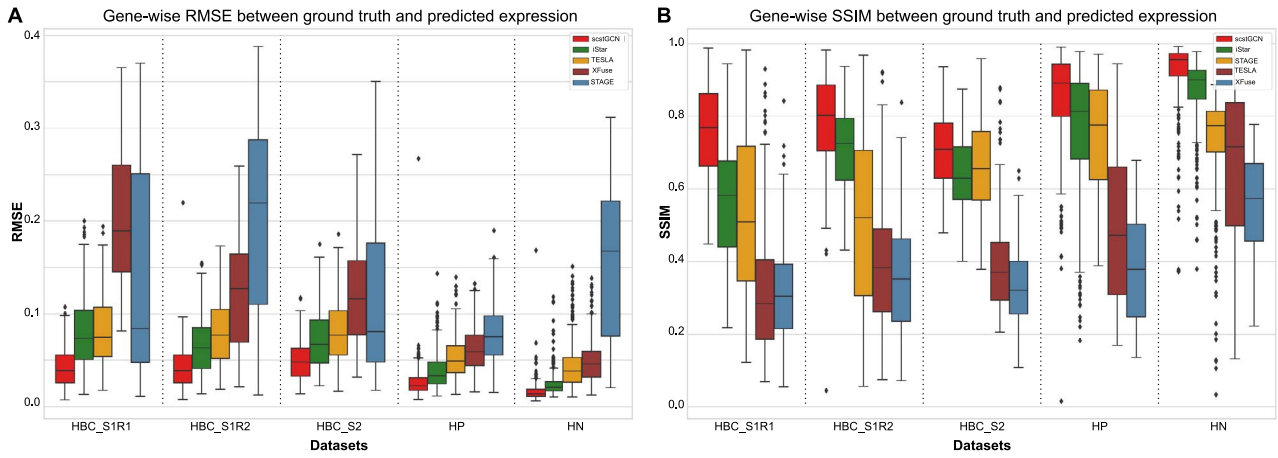


Figure 2. Results on multiple Xenium ST datasets demonstrate scstGCN can predict super-resolution gene expression with higher accuracy. (A) Calculation result in term of the RMSE metric. (B) Calculation result in term of the SSIM metric. We calculated RMSE and SSIM metrics between ground truth and predicted super-resolution gene expression using scstGCN, iStar, TESLA, XFuse, and STAGE on multiple ST datasets.

scstGCN can better predict super-resolution gene expression profiles

Due to the lack of single-cell resolution in spot-based ST data, it is not possible to perform precise quantitative evaluation of predicted super-resolution gene expression profile. To assess the accuracy of scstGCN in predicting super-resolution gene expression, our numerical evaluation experiments were conducted on multiple simulated datasets derived from Xenium data.

First, we applied scstGCN and other methods on pseudo-Visium dataset derived from the HBC dataset. HBC comprises two samples: Sample 1 includes two consecutive sections (denoted as HBC_S1R1 and HBC_S1R2), while Sample 2 contains only one section (denoted as HBC_S2). We compared the prediction accuracy of scstGCN with iStar [27], XFuse [25], TESLA [26], and STAGE [23]. Some methods were excluded such as BayesSpace [24] and stEnTrans [22] because BayesSpace separates a spot into several sub-spots and cannot impute gene expression for unmeasured locations; stEnTrans merely interpolated the gaps between spots and did not enhance the gene expression levels to super-resolution. We calculated root mean square error (RMSE) and structural similarity index measure (SSIM) [38] between ground truth and each predicted super-resolution gene expression. The results show that scstGCN attains lower median of gene-wise RMSE and higher median of gene-wise SSIM across all sections in HBC (Fig. 2). scstGCN's 75th percentile (0.055, 0.056, and 0.062, respectively) of RMSE is lower than the minimum of 50th percentiles of iStar (0.073, 0.063, and 0.067, respectively) in HBC (Fig. 2A). What's even weirder is that the 25th percentile of SSIM for scstGCN is higher than or comparable to the maximum 75th percentile of SSIM for iStar in HBC (Fig. 2B). Additionally, to demonstrate scstGCN's broad applicability, we conducted the same experiment on HP and HN datasets (details in Table 1), and the results were consistent with those observed in HBC data, with scstGCN achieving the best performance (Fig. 2).

To intuitively analyze the predictive capabilities of different methods, we spatially visualized several genes that have different spatial patterns. Figure 3 shows the comparison of pseudo-Visium, ground truth and the predicted super-resolution gene expression using scstGCN, iStar, XFuse and TESLA about certain genes. Visually, the predictions from scstGCN are the closest to the ground truth measured by Xenium data. Despite achieving

the same super-resolution as scstGCN, iStar predicts high gene expression in certain tissue areas (such as globally highly expressed in *ERBB2*, partially highly expressed in *KRT8* and *PTPRC*), XFuse exhibits distortion in genes with weak spatial patterns, and TESLA can only roughly predict gene expression patterns in all genes.

Finally, to illustrate that scstGCN has powerful transfer learning capabilities, the pseudo-Visium data of HBC_S1R2 were used as the training data, and super-resolution gene expression profiles was obtained on HBC_S1R1 using only its histological image as the input (Supplementary Figure S2). The results show that both in terms of evaluation metrics and spatial expression analysis, scstGCN outperforms the state-of-the-art method iStar.

scstGCN enables accurately restore gene expression data of Visium HD technology

Recently, 10X Genomics has launched Visium HD technology, which enhances the resolution to the single-cell level while maintaining the advantages of high gene throughput. Compared to Xenium, Visium HD has two advantages for our study: (1) the sequencing unit of Visium HD is a square area without gaps ($8 \times 8 \mu\text{m}$), which is consistent with the output of our task. This character eliminates the need to map scattered cellular gene expression onto a rectangle grid of superpixels, enhancing the reliability of the experimental results. (2) Visium HD offers the advantage of high gene throughput, allowing us to evaluate the performance of the model on a greater number of genes. In view of the above benefits that are more capable of verifying the accuracy of our method, we selected two Visium HD datasets: Human Breast Cancer (HBCHD) and Mouse Brain (MBHD).

We used the $8 \mu\text{m}$ bins of the Visium HD data as ground truth, which is approximately equal to single-cell resolution. Similar to the simulation process on Xenium data, pseudo-Visium data was obtained as the input of scstGCN and other baselines. The results show that scstGCN outperforms all baselines on MBHD and HBCHD datasets from different species. It is worth emphasizing that scstGCN's 25th percentile (0.7059 for MBHD and 0.6839 for HBCHD) of SSIM are even higher than the maximum of 75th percentile (0.6912 for MBHD and 0.6483 for HBCHD) of SSIM for state-of-the-art method iStar (Fig. 4A). Certainly, the 50th percentile of scstGCN is also significantly lower than the minimum 25th percentile of iStar in terms of RMSE. In addition,

