Computational and Structural Biotechnology Journal 23 (2024) 1786–1795

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

A contrastive learning approach to integrate spatial transcriptomics and histological images

Yu Lin [a,b], Yanchun Liang [c,d], Duolin Wang [b], Yuzhou Chang [e], Qin Ma [e], Yan Wang [a,d,*], Fei He [b,f,**], Dong Xu [b,***]

[a] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[b] Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[c] School of Computer Science, Zhuhai College of Science and Technology, Zhuhai 519041, China
[d] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[e] Department of Biomedical Informatics, Ohio State University, Columbus, OH 43210, United States
[f] School of Information Science and Technology, Northeast Normal University, Changchun 130117, China

ARTICLE INFO

Keywords:
Spatial transcriptomics
Multi-modal data integration
Tissue architecture identification
Contrastive learning
Graph neural network

ABSTRACT

The rapid growth of spatially resolved transcriptomics technology provides new perspectives on spatial tissue architecture. Deep learning has been widely applied to derive useful representations for spatial transcriptome analysis. However, effectively integrating spatial multi-modal data remains challenging. Here, we present ConGcR, a contrastive learning-based model for integrating gene expression, spatial location, and tissue morphology for data representation and spatial tissue architecture identification. Graph convolution and ResNet were used as encoders for gene expression with spatial location and histological image inputs, respectively. We further enhanced ConGcR with a graph auto-encoder as ConGaR to better model spatially embedded representations. We validated our models using 16 human brains, four chicken hearts, eight breast tumors, and 30 human lung spatial transcriptomics samples. The results showed that our models generated more effective embeddings for obtaining tissue architectures closer to the ground truth than other methods. Overall, our models not only can improve tissue architecture identification's accuracy but also may provide valuable insights and effective data representation for other tasks in spatial transcriptome analyses.

1. Introduction

Characterizing tissue architecture is crucial for understanding the biological functions and mechanisms in spatial transcriptome analysis. Recent technologies for generating spatial transcriptomics data using platforms such as Visium from 10X Genomics [1,2] and MERFISH [3,4] are effective for studying tissue architecture heterogeneity in the spatial context. Morphology images, RNA-seq gene expression, and various other multi-modal data profiles provide complementary information on tissue architecture. However, using computational methods to properly integrate spatial multi-modal data for identifying tissue architecture is still challenging in spatial transcriptome analysis.

Recently, several computational methods have combined gene expression, spatial location, and morphology information to identify spatial tissue architecture. BayesSpace [5] employs a Bayesian statistical method using prior knowledge to cluster spots into distinct domains. SpaGCN [6] integrates gene expression, spatial location, and histology to identify spatial domains by deriving spatial dependency through graph convolution. stLearn [7] applies gene expression normalization to incorporate tissue morphology information into spatial clustering analysis. stMVC [8] uses attention-based multi-view graph collaborative learning to analyze spatial tissue heterogeneity by leveraging histology, spatial location, and gene expression. DeepST [9], a flexible deep learning model, offers various graph neural networks to integrate

* Corresponding author at: School of Artificial Intelligence, Jilin University, Changchun 130012, China.
** Corresponding author at: Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA.
*** Corresponding author.
E-mail addresses: wy6868@jlu.edu.cn (Y. Wang), hefe@umsystem.edu (F. He), xudong@missouri.edu (D. Xu).

https://doi.org/10.1016/j.csbj.2024.04.039
Received 31 December 2023; Received in revised form 14 April 2024; Accepted 15 April 2024
Available online 17 April 2024

morphological images, gene expression, and spatial location data for spatial transcriptome analysis. These methods integrate spatial multi-modal data using a deep learning model to derive spatially embedded features that capture underlying biological functions and spatial tissue patterns. A common strategy of these methods is to construct distances or adjacency relationships between spots by referencing features from morphology images for training models. However, current spatial transcriptomics technologies still have limitations and cannot generate comprehensive data covering the entire transcriptome and capturing the complete tissue area [10,11]. Spots on the spatial transcriptomic slide with a lower resolution cannot perfectly map the morphological image's corresponding position, such as the inconsistent alignment between panels (A) and (B) in Fig. 1 and Figs. S1–S3. This may lead to poor performance in related analyses. Such inconsistent alignments often result from low-quality staining images, spot shifting, or low capture efficiency.

One potential method to address this challenge is contrastive learning, an efficient self-supervised learning method for various representation learning tasks, such as image classification and natural language processing. This method is effective for learning representations through self-supervision [12]. SimCLR [13,14] employs data augmentation and learnable nonlinear transformations of the input's two views to learn embedding for image classification. MoCo [15,16] builds a large and consistent dictionary with a queue using a momentum updating mechanism for unsupervised representation learning. There are also clustering-based contrastive learning methods, such as DeepCluster [17] and SwAV [18]. Distillation-based methods like BYOL [19] and SimSiam [20] have also been proposed for self-supervised learning of effective visual representations. Because contrastive learning has a strong feature representation capacity that may better tolerate noises in mapping among multiple input modalities, exploiting a contrastive learning framework model to learn embeddings by integrating spatial multi-modal data is a promising approach for spatial domain identification.

This study presents a contrastive learning-based model, contrastive learning with convolutional neural network (GCN) and ResNet, short for ConGcR, for identifying spatial tissue architecture. Graph convolutional neural network is used as the encoder for learning features from gene expression [21]. ResNet [22] is used as the encoder for the hematoxylin-eosin (H&E) stained image patch. We also integrate ConGcR with a graph auto-encoder (GAE), resulting in a new method called contrastive learning with GAE and ResNet (ConGaR, for short), to further generate spatially embedded representations. Sixteen human brains, four chicken hearts, eight breast tumors, and 30 human lung spatial transcriptomics samples [1,23–25] involving multi-modal data are used to validate our method. According to the experimental results, the proposed contrastive learning-based models, ConGcR and ConGaR, can effectively integrate spatial multi-modal data, including gene expression, spatial location, and morphology images, to produce embedded representations for accurate spatial tissue architecture identification. The results also highlight our models' potential value in integrating multi-modal data into other spatial transcriptome analyses.