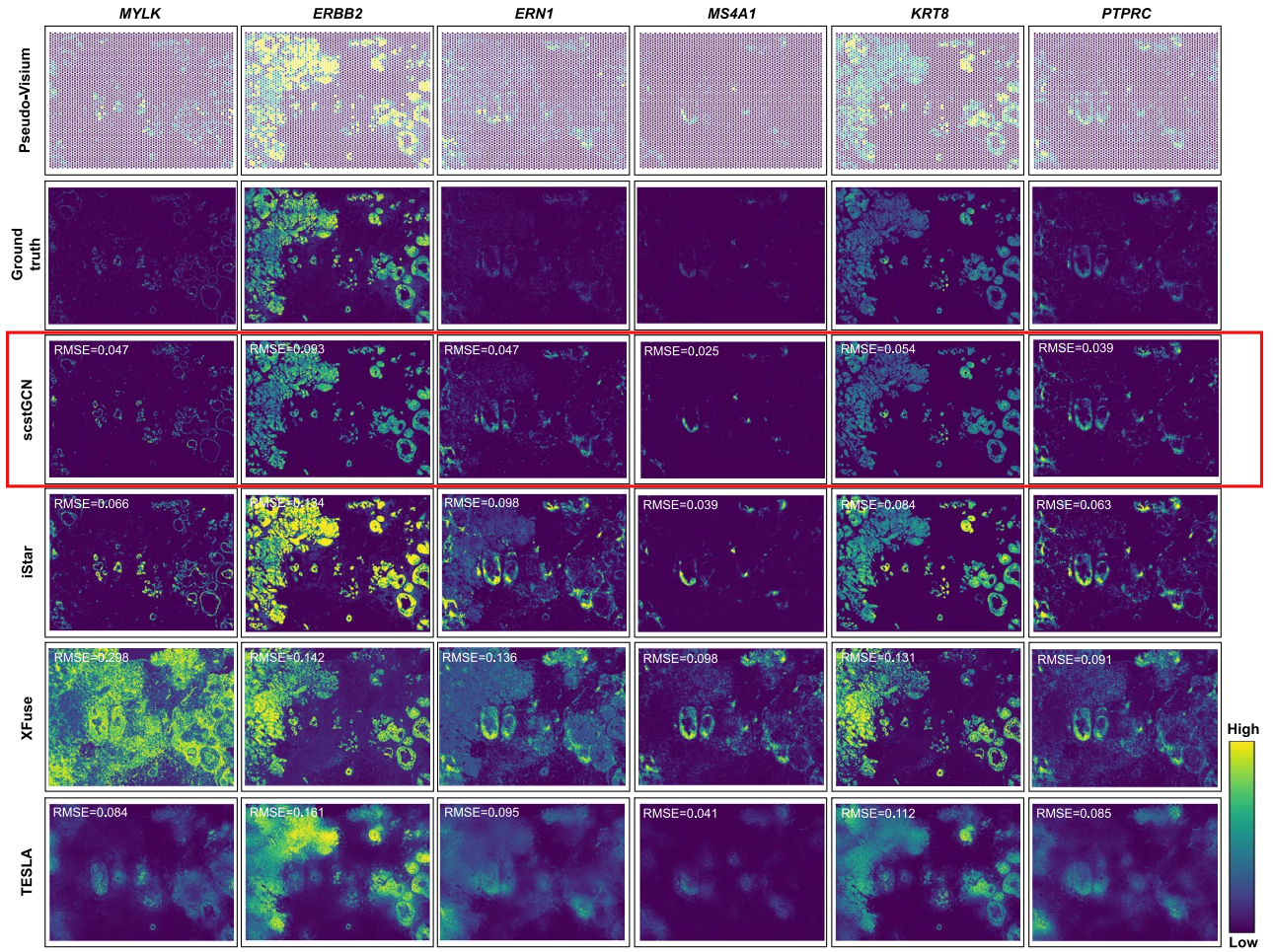


Figure 3. Spatial expression analysis of multiple groups of genes with different spatial patterns in HBC_S1R1 data further demonstrate the superiority of scstGCN. The analysis are based on pseudo-Visium data, super-resolution ground truth, and predicted data using scstGCN, iStar, XFuse, and TESLA. Each column corresponds to a gene, with the first two rows from the top displays the pseudo-Visium and ground truth, while the subsequent rows show the predicted data using different methods.

the super-resolution gene expression predicted by iStar on the Visium HD datasets exhibit many extreme values, which may be linked to the number of genes (Fig. 4A). As the number of genes increases, the gene expression data becomes sparser, making it more challenging to capture sufficient statistical information, which can lead to outliers for other baselines that don't establish complex communication relationships between neighboring cells. This results in a more pronounced advantage for scstGCN in terms of the mean evaluation metrics, with a mean RMSE of 0.0462 and 0.0507 and SSIM of 0.7919 and 0.7692 for scstGCN across MBHD and HBCHD datasets, compared to RMSE of 0.0881 and 0.1267 and SSIM of 0.5735 and 0.4874 for iStar.

In addition, due to the strong sense of pattern in the mouse brain, we spatially visualized several genes with different spatial patterns in ground truth and predicted profiles by scstGCN, iStar, XFuse, and TESLA in MBHD data (Fig. 4B). Visually, the single-cell resolution gene expression profiles predicted by scstGCN is closer to the ground truth. iStar only performs a simple mapping from hierarchical histological features to gene expression, without considering the complex relationship between adjacent cells, which results in discontinuous predicted gene mapping and prone to blurred phenomenon. XFuse is misled by intense morphological similarities between different regions in histology image, resulting in poor prediction of low-expression regions of genes.

TESLA comprehensively considers the physical distance between superpixels and the histological similarity, but does not deeply integrate gene expression with histological image, resulting in the overall smoothness of the predicted gene expression profiles, which deviates from the ground truth.

scstGCN demonstrates superior overall performance on DLPFC datasets

Here, we evaluated the ability of scstGCN to predict super-resolution gene expression on the DLPFC tissue with 12 sections from 10X Visium platform. Since DLPFC data lacks super-resolution gene expression as labels, we conducted a post-processing step to precisely evaluate the performance of scstGCN and iStar numerically. Specifically, we aggregated the predicted super-resolution gene expression data by summing the gene expression values of superpixels covered by each 10X Visium spot, thus obtaining predicted gene expression at the spot-level. As we have conducted comprehensive comparisons using three Xenium datasets and two Visium HD datasets that have single-cell resolution gene expression as labels, and considering the long runtimes of XFuse and TESLA and the need for post-processing steps on the DLPFC data, so we only consider comparison with the state-of-the-art method iStar on the DLPFC data. We evaluated the prediction performance numerically by calculating the RMSE,

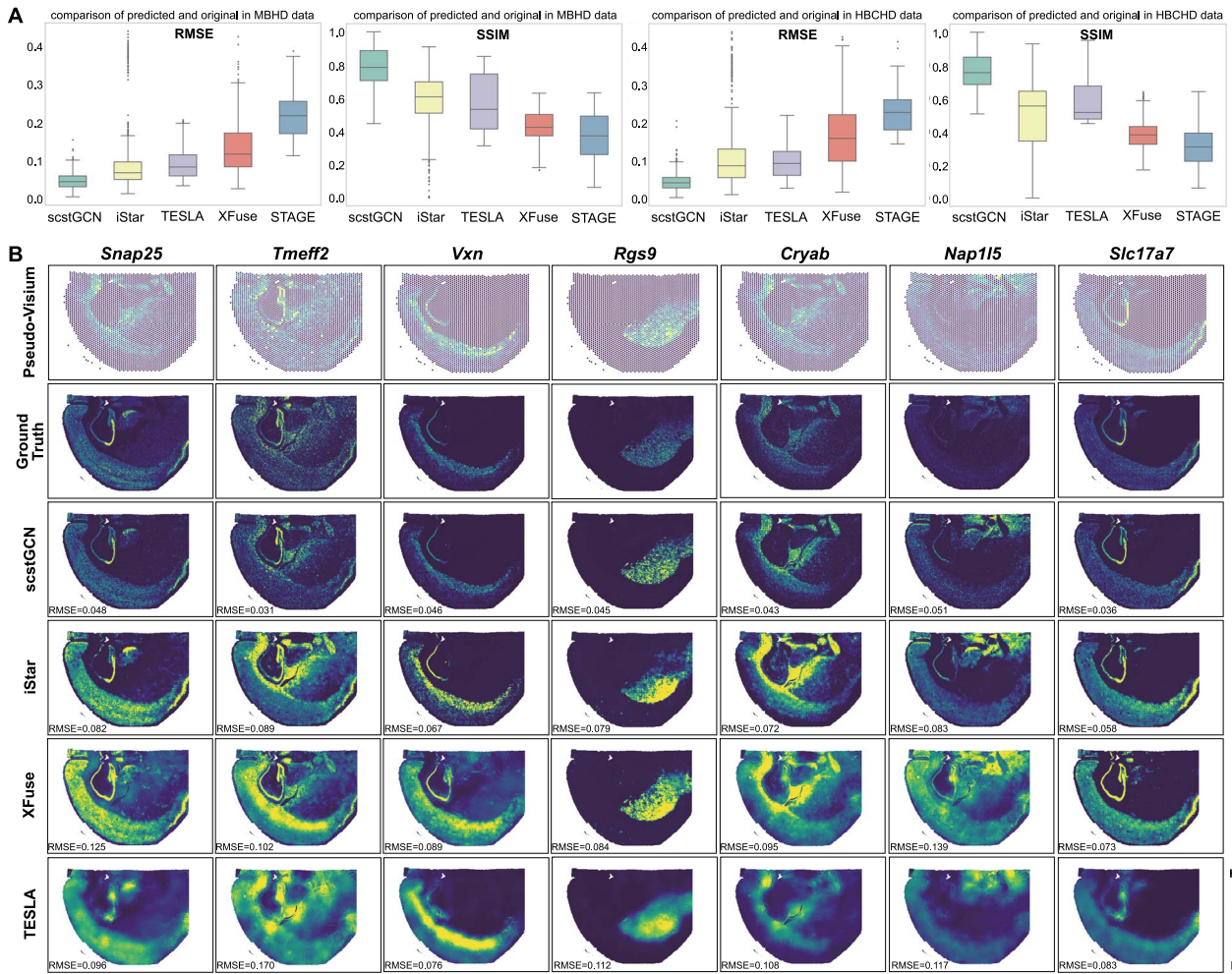


Figure 4. scstGCN enables accurately restore gene expression data for Mouse Brain (MBHD) and Human Breast Cancer (HBCHD) datasets from Visium HD technology. (A) Comparison of RMSE and SSIM between ground truth and predicted single-cell resolution gene expression by scstGCN, iStar, TESLA, XFuse, and STAGE on the MBHD and HBCHD datasets. (B) Spatial visualization of several genes having different spatial patterns for the ground truth and predicted data by scstGCN and other baselines in MBHD data.

Pearson correlation coefficient (PCC), and mean absolute error (MAE) between the original data and the predictions generated by scstGCN and iStar on the DLPFC dataset with 12 sections (Fig. 5A and Supplementary Figure S3). The results indicate that scstGCN outperformed iStar across almost all genes in 12 sections with statistical significance (median: PCC = 0.83, RMSE = 0.08, and MAE = 0.06 for scstGCN; PCC = 0.67, RMSE = 0.17, and MAE = 0.09 for iStar).

Similarly, we selected some highly variable genes that have different spatial patterns for spatial visualization to illustrate the advantages of scstGCN. The results show that the gene expressions predicted by scstGCN were closer to the ground truth, accurately recovering the original spatial patterns of the genes (Fig. 5B). In contrast, the results from iStar visually restored the spatial patterns but showed high expression levels throughout the tissue compared to the ground truth. iStar exhibits excessive smoothing in local regions of gene expression, resulting in blurriness, while scstGCN provides more finer depiction of gene expression. iStar predicts gene expression by considering only the features of the superpixel itself, without leveraging the effective knowledge from neighboring superpixels. On the other hand, scstGCN enables the neural network to adaptively learn different gene expression levels corresponding to different regions by

integrating spatial positional information into histological features. In addition, scstGCN can not only improve the resolution within the measured spot, but also predict single-cell resolution gene expression in non measured spot areas of histological images (Supplementary Figure S4).

To further demonstrate the superiority of scstGCN in maintaining the spatial structures, we performed clustering on the data obtained from scstGCN and iStar with spatial clustering methods AVGN, STAGATE, and GraphST, using Adjusted Rand Index (ARI) as the evaluation metric. The results show that the 'spot-level' data obtained from scstGCN is better than iStar in all spatial clustering algorithms, ARI has a significant improvement (Supplementary Table S2). scstGCN retains the structure of the spatial domain tightly when enhancing the resolution of spatial gene expression.

scstGCN can enhance the spatial patterns in the human DLPFC tissue

Next, we tested the ability of scstGCN to enhance the spatial patterns of genes in ST. We applied a tool called Sepal [39], which employs a diffusion-based model to identify transcripts exhibiting spatial patterns in the transcriptome, to DLPFC datasets before and after super-resolution prediction using scstGCN. Following the tutorial of Sepal, we ranked the original and super-resolution