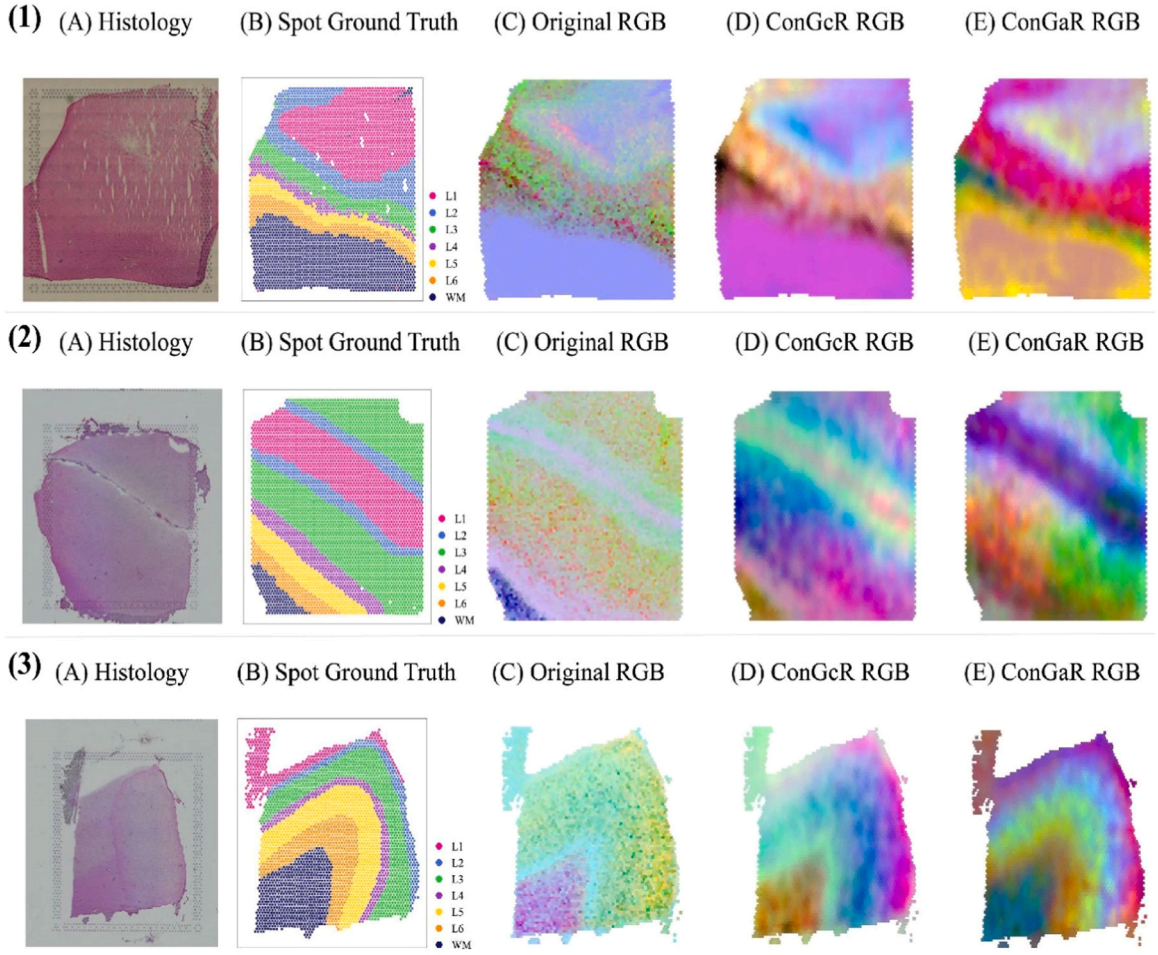

Fig. 1. Comparison results of (A) histology image, (B) spatial transcriptome spots colored by ground truth labels, (C) RGB image of original preprocessed gene expression with PCA method, (D) RGB image of ConGcR embedding with PCA method, and (E) RGB image of ConGaR embedding with PCA method in the subfigures **(1)**, **(2)** and **(3)** of samples 2–5, 151509 and 151675, respectively.

2. Materials and methods

The main workflow of ConGcR and ConGaR is used to accurately identify tissue architecture by integrating spatial multi-modal data, including tissue morphologies and transcriptomes. Fig. 2 shows ConGcR's and ConGaR's frameworks for generating spatially embedded representations by integrating morphological image, spatial location information, and gene expression. ConGcR and ConGaR identify the spatial domains in three steps: (A) data preprocessing on histological images, spatial locations, and gene expressions as the inputs, (B) generating proper embedding with effective spatial information using multi-modal data, and (C) identifying tissue architecture by clustering spatially embedded representations. The clustering result is applied to the comparison of different models for performance evaluation.

2.1. Data preprocessing

Our models take three types of spatial multi-modal data as input, that is, gene expression, spatial location, and morphology image. For gene expression data generated by different technologies, we use raw or log-transformed counts per million reads (LogCPM) [26] normalized expression to transform the data dimension and select the top 2000 variable genes as done in RESEPT [27]. The selected top 2000 highly variable genes can remove the most inessential gene expression and keep the most important gene information to generate an effective representation at the gene level. The spatial location data are used to create a spatial GCN graph by selecting the k-nearest neighbors to create edges. For each morphological image, we crop the image patches with the spot diameter in the full resolution provided by the technology platform before implementing ResNet. A square RGB (red, green and blue) patch with the same height and width is generated to match each spot for conducting contrastive learning.

2.2. Gene expression and image encoders

We use a classical GCN [28] as the gene expression encoder to extract features from the preprocessed gene expression $X \in \mathbb{R}^{n \times d}$, where n is the number of spots and d is the selected highly variable gene number. The gene expression encoder employs a two-layer GCN with message passing to derive a high-order graph representation Z*gene* as follows:

$$Z_{gene} = GCN\left(GCN\left(X_{gene}, A\right), A\right) \tag{1}$$

where X*gene* is the gene expression of highly variable genes, and A is the spatial adjacency matrix of the KNN graph. The GCN part is denoted as follows:

$$GCN\left(X_{gene}, A\right) = \tanh\left(\widehat{A} X_{gene} W^{(l)}\right) \tag{2}$$

where $\widehat{A}$ is the normalized adjacency matrix by $D^{-1/2}AD^{-1/2}$. D is a degree matrix generated from the KNN graph, and $W^{(l)}$ is the weight matrix of the l-th layer. For the learned graph embedding from gene expression, a simple fully connected (FC) neural network is applied to map the graph embedding into a shared space for contrastive learning, which is denoted as follows:

$$H_{gene} = FC\left(Z_{gene}\right) \tag{3}$$

The image encoder is used to learn the morphological features from the cropped image patch of each spot. We adopt ResNet-18 as the image feature extractor. Other different architecture ResNet models are also optional in our model setting. We modify the convolutional configurations of the first convolutional layer in ResNet, such as kernel size, stride,