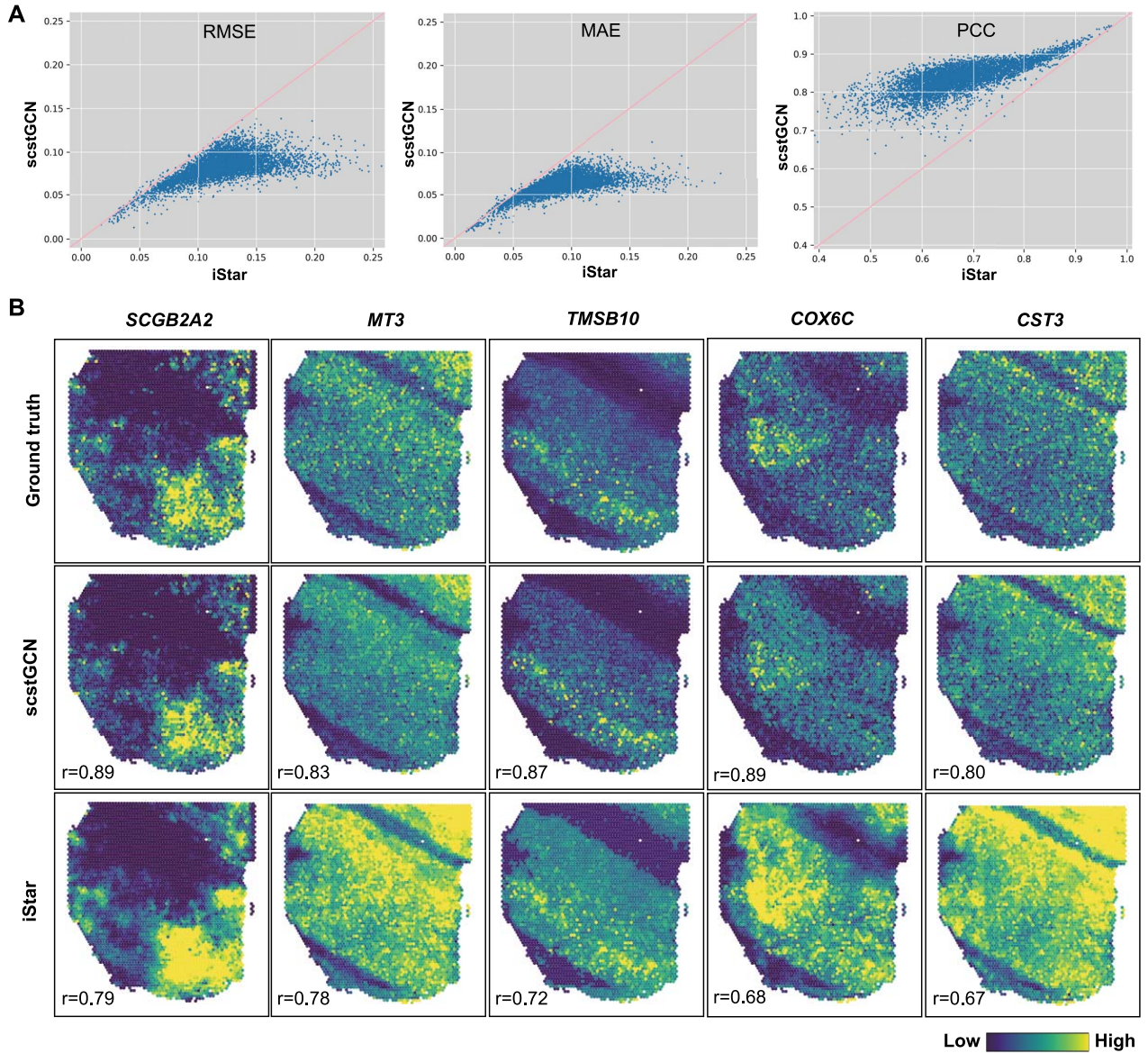


Figure 5. scstGCN demonstrates superior performance in the human DLPFC tissue compared to state-of-the-art method iStar. (A) Scatter plots of RMSE, MAE, and PCC between the original spot-level gene expression and ‘spot-level’ gene expression obtained from the enhanced expression generated by scstGCN and iStar for the top 1000 highly variable genes selected across all sections in DLPFC tissue. Each dot in plots represents one of the 1000 genes. (B) Spatial visualization of several genes having different spatial patterns for the ground truth and predicted data by scstGCN and iStar in section 151507, respectively.

gene expression profiles in each section by the degree of spatial structure from distinct to random patterns respectively; then grouped top-ranked genes that had distinct spatial patterns into pattern families based on the similarity of their spatial organization. Finally, we conducted functional enrichment analysis for each section in DLPFC datasets using the Gene Ontology: Biological Processes (GO: BP) database [40]. Figure 6A shows a subset of transcript profiles in 151510 section that top-ranked in the super-resolution data but lower rankings in original data. The super-resolution expressions of these genes have distinct spatial structure. In contrast, the spatial structures are more diffuse in the original data. *TPT1* is pivotal in cell growth regulation and cancer progression, making it a promising target for therapies aimed at inhibiting tumor development and metastasis [41]. *COX6C* is vital for mitochondrial function, influencing cellular energy production and potentially impacting diseases like neurodegeneration and cancer [42]. *PTGDS* is crucial for inflammation

and neuroprotection, making it a promising target for treating neurological and inflammatory conditions [43]. *RTN1* regulates neuronal morphology and function, impacting neurodegenerative disorders and neuronal regeneration [44]. The super-resolution data predicted by scstGCN helps researcher better discover spatial patterns of these genes, significantly improving their rankings.

Figure 6B shows that the DLPFC tissue has six cortical layers and white matter annotated manually. Next, we partitioned the top 100 ranked genes into pattern families using Sepal in both the original data and the super-resolution data across all sections (Fig. 6C). In the original Visium data, the number of families partitioned by Sepal is much greater than the manually annotated count because low-resolution expression blurs the family patterns. The super-resolution gene expression data brings the number of families closer to manual annotations because it maximally restores the actual biological features of the tissue and enhances spatial patterns.

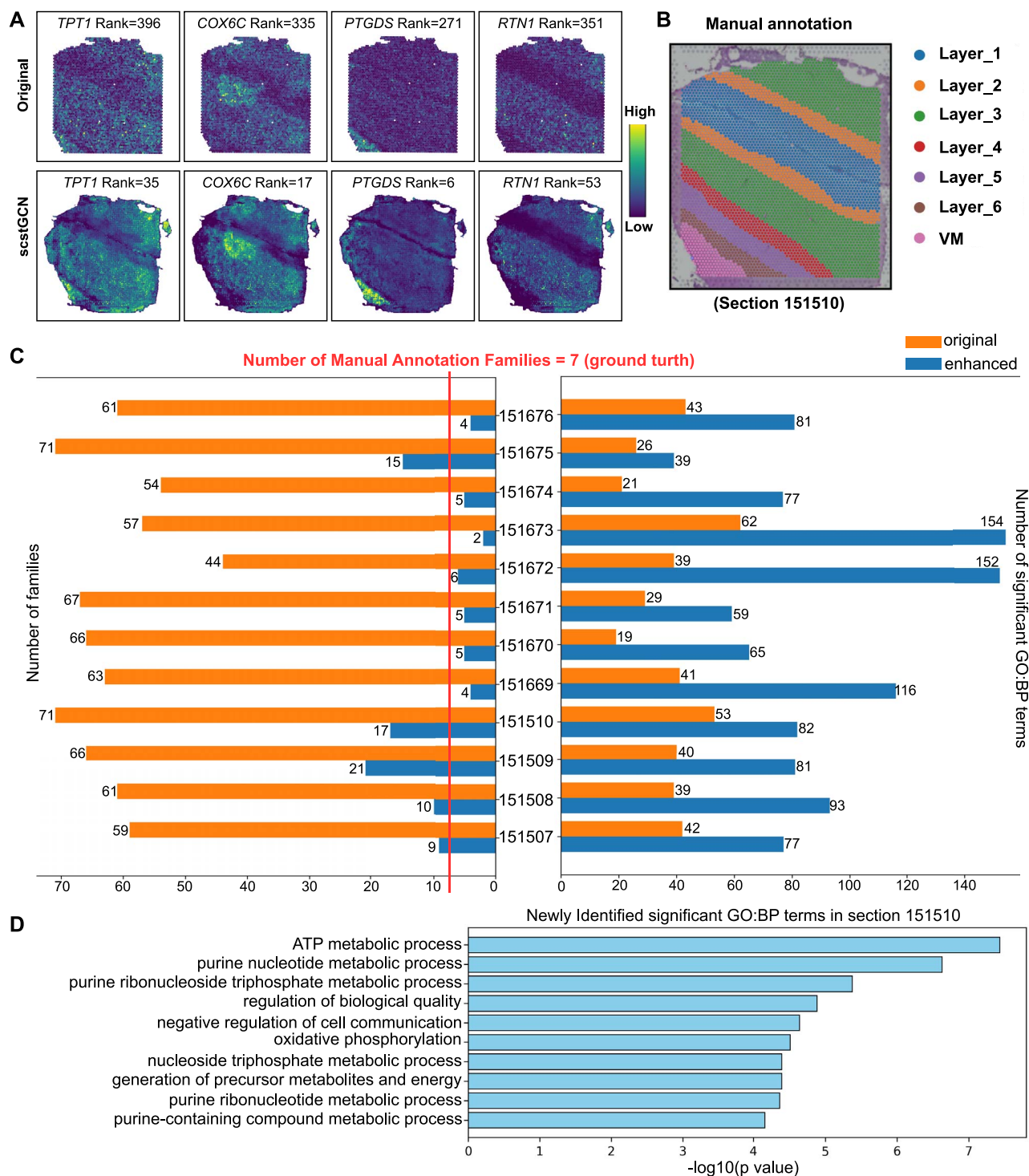


Figure 6. scstGCN can help find spatial patterns for DLPFC tissue and enrich more biologically significant pathways. (A) Some disease-related genes rank higher in super-resolution data predicted by scstGCN (below) because of distinct spatial patterns but rank lower in original Visium data (above). The head of each gene expression profile provides the names and specific ranking of the genes. (B) Manual annotation of hematoxylin-and-eosin-stained tissue image in section 151510. (C) Number of families divided according to similarity in spatial structure (left) and significant (P -value<0.05) GO:BP terms (right) for all sections. Orange represents the original Visium data, while blue represents the super-resolution data. (D) Top-10 significant GO:BP terms were only identified in the super-resolution data in section 151510.