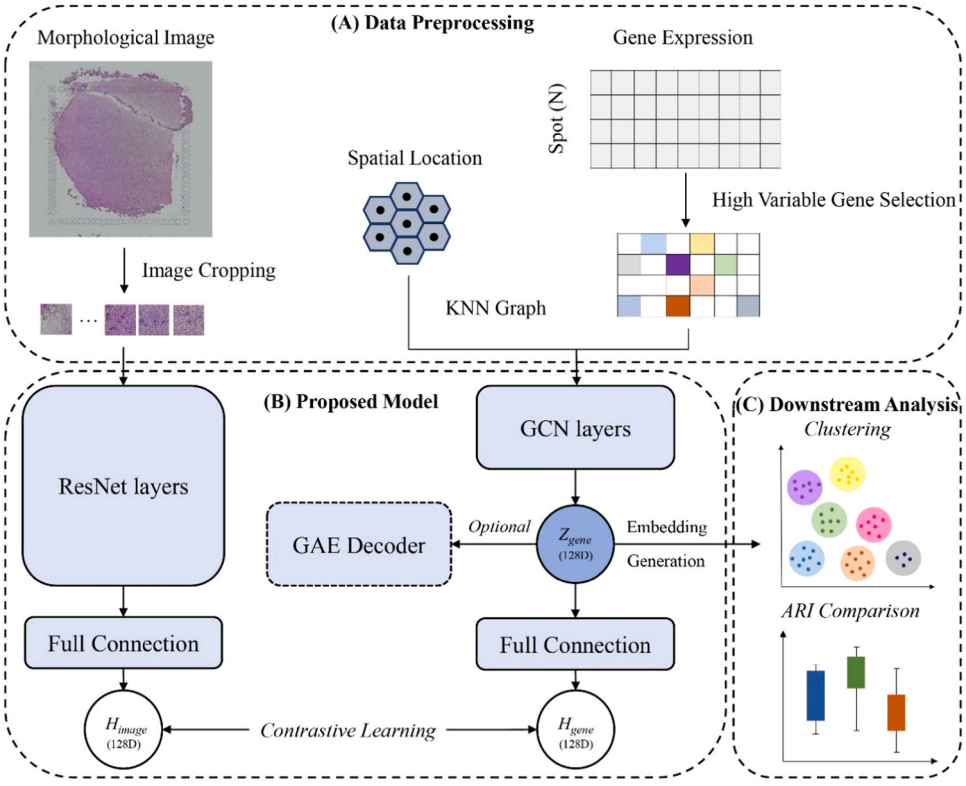

Fig. 2. The workflow of the proposed models ConGcR and ConGaR for tissue architecture identification with integrating spatial multi-modal data. The process details of the three experimental steps are shown in panels **(A)**, **(B)**, and **(C)**, respectively. (A) is the data preprocessing on the three types of spatial multi-modal data, that are histology image, spatial location, and gene expression. ConGcR and ConGaR take the preprocessed spatial multi-modal data in panel (A) as inputs. (B) shows the model architecture of our models to generate spatially embedded representations for identifying spatial domains. (C) details the downstream analysis tasks. The clustering method is applied to the generated embedding for tissue architecture identification, and ARI is used as the metric to evaluate the performance.

and padding settings, to better extract image feature details. Similarly, the output dimension of the last FC layer in ResNet is the same as the gene expression encoder for taking morphological features into the shared contrastive learning space.

2.3. Loss functions

To learn the joint embeddings that extract features from two modalities of gene expression and morphological image simultaneously, we employ NT-Xent loss in SimCLR [13] to conduct contrastive learning on the shared hypersphere space. Contrastive learning aims to pull together positive pairs between RNA and H&E representations of the same spot and push away negative pairs that do not match each other. NT-Xent loss, also called normalized temperature-scaled cross-entropy loss, guides the projected representations from the two modalities to learn the joint features. The mini-batch of M spots' gene expressions and M morphological features of the spots' corresponding cropped image patches are used to calculate each loss value. When a positive pair is defined in one batch, the other 2(M-1) spots' gene expressions or projected image features are taken as the negative samples. Then, the contrastive loss function for a positive pair of gene and image features of the same spot can be formed as follows:

$$L_{H^i_{gene},H^j_{image}} = -\log \frac{\exp\left(\frac{sim\left(H^i_{gene},H^j_{image}\right)}{\tau}\right)}{\Sigma^{2N}_{k=1} 1_{[k\neq i]} \exp\left(\frac{sim\left(H^i_{gene},H^k_{image}\right)}{\tau}\right)} \tag{4}$$

where $1_{[k\neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 if $k \ddagger i$, and τ is a temperature parameter used to control the degree of pushing away negative samples. $sim(H^i_{gene}, H^j_{image})$ means the dot product between l_2 normalized projected gene and image features, and it is denoted as follows:

$$sim\left(H^i_{gene}, H^j_{image}\right) = \frac{\left(H^i_{gene}\right)^T H^j_{image}}{\left\|H^i_{gene}\right\| \left\|H^j_{image}\right\|} \tag{5}$$

The final contrastive loss is calculated using all positive pairs with H^i_{gene}, H^j_{image} andH^j_{gene}, H^i_{image} within batch data.

To better learn the spatial topology and location relationship, we apply GAE on the graph embedding in the decoder part, resulting in a new ConGaR model. ConGaR formulates the model by not only using the morphological features for distillation learning but also making learned embeddings to maintain spatial adjacency relationships. The graph decoder is an inner product between graph embeddings by a sigmoid activation function:

$$\widetilde{A} = sigmoid\left(Z_{gene}\left(Z_{gene}\right)^T\right) \tag{6}$$

where $\widetilde{A}$ is the reconstructed adjacency matrix, and the decoder aims to minimize the cross-entropy L between the spatial adjacency matrix and the reconstructed adjacency matrix:

$$L(A,\widetilde{A}) = -\frac{1}{N\times N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(a_{ij} * \log\left(\widetilde{a}_{ij}\right) + \left(1-a_{ij}\right) * \log\left(1-\widetilde{a}_{ij}\right)\right) \tag{7}$$

where$a_{ij}$and$\widetilde{a}_{ij}$are the spatial and reconstructed adjacency matrix elements, respectively. N is the total number of spots in the KNN graph.

Then, the final total loss of our model is defined as follows:

$$L_{final} = \lambda_1 L_{contrastive} + \lambda_2 L_{graph-auto} \tag{8}$$

where λ_1 and λ_2 are the hyper-parameters to control the weights of contrastive learning and GAE losses in the final loss. ConGcR or ConGaR is iteratively trained to obtain the final graph embedding Z for the downstream analysis.

2.4. Dataset and experimental settings

This study uses 12 published and four private spatial multi-modal samples generated from the human brain [29–31] to validate our models. To further test model generalizability, four chicken heart spatial multi-modal samples are applied. The 12 published samples are named from 151507 to 151676. Four private samples are divided into non-AD cases at Braak stages I–II (samples 2–5 and 18–64) and early-stage AD cases at Braak stages III–IV (samples 2–8 and T4857). Four chicken heart samples (D4–D14) are profiled at four development stages from day 4 to day 14. The 16 human brain and four chicken heart datasets are generated from the 10X Genomics Visium platform. Table S1 provides more details. To enhance the model usage scenario, we also conduct experiments on a HER2-positive breast tumor dataset generated using spatial transcriptomics [32] and a human lung dataset generated using CosMx Spatially Molecular Imaging (SMI) [33,34]. The spatial transcriptomics dataset includes eight spatial multi-modal samples (A1–H1) with annotation labels. Tables S2 and S3 show the specific dataset details. The CosMx SMI dataset includes 30 spatial multi-modal samples (Fov1 to Fov30) with cell labels. Tables S4 and S5 provide further details. The proposed models, ConGcR and ConGaR, are implemented by Python 3.8.5 and Pytorch 1.13.0. They are trained on an NVIDIA TU102 [TITAN RTX] GPU. All the methods are conducted on a computing server running Ubuntu 18.04 operating system with 2.2 GHz, 144-core CPU, and 503 GB RAM.

For the 16 human brain multi-modal samples, GCN's gene expression encoder architecture has a dimension of 512 for the first layer, a dimension of 128 for the second layer, and an FC layer with two modalities with a dimension of 128. The number of k in the KNN graph is set as eight. For the four chicken heart samples and enhancement of the model usage scenario using spatial transcriptomics and CosMx SMI dataset, the hyper-parameter settings are kept the same as those used for the 16 samples—except that the first-layer dimension of GCN is 256 and the number of k in the KNN graph is four. Our model is trained by the Adam optimizer with a learning rate of 0.001. The temperature parameter in NT-Xent loss is 0.1, the weight of λ_1 is 1, and the weight of λ_2 is 100 in the final loss. We implement several training batch sizes, such as 64, 128, and 256.

To better show the potential spatial domain information in the embeddings of different methods, three embedding dimensional reduction methods of PCA, t-SNE, and UMAP [35–37] are applied to transform the embeddings into three-dimensional features used to generate RGB images in RESEPT. The parameter settings are all set as the default in the Python package sklearn. Differential expression gene (DEG) analysis is conducted using Python package scanpy and the cluster labels of ConGcR. Based on these differentially expressed genes, the enrichment analysis of GO terms (Biological Process) is performed by the R package clusterProfile using the enrichGO function.

2.5. Evaluation metrics

We apply the K-means algorithm [38] for clustering on the embeddings learned from different methods to identify spatial tissue architecture. K-means is the most widely used clustering algorithm in spatial transcriptome analysis, and it has been widely used in spatial methods, such as BayesSpace, SpaGCN and RESEPT [5,6,27]. The number of spatial domains is set as the number of ground truth labels in each sample. After clustering, we use the adjusted Rand index (ARI) [39] to evaluate clustering results. ARI is used to measure the similarity between two partitions. This index reflects the consistency between the clustering labels and the ground truth spatial domains as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}} \quad (9)$$

where n_{ij} means the common samples in both i-th predicted label and j-th ground truth label, and a_i and b_j are the number of samples in the i-th predicted label and the j-th ground truth label, respectively. The index range of ARI is [− 1,1], and a larger ARI means a higher consistency. n_{ij} means the common samples in both i-th predicted label and j-th ground truth label. a_i and b_j are the number of samples in the i-th predicted label and the j-th ground truth label, respectively. The index range of ARI is [− 1,1], and a larger ARI means a higher consistency.

To evaluate RGB image quality, the three metrics of mean square error (MSE), peak signal-to-noise ratio (PSNR), and structure similarity index measure (SSIM) are used [40–42]. MSE is applied to calculate the pixel difference at each position of two RGB images. A smaller MSE value means a higher RGB image quality. PSNR is commonly used to assess the similarity between the color distribution of two RGB images based on the MSE metric as follows:

$$PSNR = 10 * \log_{10}\frac{L^2}{MSE} \quad (10)$$

where L is the maximum pixel value in an RGB image, and the larger the PSNR value, the higher the RGB image quality. SSIM is another useful RGB image quality assessment metric to consider all three image aspects, that is, luminance, contrast, and structure. It focuses more on the influence of structural information, especially the local image information of each region, to assess the RGB image quality. When SSIM calculates the difference between two images at each location, it uses the pixels from a region in the image rather than taking a single pixel from each position to derive the evaluation metric, as done in PSNR and MSE. A larger SSIM value means a higher similarity between two RGB images.

2.6. Benchmark methods

For the 16 human brain samples, we compare ConGcR and ConGaR with the original clustering methods, the baseline methods of integrating spatial multi-modal data, and the models of BayesSpace, SpaGCN, DeepST, and conST [43]. For the raw gene expression, we adopt the same experimental procedures as the data preprocessing in ConGcR and ConGaR. For the raw morphological image, the pixel values in a cropped square patch are flattened into one dimension, and we project them into 2000 dimensions using the PCA method. Then, min-max normalization is used to map the features of the two modalities into the same scale. We use K-means clustering directly on these features of the two modalities. We add and concatenate the preprocessed gene expression and morphological features as the baseline integration methods. We consider gray-scale and RGB images for each piece of H&E data in this study. For the BayesSpace, SpaGCN, DeepST and conST models, we keep the default hyper-parameter model settings, use the same data preprocessing in ConGcR and ConGaR, and apply K-means as the initial method in the training process for BayesSpace and SpaGCN to identify spatial domains. For testing the model's generalizability on four chicken heart samples, the original and baseline method settings are kept the same as those used in the model validation on 16 samples—except that the number of selected highly variable genes and PCA projected dimensions are both 500. For enhancing model usage scenario on eight HER2-positive breast tumor and 30 human lung samples, all the default hyper-parameter model settings are used for the BayesSpace, SpaGCN, DeepST, and conST models.

3. Results

We evaluate our models' effectiveness and generalizability using 16 human brains' and four chicken hearts' multi-modal data from 10X Genomics Visium. An eight ST breast tumor and 30 CosMx SMI lung multi-modal dataset are used to enhance model usage scenarios. The experimental results show that ConGcR and ConGaR not only can successfully integrate spatial multi-modal data but also can be applied and generalized to the new test data for spatial tissue architecture identification.

3.1. Model validation using 16 human brain spatial multi-modal samples

Table 1 and Table S6 show the tissue architecture identification comparison results of different methods on the spatial multi-modal dataset of 16 human brain samples using raw and LogCPM normalized gene expression, respectively. To better show our models' identification capacity, we conduct the ConGcR and ConGaR experiments with different hyper-parameter settings for batch size and epoch number. Tables S7 and S8 show the comparison results. Tables S9 and S10 show the experimental results of GCN as a gene expression encoder for our contrastive learning model. Table S11 shows the comparison results of the individual encoders of ResNet on morphological image patches and GCN on spatial gene expression. Tables S12 and S13 show the comparison results of GAE and ConGaR to demonstrate the effectiveness of using histological image data. Tables S14 and S15 compare the BayesSpace, SpaGCN, DeepST, conST, ConGcR, and ConGaR models.