Finally, we conducted enrichment analysis in DLPFC tissue among the top 100 ranked genes. The results show that super-resolution data were enriched more GO: BP term across all sections (Fig. 6C). Specifically, section 151672 enriched 113 new pathways, section 151673 enriched another 92 new pathways, and the least enriched section 151675 also identified 13 new pathways.

Moreover, high-resolution DLPFC data revealed a variety of biological processes closely related to metabolic activities and regulatory networks (Fig. 6D and Supplementary Figure S5), among which adenosine triphosphate metabolic production, precursor and energy metabolites, and oxidative phosphorylation were closely related to schizophrenia [45, 46], which provides new

perspectives and strategies for studying disease mechanisms and identifying therapeutic targets. By inferring super-resolution gene expression to accurately capture and analyze tissue structure details, scstGCN brings new advances to systems biology and biomedical research.

Explore insights by super-resolution gene expression

We further analyzed the capability of scstGCN in super-resolution tissue structure segmentation and annotation. We applied scstGCN to the HER2ST data [29], which has a lower resolution than Visium. Specifically, we considered three consecutively cut sections from sample *H*, where manual annotations were only provided for section *H_1*. We used the spot-based ST data and histology image from section *H_1* for training data, and then we used only the histology images from three sections as inputs to obtain the super-resolution gene expression. We used the output of the GCN as features for the k-means algorithm [47], and the resulting segmentation showed a strong concordance with the manual annotations (Fig. 7A). scstGCN successfully separated *in situ* cancer, Breast glands, and Adipose tissue from manually annotated section *H_1*. Additionally, the tissue segmentation results have similar structure across the three sections, demonstrating that scstGCN has robust transfer learning capabilities and maintains consistency across different sections from the same sample.

Next, we explored the potential of scstGCN in detecting multicellular structures. Tertiary lymphoid structures (TLSs) refer to specific types of lymphoid tissue structures that form in non-lymphoid tissues [48]. Research indicates that TLSs may play a significant role in tumor immune surveillance and therapeutic responses. However, the resolution of ST data may not be sufficient to accurately capture TLSs because they are typically small and sparsely distributed lymphoid structures, especially under low signal-to-noise conditions. We calculated TLS scores separately on the super-resolution gene expression and Visium data of the section *G_1* of another sample *G* in the HER2ST dataset by normalizing and averaging the TLS marker genes (Supplementary Table S3). The results show that TLSs detected from the Visium data are unusually sparse and have low density. On the contrary, super-resolution data predicted by scstGCN can detect more clearer and densely packed TLSs at a fine-grained characteristics (Fig. 7B).

Finally, we applied scstGCN to MBS and MBC datasets to demonstrate that scstGCN is a generic tool not limited to a specific type of cancer or healthy tissue. We used the same method as in Fig. 7A to perform tissue segmentation, and the results show that scstGCN could characterize fine-grained tissue structure with high-resolution (Fig. 7C and D). The fine-grained tissue segmentation based on super-resolution gene expression predicted by scstGCN closely aligns with Allen Brain Atlas annotations. This further demonstrates the accuracy and reliability of scstGCN. When Allen Brain Atlas Annotation does not fully cover certain data, scstGCN serves as a new effective tool for brain research, opening new avenues for disease research and therapeutic strategies.

In addition, we also compared the ability of the spatial clustering algorithms STAGATE and GraphST to identify spatial domains with scstGCN, although both methods only identify spatial domains at the spot-level. The results show that STAGATE and GraphST did not show excellent spatial domain recognition performance on HER2ST data (Supplementary Figure S6). In contrast, scstGCN provides a better fine-grained super-resolution segmentation effect, which is consistent with the rough manual

annotation. STAGATE and GraphST exhibited more pronounced effects in the specific regions identified in the MBC data, but the regions scstGCN identifies in MBS data may be more obvious (Supplementary Figure S7). scstGCN is not a tool specifically designed to identify spatial domains, which provides an idea for fine-grained high-resolution tissue segmentation.

Finally, we assessed the predictive performance of super-resolution gene expression in all of the aforementioned datasets. scstGCN outperformed iStar for virtually all genes across all datasets (Supplementary Figure S8).

Assessing the impact of each module in scstGCN on predicted single-cell resolution gene expression results

In order to comprehend the reasons behind the performance improvement of scstGCN compared to other methods, we assessed the contribution of each module to scstGCN, including various feature map from multimodal feature map (denoted as 'UNI', 'LOC', and 'RGB', respectively) and the GCN module to scstGCN, as shown in Table 2. We systematically removed various combinations of four modules, including individual and combined removals. It is worth noting that the histological feature maps must retain at least one of the 'UNI', 'LOC', and 'RGB' modules, and when the GCN is removed, we replace it with a feedforward neural network with four hidden layers, each containing 512 nodes, and ReLU as the activation function for the hidden layers was employed. From the numerical evaluation results of RMSE and SSIM in five Xenium datasets, keeping all modules intact provides the best alignment of the predicted super-resolution gene expression with the ground truth. Notably, removing the GCN module resulted in a significant performance drop across all datasets, only slightly better than iStar. Removing the UNI module led to an even more substantial decline.

But one key point to emphasize is that removing UNI and relying solely on 'LOC' and 'RGB' modules for gene expression prediction will inevitably lead to a significant drop in performance. In contrast, if UNI is replaced with other pre-trained large models for extracting histological features, the differences will be minimal. To further demonstrate the contribution of our study, we compare scstGCN to state-of-the-art method iStar where both are using the same underlying ViT. We replaced the histological feature extraction of scstGCN with the HIPT hierarchical Transformer to make it consistent with iStar, and we refer to this method as 'scstGCN+HIPT'. On the other hand, we also replaced iStar's hierarchical histological feature extractor HIPT with UNI to make it consistent with scstGCN, which we refer to as 'iStar+UNI'. We conducted a series of numerical evaluation experiments on the HBC datasets from Xenium platform and the MBHD and HBCHD datasets from Visium HD platform by comparing the prediction accuracy of those methods. The results indicate that replacing UNI with HIPT hierarchical Transformer did not significantly decrease the model's performance compared to the original scstGCN. Even with the replacement of UNI with HIPT for feature extraction, our method ('scstGCN+HIPT') still achieved higher median and mean values in RMSE and SSIM compared to iStar (Table 3 and Supplementary Figure S9). Specifically, compared to iStar on the mean, 'scstGCN+HIPT' has increased by 42.5%, 55.0% on the RMSE, and 35.9%, 55.5% on the SSIM, respectively, on MBHD and HBCHD datasets, with similar improvements observed in three sections of HBC dataset (details in Table 3). In addition, the median comparison with iStar yielded results consistent with the mean. We also analyzed the difference in