As Tables 1 and S6 show, our mdoels ConGcR and ConGaR have a higher ability to identify spatial tissue architecture than the methods of original clustering on the features of gene expression or H&E data and the baseline integration methods of adding or concatenating the features of two modalities. The contrastive learning-based models can be useful for integrating multi-modal data of morphological images and gene expressions with spatial locations in this study. ConGcR can obtain competitive ARIs higher than 0.5, such as sample 2–5. After integrating the GAE framework and the spatial location information, ConGaR outperforms ConGcR in most cases. Adding or concatenating the two modalities' features can effectively help with spatial domain identification. This demonstrates that the preprocessed features of the two modalities contain the potential information for mutual assistance and benefit. Compared to the integration method of adding features, concatenating features can be more helpful in both cases of adopting a morphological image in gray-scale or RGB format. On sample 151507, the performance of concatenating features can improve by at least 10% compared to that of the original gene expression in all cases. The performance is also better than that of both original features of gene expression and H&E data. The features in RGB images are more effective for integrating gene expression to generate embeddings. As for the original H&E features, their ability to identify spatial tissue architecture is weaker. Some original H&E features contain mostly noises, such as samples 151670 and 151671.

Tables S7 and S8 compare several batch sizes (64, 128, and 256) where models are trained for different numbers of epochs. According to the overall average ARI, ConGcR obtains the optimal value when the batch size is 64 and 128 using raw and LogCPM normalized gene expression, respectively. ConGaR has the highest capacity to identify spatial domains when batch size is 256 or 128 in two corresponding cases. ConGaR can obtain superior performance compared to ConGcR. As for ConGcR, some samples (151508, 151509,151510, 2–5, and T4857) benefit from the smaller training batch size and can obtain the highest ARI at batch size 64. As Tables S9 and S10 show, ConGcR can be superior to the contrastive learning model (ConMR) using MLP as the gene expression encoder. ConGcR outperforms ConMR in identifying spatial domains when different batch sizes are applied in most cases. This shows that the GCN model involving spatial location information can be more suitable for extracting features from gene expression for

Table 1
Tissue architecture identification comparison results of original preprocessed gene expression and H&E methods, spatial multi-modal data integration baseline methods, ConGcR and ConGaR using raw gene expression on 16 human brain samples. The best performances of each sample and overall average ARI are marked in bold.

Sample	*RNA_Original*	*H&E_Original (Gray/RGB)*	*RNA_H&E_Add (Gray/RGB)*	*RNA_H&E_Concatenate (Gray/RGB)*	*ConGcR*	*ConGaR*
151507	0.109	0.111/0.148	0.145/0.124	0.143/0.149	0.374	**0.477**
151508	0.118	0.267/0.267	0.218/0.215	0.204/0.216	0.304	**0.399**
151509	0.164	0.193/0.297	0.204/0.240	0.231/0.202	0.412	**0.432**
151510	0.131	0.220/0.240	0.199/0.225	0.219/0.221	0.335	**0.481**
151669	0.125	0.069/0.093	0.118/0.115	0.113/0.118	**0.330**	0.285
151670	0.139	0.000/0.000	0.147/0.151	0.142/0.155	**0.346**	0.284
151671	0.151	0.000/0.000	0.157/0.152	0.154/0.150	0.392	**0.431**
151672	0.130	0.000/0.015	0.130/0.131	0.130/0.131	0.398	**0.488**
151673	0.170	0.160/0.163	0.193/0.204	0.199/0.205	0.337	**0.472**
151674	0.170	0.154/0.156	0.185/0.188	0.184/0.190	0.325	**0.388**
151675	0.174	0.190/0.191	0.211/0.231	0.234/0.230	0.297	**0.430**
151676	0.190	0.171/0.163	0.225/0.220	0.223/0.220	**0.427**	0.406
18_64	0.088	0.209/0.209	0.124/0.127	0.127/0.126	0.396	**0.406**
2_5	0.548	0.301/0.257	0.541/0.539	0.540/0.539	**0.714**	0.654
2_8	0.315	0.103/0.097	0.348/0.343	0.341/0.349	0.324	**0.538**
T4857	0.206	0.133/0.135	0.235/0.238	0.234/0.235	0.445	**0.513**
Ave±Std	0.183 ± 0.11	0.143 ± 0.09/0.152 ± 0.09	0.211 ± 0.10/0.215 ± 0.10	0.214 ± 0.10/0.215 ± 0.10	0.384 ± 0.10	**0.443**± 0.09

contrastive learning.

As Table S11 shows, compared to individual ResNet or GCN encoders, better identification results can be obtained when contrastive learning is applied in our models. The GCN encoder using spatial gene expression has a higher capacity than the ResNet encoder in regard to using morphological image patches for spatial domain identification. ConGaR is the best method in both cases of using raw and LogCPM normalized gene expression. Even if the overall average ARI of GCN encoder is higher than ConGcR in the case of using LogCPM normalized gene expression, ConGcR has higher ARIs than the GCN encoder in more samples. The better capacity of ConGcR to identify spatial domains can be shown when using other hyper-parameter settings in Table S8. Tables S12 and S13 show GAE's comparison results with the weight of λ_1 being 0. When histological image data are applied to the contrastive learning-based model, ConGaR has a higher capacity to identify spatial domains, and the highest overall average ARI can be obtained when the batch size settings are 256 and 128 in the case of using raw and LogCPM normalized gene expression, respectively. The histological image is effectively used in our model to integrate spatial multi-modal data. As Tables S14 and S15 show, ConGaR outperforms the BayesSpace, SpaGCN, DeepST, conST, and ConGcR models. Compared to other models, it obtains the highest overall average ARI and has more samples with higher identification ability in both tables. Further, ConGcR has a higher ability than SpaGCN and DeepST to identify spatial tissue architecture. In seven samples, ConGaR obtains the highest ARI among six comparison models.

3.2. Application of spatially resolved transcriptomics samples to the new test dataset

To further evaluate our models' spatial domain identification capacity on the new test dataset, we conduct the same experimental comparisons using 16 human brain samples on the new spatial multi-modal dataset consisting of four chicken heart samples. Table 2 and Table S16 show the tissue architecture identification comparison results of different methods on the new test dataset using raw and LogCPM normalized gene expression, respectively. Tables S17 and S18 show the experimental results of ConGcR and ConGaR with different hyper-parameter settings of batch size and epoch number.

Tables 2 and S16 show that our models can be effectively applied to the new test 10X Visium dataset to integrate spatial multi-modal data and develop a higher capacity to identify spatial domains than other comparison methods. ConGcR obtains the highest ARI among comparison methods on sample D7 in Table 2. After integrating the GAE framework with learning spatial location information, ConGaR outperforms ConGcR in most cases. The baseline methods can still improve the identification performance based on the two original features of gene expression and H&E data. Compared to adding features, concatenating features of the two modalities is more useful for downstream analysis. The baseline methods of adding and concatenating features are more effective when the morphological images of gray-scale and RGB formats are used, respectively. As for the original preprocessed features, a single H&E feature of the morphological image has no obvious usefulness, and the LogCPM normalized feature of gene expression contributes more significantly than the raw feature to identifying spatial domains.