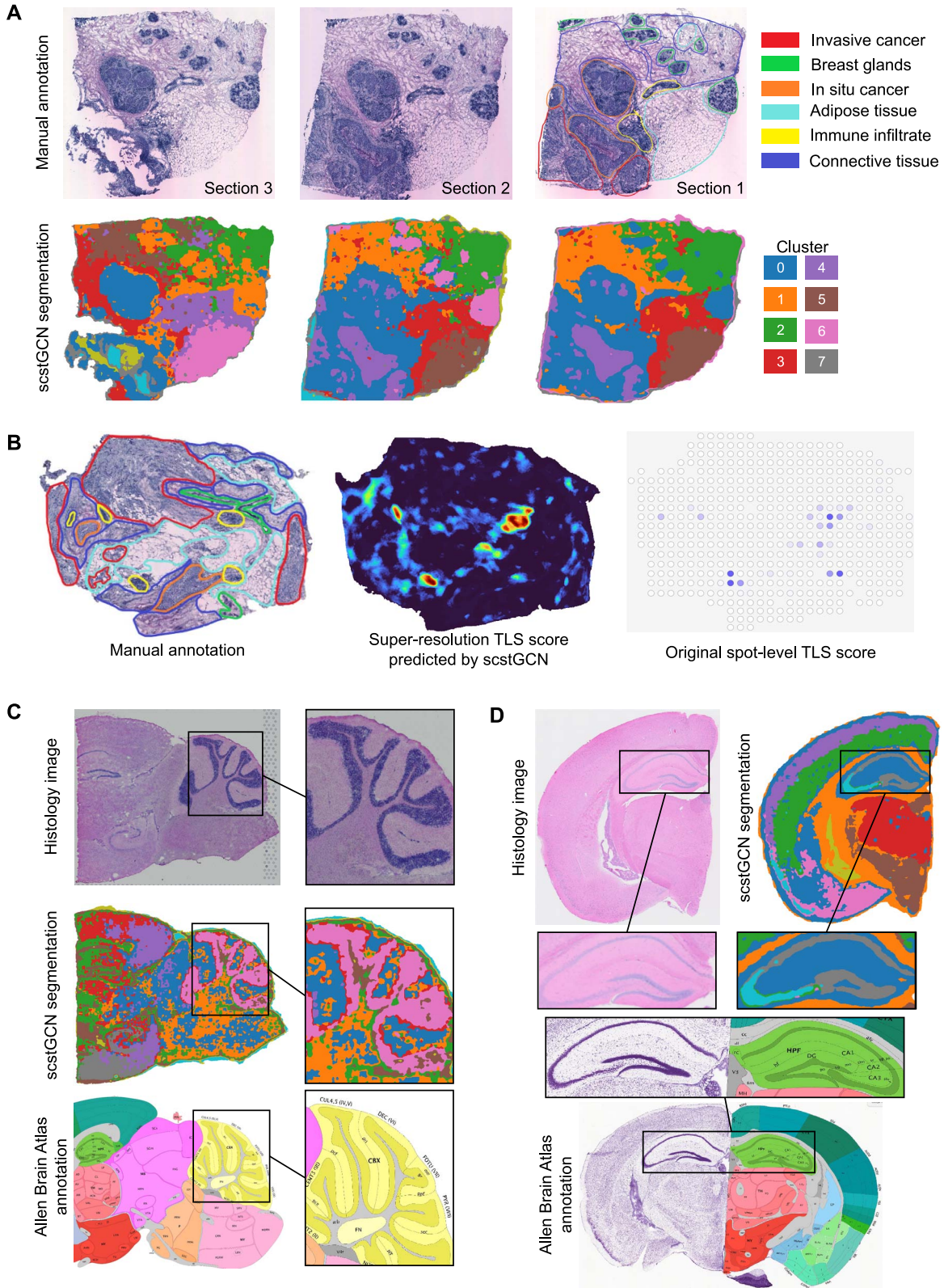


Figure 7. Exploring the capability of scstGCN for segmenting and annotating tissue structures at super-resolution across different types of datasets. (A) Annotating tissue architecture at superpixel-level using scstGCN on three consecutive tissue sections from sample H in the HER2ST breast cancer dataset. (B) Super-resolution gene expression obtained from section G_1 of HER2ST breast cancer dataset by scstGCN can facilitate the discovery of TLSs. (C, D) The tissue segmentation results based on super-resolution gene expression closely resemble the details of the Allen Brain Atlas. (C) mouse brain coronal cut data. (D) mouse brain sagittal cut posterior data.

Table 2. Ablation experiments on the HBC, HP, HN datasets from Xenium platform. The results showed the average RMSE and SSIM for scstGCN and its variants after removing specific modules in each dataset.

UNI	LOC	RGB	GCN	HBC_S1R1		HBC_S1R2		HBC_S2		HP		HN	
				RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
✓	✓	✓	✓	0.0420	0.7579	0.0419	0.7835	0.0495	0.7395	0.0259	0.8462	0.0231	0.9051
✗	✓	✓	✓	0.1487	0.2723	0.1525	0.3099	0.1929	0.1740	0.2298	0.1056	0.2363	0.1019
✓	✗	✓	✓	0.0441	0.7235	0.0445	0.7493	0.0511	0.7339	0.0271	0.8219	0.0277	0.8942
✓	✓	✗	✓	0.0432	0.7423	0.0423	0.7629	0.0498	0.7095	0.0269	0.8214	0.0245	0.9006
✓	✓	✓	✗	0.0642	0.6564	0.0618	0.6934	0.0699	0.6412	0.0385	0.7627	0.0278	0.8208
✗	✗	✓	✓	0.1459	0.2672	0.1378	0.3209	0.1476	0.3519	0.2364	0.0924	0.4605	0.0363
✗	✓	✗	✓	0.3405	0.1176	0.3242	0.1821	0.3880	0.0898	0.5594	0.0026	0.3996	0.0171
✗	✓	✓	✗	0.1203	0.3137	0.1211	0.3547	0.1204	0.3304	0.3807	0.0688	0.1562	0.1399
✓	✗	✗	✓	0.0432	0.7259	0.0441	0.7494	0.0519	0.6909	0.0281	0.8106	0.0251	0.8936
✓	✗	✓	✗	0.0627	0.6636	0.0607	0.6913	0.0687	0.6425	0.0386	0.7606	0.0286	0.8149
✓	✓	✗	✗	0.0636	0.6581	0.0618	0.6949	0.0700	0.6399	0.0385	0.7641	0.0290	0.8126
✓	✗	✗	✗	0.0625	0.6653	0.0607	0.6917	0.0686	0.6452	0.0277	0.8426	0.0280	0.8147
✗	✗	✓	✗	0.1560	0.1909	0.1587	0.2618	0.1081	0.4215	0.3438	0.0386	0.3777	0.0892
✗	✓	✗	✗	0.1785	0.2601	0.1971	0.2506	0.2251	0.1669	0.4154	0.0659	0.2013	0.0713

Table 3. Comprehensively evaluate the performance of scstGCN and iStar, under the condition that both use the same pretrained ViT. ‘Increased proportion’ represents the percentage of performance improvement of scstGCN+HIPT compared to iStar.

Mean prediction error	HBC_S1R1		HBC_S1R2		HBC_S2		MBHD		HBCHD	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
scstGCN	0.0420	0.7579	0.0419	0.7835	0.0495	0.7395	0.0462	0.7919	0.0507	0.7692
scstGCN+HIPT	0.0447	0.7219	0.0435	0.7704	0.0505	0.7303	0.0507	0.7795	0.0570	0.7577
iStar	0.0787	0.5715	0.0660	0.6853	0.0733	0.6379	0.0881	0.5735	0.1267	0.4874
iStar+UNI	0.0679	0.6153	0.0666	0.6449	0.0736	0.6141	0.0816	0.6185	0.1118	0.5458
Increased proportion	43.2%	26.3%	34.1%	12.4%	31.1%	14.5%	42.5%	35.9%	55.0%	55.5%
Median prediction error	HBC_S1R1		HBC_S1R2		HBC_S2		MBHD		HBCHD	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
scstGCN	0.0389	0.7688	0.0389	0.8014	0.0481	0.7386	0.0447	0.7840	0.0489	0.7546
scstGCN+HIPT	0.0405	0.7272	0.0412	0.7891	0.0474	0.7269	0.0475	0.7794	0.0538	0.7440
iStar	0.0736	0.5811	0.0632	0.7256	0.0679	0.6295	0.0679	0.6057	0.0986	0.5537
iStar+UNI	0.0611	0.6262	0.0652	0.6579	0.0708	0.6018	0.0692	0.6247	0.0944	0.5400
Increased proportion	45.0%	25.1%	34.8%	8.8%	30.2%	15.5%	30.0%	28.7%	45.4%	34.4%

performance between iStar and ‘iStar+UNI’. For numerical evaluation metrics, ‘iStar+UNI’ has a slight improvement in HBC_S1R1 but even a slight decrease in HBC_S1R1 and HBC_R2 compared to iStar. To further verify the importance of GCN module, we set up a method to improve iStar, introducing GCN module to capture complex relationships between neighboring cells while keeping the upstream feature map acquisition part unchanged. The experimental results show that the introduction of GCN to capture complex relationships between neighboring cells significantly improves the prediction accuracy in iStar (Supplementary Figure S10 and Table S4).