As Tables S17 and S18 show, our models can obtain the optimal value when the batch size is 256, and the highest overall average ARI can be obtained by ConGaR in both tables. The larger training batch size is helpful for ConGcR and ConGaR on the four chicken heart samples. Like the 16 human brain samples, ConGaR also has a higher capacity than ConGcR to identify spatial tissue architecture when different batch sizes are used in most cases.

Table 2
Model generalizability test by tissue architecture identification comparison results of original preprocessed gene expression and H&E methods, spatial multi-modal data integration baseline methods, ConGcR and ConGaR using raw gene expression on the new dataset of four chicken heart samples. The best performances of each sample and overall average ARI are marked in bold.

Sample	*RNA_Original*	*H&E_Original (Gray/RGB)*	*RNA_H&E_Add (Gray/RGB)*	*RNA_H&E_Concatenate (Gray/RGB)*	*ConGcR*	*ConGaR*
D4	**0.227**	0.124/0.112	0.171/0.165	0.171/0.173	0.137	0.143
D7	0.109	0.078/0.078	0.113/0.113	0.117/0.116	**0.371**	0.267
D10	0.043	0.067/0.070	0.086/0.094	0.089/0.096	0.243	**0.275**
D14	0.055	0.111/0.114	0.094/0.088	0.089/0.090	0.212	**0.380**
Ave ± Std	0.108 ± 0.08	0.095 ± 0.03/0.093 ± 0.02	0.116 ± 0.04/0.115 ± 0.03	0.116 ± 0.04/0.119 ± 0.04	0.241 ± 0.08	**0.266** ± 0.08

3.3. Analysis of case study results

We select samples 2–5 and 151509, which show promising results, and samples 151675 and 2–8, which show poor results, to further analyze the effectiveness and limitations of integrating methods using spatial multi-modal data. Fig. 3 shows the tissue architecture identification results of model ConGcR with the hyper-parameter settings of raw gene expression and 64 batch sizes for the four samples. Fig. S4 shows the same identification results for the 16 samples. In Fig. 1, we select samples 2–5, 151509 and 151675, which are the best, general, and worst identification results in the four samples, to illustrate the comparison results of the original histological image, spatial transcriptome spots colored by ground truth labels, RGB image of original preprocessed gene expression with PCA method, and RGB images of ConGcR and ConGaR embeddings from the gene expression encoder with the PCA method. For samples 2–8, Fig. S1 shows the relatively poor identification results for the four samples. The same comparison results using t-SNE and UMAP methods on these four samples are shown in Figs. S2 and S3, respectively. To select the most effective dimensional reduction method, we evaluate the RGB image quality between model embedding and original preprocessed embedding using the PSNR, SSIM, and MSE assessment metrics. Tables S19–S22 detail the RGB image quality comparison results of using PCA, t-SNE, or UMAP for transforming different embeddings into three-dimensional features on samples 2–5, 151509, 151675, and 2–8, respectively.

As Fig. 3 illustrates, the tissue architecture identifications of samples 151509 and 2–5 are significantly more accurate than those of samples 151675 and 2–8. ARI improves gradually as the number of training epochs on samples 2–5 and 151675 are increased. Within several epochs, the highest ARI can be obtained, after which the changes become stable in the training process for samples 151509 and 2–8. Additionally, when ConGcR is trained for five epochs on samples 2–5, it can generate representative embedding with spatial domain identification of ARI higher than 0.7. Most of the 16 samples have the same overall trend: that is, the results become relatively stable after five epochs. Meanwhile, a few samples have a decreasing trend, as Fig. S4 shows.

As Fig. 1 shows, the degrees of consistency between the original morphological image and the spatial transcriptome spot with the ground truth labels in panels (A) and (B) are stronger on samples 2–5. The general and low consistency degrees between panels (A) and (B) are illustrated on samples 151509 and 151675, respectively. The color and texture changes align with some spot labels at the layer levels on the samples with promising results, such as in the layer of L1 and WM on samples 2–5 and in the layer of L1 on sample 151509. However, the ground truth spot labels cannot be appropriately aligned with the image color and texture changes on sample 151675 with poor identification results. For example, there is a vertical texture in the middle part of the H&E image. Generally, there are some blank areas across the transcriptome layer labels in panel (B) of the spot ground truth. As panels (A) and (B) of Fig. S1 show, the blank areas are large and continuous at the intersection of L5, L6, and WM on samples 2–8. The promising identification results benefit from the higher alignment of the color and texture layout in the histology and ground truth labels of the transcriptome layer, which allows for different data integration methods to extract mutually supportive and effective features from the two modalities. The data qualities of the morphological images and the gene expression of spatial transcriptome spots in some cases limit our current models, requiring further research.

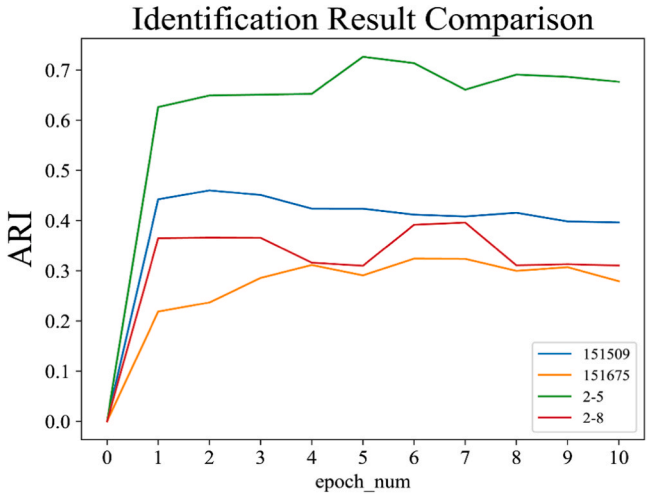

Fig. 3. Tissue architecture identification comparison of ARI changes with a number of epochs in the hyper-parameter setting of batch size 64 using raw gene expression on samples 151509, 151675, 2–5 and 2–8.

As panels (C), (D), and (E) of Figs. 1 and S1 show, the RGB images of original preprocessed embedding, ConGcR embedding, and ConGaR embedding contain roughly similar patterns and texture layouts with the transcriptome spots colored by the ground truth. Compared to the original RGB image, our models' RGB images have less noise and fragmented areas with smoother color contrast and clearer patterns. They benefit from the contrastive learning framework by integrating histology information and gene expression, as well as the encoder architecture of GCN with message passing. The more accurate spatial domain identification of generative embedding with higher ARI is, the more obvious the pattern the RGB image contains (e.g., the images in panel [E] of sample 2–5). These figures achieve similar results using dimensional reduction methods of t-SNE and UMAP.