The experimental results indicate that there is a minimal performance difference in predicting single-cell resolution spatial gene expression between using UNI and other pre-trained large models as feature extractors for histology images. On the contrary, establishing complex communication relationships between neighboring cells using the GCN module is crucial for enhancing resolution of gene expression. Relying on histological features to predict gene expression is the cornerstone, and the GCN module of establishing complex communication relationships between neighboring cells plays a crucial role in improving accuracy. Additionally, removing the ‘LOC’ and ‘RGB’ modules from the

histological feature maps also reduced performance to varying degrees.

Conclusion

ST is a revolutionary technology with tremendous potential for elucidating cellular composition, understanding molecular interactions between tissue components, and advancing disease research [49]. Despite its advantages, the low resolution or inability to cover the entire transcriptome limits their application. Computational predictions of high-resolution gene expression are effective solution, but existing methods often fail to accurately capture the full complexity of cellular features. Here, we developed scstGCN, a GCN-based method that leverages a weakly supervised learning framework to integrate multimodal information and then infer super-resolution gene expression at single-cell level. scstGCN can accurately enhance gene expression from the spot-level to the superpixel-level, and it can predict expression both outside the spots and in external tissue sections.

However, the datasets used in this study are limited to a narrow range of sequencing platform types. In future studies, we consider introducing diverse data types from multiple sequencing

platforms. Our ideal solution is to develop a unified framework designed to enhance the spatial resolution of gene expression across various data types derived from different sequencing technologies. In addition, scstGCN currently lacks the capability to integrate multi-omics data. In future studies, we aim to incorporate single-cell RNA sequencing data and effectively fuse it with ST data and histological images using an attention mechanism. By combining multi-omics data at the informative feature representation level, we expect to enhance the accuracy and robustness of scstGCN's single-cell resolution gene expression predictions.

Identifying genes that display spatial expression patterns is a crucial step in characterizing the complex tissue landscapes in ST studies. Due to low resolution and high noise levels, original transcriptomics data makes it challenging to identify biologically significant genes with spatial patterns. We demonstrated that scstGCN can help identify spatial patterns of disease-related genes, which is of significant importance for understanding disease mechanisms by analyzing all sections from DLPFC tissue. Enrichment analysis in DLPFC tissue has shown that scstGCN allows for the enrichment of more biologically significant pathways, thereby providing deeper insights into biological processes. The ability of scstGCN to annotate tissues at super-resolution demonstrates great potential, showing consistency with manual annotations while providing higher granularity.

In summary, scstGCN provides a robust approach for ST data enhancement, including super-resolution gene expression generation, identification of spatial patterns of genes, enrichment of biologically significant pathways, and tissue structures segmentation. In addition, we quantified the runtime and memory consumption of scstGCN. scstGCN is the second fastest method, only slightly higher than iStar. And scstGCN consumes the least Video Random Access Memory (Supplementary Table S5).

Key Points

- scstGCN combines multi-modal information including histology image, spot-based spatial transcriptomics (ST) data, and physical spatial location through deep learning methods to achieve single-cell resolution of spot-based ST data without requiring single-cell references.
- scstGCN employs GCN to capture complex relationships between neighboring cells, facilitating the integration of multimodal feature information based on single-cell level, and then accurately infers single-cell resolution spatial gene expression.
- scstGCN can infer single-cell resolution gene expression across the entire tissue region. Through transfer learning, gene expression in three-dimensional tissues can be characterized efficiently. Furthermore, it demonstrates outstanding performance in spatial patterns enhancement, functional enrichment analysis, and annotate tissues at the high-resolution.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

The work was supported in part by the National Natural Science Foundation of China (62262069 and 12101026), Yunnan

Fundamental Research Projects (202301AT070230), and Young Talent Program of Yunnan Province (C619300A067).

Data availability

All datasets in this study are available from their original publications. The Xenium human breast cancer data can be found at <https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast>. The Xenium human pancreas data are available at <https://www.10xgenomics.com/datasets/ffpe-human-pancreas-with-xenium-multimodal-cell-segmentation-1-standard>. The Xenium hHeart Non-diseased data can be downloaded from the Genomics website <https://www.10xgenomics.com/datasets/human-heart-data-xenium-human-multi-tissue-and-cancer-panel-1-standard>. The Visium HD mouse brain (fresh frozen) data are available from <https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-mouse-brain-fresh-frozen>. The Visium HD human breast cancer (fresh frozen) data can be found at <https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-human-breast-cancer-fresh-frozen>. The human dorsolateral prefrontal cortex tissue datasets are available in the spatialLIBD package <http://spatial.libd.org/spatialLIBD>. The human HER2-positive breast tumor ST data are available in the her2st package <https://github.com/almaan/her2st>. The 10X Visium mouse brain coronal cut data can be found at <https://www.10xgenomics.com/resources/datasets/mouse-brain-coronalsection-2-ffpe-2-standard>. The 10X Visium mouse brain sagittal cut data can be downloaded from the Genomics website <https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-2-sagittal-posterior-1-standard>.

Code availability

All source codes of the scstGCN algorithm have been deposited at <https://github.com/wenwenmin/scstGCN>.

References

1. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;**19**:534–46. <https://doi.org/10.1038/s41592-022-01409-2>.
2. Rosenberg AB, Roco CM, Muscat RA. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 2018;**360**:176–82. <https://doi.org/10.1126/science.aam8999>.
3. Han X, Wang R, Zhou Y. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* 2018;**172**:1091–1107.e17. <https://doi.org/10.1016/j.cell.2018.02.001>.
4. Rao A, Barkley D, França GS. et al. Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;**596**:211–20. <https://doi.org/10.1038/s41586-021-03634-9>.
5. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* 2020;**17**:193–200. <https://doi.org/10.1038/s41592-019-0701-7>.
6. Lohoff T, Ghazanfar S, Missarova A. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat Biotechnol* 2022;**40**:74–85. <https://doi.org/10.1038/s41587-021-01006-2>.
7. Piwecka M, Rajewsky N, Rybak-Wolf A. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nat Rev Neurol* 2023;**19**:346–62. <https://doi.org/10.1038/s41582-023-00809-y>.

8. Valdeolivas A, Amberg B, Giroud N. et al. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *NPJ Precis Oncol* 2024;**8**:10. <https://doi.org/10.1038/s41698-023-00488-4>.
9. Alon S, Goodwin DR, Sinha A. et al. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* 2021;**371**:eaax2656. <https://doi.org/10.1126/science.aax2656>.
10. Wang X, Allen WE, Wright MA. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**. <https://doi.org/10.1126/science.aat5691>.
11. Zhang M, Eichhorn SW, Zingg B. et al. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature* 2021;**598**:137–43. <https://doi.org/10.1038/s41586-021-03705-x>.
12. Chen KH, Boettiger AN, Moffitt JR. et al. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**:aaa6090. <https://doi.org/10.1126/science.aaa6090>.
13. Eng C-HL, Lawson M, Zhu Q. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 2019;**568**: 235–9. <https://doi.org/10.1038/s41586-019-1049-y>.
14. Rodriguez R, Krishnan Y. The chemistry of next-generation sequencing. *Nat Biotechnol* 2023;**41**:1709–15. <https://doi.org/10.1038/s41587-023-01986-3>.
15. Lei L, Han K, Wang Z. et al. Attention-guided variational graph autoencoders reveal heterogeneity in spatial transcriptomics. *Brief Bioinform* 2024;**25**. <https://doi.org/10.1093/bib/bbae173>.
16. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 2022;**13**:1739. <https://doi.org/10.1038/s41467-022-29439-6>.
17. Min W, Fang D, Chen J. et al. Dimensionality reduction and denoising of spatial transcriptomics data using dual-channel masked graph autoencoder. *bioRxiv*. 2024;01–20. <https://www.biorxiv.org/content/10.1101/2024.05.30.596562v1>.
18. Wan X, Xiao J, Tam SST. et al. Integrating spatial and single-cell transcriptomics data using deep generative models with spatialscope. *Nat Commun* 2023;**14**:7848. <https://doi.org/10.1038/s41467-023-43629-w>.
19. Li X, Zhu F, Min W. SpaDiT: diffusion transformer for spatial gene expression prediction using scRNA-seq. *Brief Bioinform* 2024;**25**:bbae571.
20. Xie R, Pang K. et al. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Adv Neural Inf Process Syst* 2024;**36**:1–12.
21. Min W, Shi Z, Zhang J. et al. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Brief Bioinform* 2024;**25**. <https://doi.org/10.1093/bib/bbae551>.
22. Xue S, Zhu F, Wang C. et al. stEnTrans: transformer-based deep learning for spatial transcriptomics enhancement. In: *International Symposium on Bioinformatics Research and Applications*. Springer, 2024, 63–75. https://doi.org/10.1007/978-981-97-5128-0_6.
23. Li S, Gai K, Dong K. et al. High-density generation of spatial transcriptomics with stage. *Nucleic Acids Res* 2024;**52**:4843–56. <https://doi.org/10.1093/nar/gkae294>.
24. Zhao E, Stone MR, Ren X. et al. Spatial transcriptomics at sub-spot resolution with bayesspace. *Nat Biotechnol* 2021;**39**:1375–84. <https://doi.org/10.1038/s41587-021-00935-2>.
25. Bergenstr hle L, He B, Bergenstr hle J. Super-resolved spatial transcriptomics by deep data fusion. *Nat Biotechnol* 2022;**40**: 476–9. <https://doi.org/10.1038/s41587-021-01075-3>.
26. Jian H, Coleman K, Zhang D. et al. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with tesla. *Cell Syst* 2023;**14**:404–417.e4. <https://doi.org/10.1016/j.cels.2023.03.008>.
27. Zhang D, Schroeder A, Yan H. et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat Biotechnol* 2024;**42**:1372–7. <https://doi.org/10.1038/s41587-023-02019-9>.
28. Janesick A, Shelansky R, Gottscho AD. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* 2023;**14**:8353. <https://doi.org/10.1038/s41467-023-43458-x>.
29. Andersson A, Larsson L, Stenbeck L. et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* 2021;**12**:6012. <https://doi.org/10.1038/s41467-021-26271-2>.
30. Maynard KR, Collado-Torres L, Weber LM. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**:425–36. <https://doi.org/10.1038/s41593-020-00787-0>.
31. Tsai P-C, Lee T-H, Kuo K-C. et al. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun* 2023;**14**:2102. <https://doi.org/10.1038/s41467-023-37179-4>.
32. Mondol RK, Millar EKA, Graham PH. et al. hist2RNA: an efficient deep learning architecture to predict gene expression from breast cancer histopathology images. *Cancer* 2023;**15**:2569. <https://doi.org/10.3390/cancers15092569>.
33. Dosovitskiy A, Beyer A, Kolesnikov A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*, 2020.
34. Hendrycks D, Gimpel K. Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. In: *International Conference on Learning Representations (ICLR)*, 2016.
35. Chen RJ, Ding T, Lu MY. et al. Towards a general-purpose foundation model for computational pathology. *Nat Med* 2024;**30**: 850–62. <https://doi.org/10.1038/s41591-024-02857-3>.
36. Oquab M, Darcet T. et al. DINOv2: learning robust visual features without supervision. *Trans Mach Learn Res J* 2024;1–31.
37. Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). In: *International Conference on Learning Representations (ICLR)*, 2016.
38. Wang Z, Bovik AC, Sheikh HR. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 2004;**13**:600–12. <https://doi.org/10.1109/TIP.2003.819861>.
39. Andersson A, Lundberg J. Sepal: identifying transcript profiles with spatial patterns by diffusion-based modeling. *Bioinformatics* 2021;**37**:2644–50. <https://doi.org/10.1093/bioinformatics/btab164>.
40. Raudvere U, Kolberg L, Kuzmin I. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;**47**:W191–8. <https://doi.org/10.1093/nar/gkz369>.
41. Bae S-Y, Byun S, Bae SH. et al. TPT1 (tumor protein, translationally-controlled 1) negatively regulates autophagy through the BECN1 interactome and an MTORC1-mediated pathway. *Autophagy* 2017;**13**:820–33. <https://doi.org/10.1080/15548627.2017.1287650>.
42. Tian B-X, Sun W, Wang SH. et al. Differential expression and clinical significance of cox6c in human diseases. *Am J Transl Res* 2021;**13**:1–10.
43. Shunfeng H, Ren S, Cai Y. et al. Glycoprotein PTGDS promotes tumorigenesis of diffuse large B-cell lymphoma by MYH9-mediated regulation of Wnt- β -catenin-STAT3 signaling. *Cell Death Differ* 2022;**29**:642–56. <https://doi.org/10.1038/s41418-021-00880-2>.

44. Gong L, Tang Y, An R. et al. RTN1-C mediates cerebral ischemia/reperfusion injury via ER stress and mitochondria-associated apoptosis pathways. *Cell Death Dis* 2017;**8**:e3080–0. <https://doi.org/10.1038/cddis.2017.465>.
45. Sullivan CR, Koene RH, Hasselfeld K. et al. Neuron-specific deficits of bioenergetic processes in the dorsolateral prefrontal cortex in schizophrenia. *Mol Psychiatry* 2019;**24**:1319–28. <https://doi.org/10.1038/s41380-018-0035-3>.
46. Martins-de-Souza D, Gattaz WF, Schmitt A. et al. Proteomic analysis of dorsolateral prefrontal cortex indicates the involvement of cytoskeleton, oligodendrocyte, energy metabolism and new potential markers in schizophrenia. *J Psychiatr Res* 2009;**43**: 978–86. <https://doi.org/10.1016/j.jpsychires.2008.11.006>.
47. Hamerly G, Elkan C. Learning the k in k-means. *Adv Neural Inf Process Syst* 2003;**16**:281–8.
48. Fridman WH, Meylan M, Petitprez F. et al. B cells and tertiary lymphoid structures as determinants of tumour immune contexture and clinical outcome. *Nat Rev. Clin Oncol* 2022;**19**:441–57. <https://doi.org/10.1038/s41571-022-00619-z>.
49. Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. *Nat Biotechnol* 2023;**41**:773–82. <https://doi.org/10.1038/s41587-022-01448-2>.