Tables S19–S22 show that RGB images generated using the PCA method to transform embeddings into three-dimensional features have the highest quality among the RGB images generated by three different dimensional reduction methods. The PSNR and SSIM values between the original RGB and our models' RGB of PCA are larger than the t-SNE or UMAP values. The MSE values between the RGB images using PCA are smaller than the values using t-SNE or UMAP. By applying PCA, the RGB images generated from ConGcR and ConGaR are closer to the original embedding RGB images. Although the embeddings of our models integrate the features of the two modalities, they preserve more original information when PCA is used. Therefore, PCA is the most effective dimensional reduction method for transforming embedding into three-dimensional features to generate RGB images in this study.

3.4. Model usage comparison on spatial transcriptomics and CosMx SMI datasets

To enhance our model usage scenario, we compare the spatial domain identification performances of ConGcR and ConGaR with the other four models (BayesSpace, SpaGCN, DeepST, and conST) on the HER2-positive breast tumor dataset of spatial transcriptomics technology and human lung dataset of CosMx SMI technology. The details of the dataset and different model hyper-parameter settings are provided in Section 2.4 *Dataset and experimental settings*. Tables S23 and S24 show the specific identification comparison results for eight HER2-positive breast tumor and 30 human lung samples. Figs. S5–S7 show the learned labels with spatial coordinates for the tissue architecture identification results of the six comparison models on the three spatial transcriptomics samples. We also conduct the DEG analysis based on these labels of ConGcR and list the differentially expressed genes of each cluster label in Supplementary Data 1–3. The enrichment analysis results based on the DEG lists are detailed in Supplementary Data 4–6. Both DEG and enrichment results are filtered out if the adjusted p-value is greater than 0.05.

As Table S23 shows, by adopting contrastive learning to integrate gene expression, morphological image, and spatial information, ConGcR has the highest capacity to identify spatial domains. The highest overall average ARI and more stable results with lower standard deviation can

be obtained by our models among all comparison models. Compared to the BayesSpace and conST models, ConGaR demonstrates better identification capacity. Even if the overall average ARI of ConGaR is lower than SpaGCN, ConGaR can obtain a higher identification capacity on five samples of A1, E1, F1, G2, and H1. The highest ARI among all comparison models can be obtained by ConGaR on two samples of F1 and G2. Our models' superiority in identifying spatial domains can also be demonstrated in Table S24. ConGaR can outperform other comparison models and obtain the highest overall average ARI. ConGcR has a better capacity than BayesSpace, SpaGCN and conST. It can obtain the highest ARI among all comparison models, and the ARI values are larger than 0.2 in the samples of Fov9, Fov15, and Fov25.

As Figs. S5–S7 illustrate, compared to other models, the spatial patterns detected by ConGcR and ConGaR are closer to the ground truth categories. The connective tissue region in ground truth can be depicted well by the labels of L2 in ConGcR, whereas most of the other models fail to distinguish connective tissue and invasive cancer regions according to Fig. S5. Fig. S6 demonstrates that the labels of ConGcR can obtain a similar cluster pattern to the invasive cancer region in the ground truth. There are six ground truth categories with corresponding regions in sample H1. Our models identify similar biological regions and accurate layouts, such as the breast glands region in the ground truth of Fig. S7.

Supplementary Data 1–6 show the identified differentially expressed genes and enriched descriptions. As for the connective tissue regions in sample D1, the most significantly expressed gene based on the ConGcR labels of L2 is ACTG1, which participates in various types of cell movements [44]. As for the immune infiltration region in sample E1, the gene ERBB2 identified by the ConGcR labels of L3 demonstrates the malignancy of cancer cells, especially in breast cancer [45]. Additionally, breast cancer patients generally have a higher proportion of B cells, but this is highly variable [46]. In the invasive cancer region of sample E1, the results show that B cell-related functions are intensively enriched, such as B cell-mediated immunity and B cell activation. Therefore, the learned cluster label of our model can help further interpret the biological region functions in spatial transcriptome analysis.

4. Discussion and conclusions

This study proposes a contrastive learning-based model, ConGcR, that integrates spatial multi-modal data to accurately identify tissue architecture in spatial transcriptome analysis. Graph convolutional and ResNet neural networks are used as the encoders for gene expression and morphological image data, respectively. NT-Xent loss is applied as the model loss function in ConGcR. After integrating the cross-entropy loss of GAE into ConGcR, a new contrastive learning-based model, ConGaR, is used to further identify the spatial domain. Both ConGcR and ConGaR can generate effective embeddings containing promising spatial information for domain identification. Sixteen human brain spatial multi-modal samples are used to validate our models. Directly clustering, addition, and concatenation of the preprocessed features of the two modalities are used as original clustering and baseline integration methods for comparison purposes. We also compare ConGcR and ConGaR with the four models of BayesSpace, SpaGCN, DeepST and conST. Further, four chicken heart spatial multi-modal samples are used to further test model generalizability on the new dataset. In the case study, we conduct a detailed analysis of method effectiveness and limitations in integrating spatial multi-modal data by the application of our models on four samples with promising or poor results. Eight breast tumor and 30 human lung spatial multi-modal samples are used to enhance the model usage scenario.

The study demonstrates that our models can obtain superior performance compared to the original clustering method, the baseline integration methods, and the four models of BayesSpace, SpaGCN, DeepST, and conST on 16 human brain samples. After employing the cross-entropy loss of GAE, ConGcR further learns spatial location information, and higher ARIs can be obtained. Compared with the original preprocessed features of gene expression and morphological image in the original method, adding and concatenating two modalities' features in baseline methods can generate more useful features in most cases. In terms of the two baseline methods, the concatenation method obtains the more accurate predicted labels with ground truth. These results suggest that concatenated features are better suited for identifying spatial tissue architecture. Compared with BayesSpace, SpaGCN, DeepST, and conST, ConGaR can achieve the most superior performance. The competitive results can also be demonstrated in the test of model generalizability on the four chicken heart samples, where our models, as well as the baseline methods, can effectively integrate multiple profiles of spatial transcriptomics data for tissue architecture identification. Our models obtain the highest ARI in most cases. Additionally, our models can be effectively used for the HER2-positive breast tumor dataset of spatial transcriptomics technology and the human lung dataset of CosMx SMI technology. They can outperform the other comparison models and better help identify the biological regions. The learned spatial domain label is useful for further interpreting the biological function of each region through DEG and GO enrichment analyses. In conclusion, ConGcR and ConGaR can generate more accurate embeddings using spatial multi-modal data to identify tissue architecture in spatial transcriptome analysis.

Most existing data integration methods use the extracted histology features to establish adjacency relationships without updating the trained model to adopt the two modalities of morphological image and gene expression simultaneously for identifying spatial domains, such as SpaGCN and DeepST. This study's main novelty lies in its use of three profiles—that is gene expression, spatial location, and morphological image—and applying contrastive learning to distill features from gene expression and H&E data for training our models. The encoders of GCN and ResNet are used to extract the representative features for the common space of gene expression and morphological image. Two modalities' features can be fully integrated, and the explicit noises in gene expression can be efficiently removed in the shared hypersphere space through contrastive learning. The RGB images generated from three-dimensional features are useful for better displaying the potential spatial architecture and pattern in different embeddings and showing the effective function of contrastive learning in our models.

In ConGcR and ConGaR, the types of image encoders and the number of k-nearest neighbors of the spatial graph in GCN can be adjusted according to different cases. Compared to the individual encoders of GCN or ResNet, ConGcR and ConGaR achieve a higher identification capacity by implementing contrastive learning. Compared to GAE, ConGaR can effectively involve histological image data to learn more accurate features for identifying spatial domains. Compared to ConGcR, ConGaR can enhance the spatial location relationship by using the loss function of GAE, but it has a more complex model and requires more computational resources. ConGcR model is more suitable when the morphological image and transcriptome slide are better aligned; otherwise, ConGaR model can be a better choice. Because of the comprehensive representation of spatial multi-modal data and model flexibility, ConGcR and ConGaR can be extended to various applications using different types of spatial transcriptomics data.

Although ConGcR and ConGaR obtain competitive performance in identifying spatial domains compared to the other multi-modal data integration methods in this study, the best ARIs of the two models are not superior enough and some cases show poor results, such as samples 151672, 151675, and 2–8. There is still potential to improve the model's capacity to learn valuable features with spatial information for identifying spatial domains. The limitation of existing models is that the alignment and consistency of the morphological images and spatial transcriptome spots with some missing parts influences the exploration of the spatial domain, which may reduce the effectiveness of extracting and learning features from both modalities. To address this limitation, methods to sufficiently apply the morphological image with complete

modality information need to be explored. Future research will improve the model design in mutual learning features using gene expression and H&E data.

CRediT authorship contribution statement

Yu Lin: Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yanchun Liang:** Data curation, Formal analysis, Funding acquisition, Project administration. **Duolin Wang:** Data curation. **Yuzhou Chang:** Data curation. **Qin Ma:** Conceptualization, Methodology, Writing – review & editing. **Yan Wang:** Data curation. **Fei He:** Data curation. **Dong Xu:** Conceptualization, Methodology, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62372494 and 62072212), Development Project of Jilin Province of China (No. 20220508125RC), National Key R&D Program (No. 2018YFC2001302), Jilin Provincial Key Laboratory of Big Data Intelligent Cognition (No. 20210504003GH), Guangdong Universities' Innovation Team (No. 2021KCXTD015), Key Disciplines Projects (No. 2021ZDJS138), and the Paul K. and Diane Shumaker Endowment Fund at the University of Missouri.

Data and Code Availability

The 12 published human brain samples generated by 10X Visium can be accessed at http://research.libd.org/spatialLIBD. The four private human brain samples (2–5, 2–8, 18–64, T4857) of non-AD or AD cases are available upon request. The four published chicken heart samples generated by 10X Visium can be accessed at https://github.com/madhavmantri/chicken_heart. The eight published HER2-positive breast tumor samples generated by spatial transcriptomics technology can be accessed at https://github.com/almaan/her2st. The 30 published human lung samples generated by CosMx SMI technology can be accessed at https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/. Last, the source code with the application demo is freely available on GitHub at https://github.com/YuLin-code/Spatial-ConGR.

Supplementary Files

The supplementary figures and tables in this study can be found in the Supplementary Materials section. Supplementary Data 1–3 are the DEG lists based on the cluster labels of ConGcR on samples D1, E1, and H1, respectively. Supplementary Data 4–6 are the enrichment analysis results based on DEG lists on samples D1, E1, and H1, respectively.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.039.

References

[1] Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci 2021;24(3):425–36.
[2] Burgess DJ. Spatial transcriptomics coming of age. Nat Rev Genet 2019;20(6). 317-317.
[3] Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science 2018;362(6416):792.
[4] Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348(6233). 412-412.
[5] Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol 2021;39(11):1375–84.
[6] Hu J, Li X, Coleman K, Schroeder A, Ma N, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods 2021;18(11): 1342–51.
[7] Pham D.T., Tan X., Xu J., Grice L.F., Lam P.Y., et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv 2020: 2020.2005.2031.125658.
[8] Zuo, Zhang C, Cao Y, Feng C, Jiao J, et al. Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. Nat Commun 2022;13(1). 5962-5962.
[9] Xu C, Jin X, Wei S, Wang P, Luo M, et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. Nucleic Acids Res 2022;50(22). e131-e131.
[10] Hu J, Schroeder A, Coleman K, et al. Statistical and machine learning methods for spatially resolved transcriptomics with histology. Comput Struct Biotechnol J 2021;19:3829–41.
[11] Zhang D, Schroeder A, Yan H, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. Nat Biotechnol 2024:1–6.
[12] Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. Int Conf Mach Learn (PMLR) 2020: 9871–81.
[13] Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. Int Conf Mach Learn (PMLR) 2020:1597–607.
[14] Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G. Big Self-Supervised Models are Strong Semi-Supervised Learners. Adv Neural Inf Process Syst 2020;33: 22243–55.
[15] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. Proceedings of. IEEE/CVF Conf Comput Vis Pattern Recognit 2020:9726–35.
[16] Chen X., Fan H., Girshick R., He K. Improved baselines with momentum contrastive learning. arXiv preprint 2020: arXiv:2003.04297.
[17] Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. Proc Eur Conf Comput Vis (ECCV) 2018:132–49.
[18] Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, et al. Unsupervised learning of visual features by contrasting cluster assignments. Adv Neural Inf Process Syst 2020;33:9912–24.
[19] Grill JB, Strub F, Altché F, Tallec C, Richemond P, et al. Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 2020;33: 21271–84.
[20] Chen X, He K. Exploring simple siamese representation learning. Proceedings of. IEEE/CVF Conf Comput Vis Pattern Recognit 2021:15750–8.
[21] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification. IEEE Trans Pattern Anal Mach Intell 1996;18(6):607–16.
[22] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of. IEEE Conf Comput Vis Pattern Recognit 2016:770–8.
[23] Chen S., Chang Y., Li L., Acosta D., Morrison C., et al. Spatially resolved transcriptomics reveals unique gene signatures associated with human temporal cortical architecture and Alzheimer's pathology. bioRxiv 2021: 2021.2007.2007.451554.
[24] Mantri M, Scuderi GJ, Abedini-Nassab R, Wang MF, McKellar D, et al. Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. Nat Commun 2021; 12(1). 1771-1771.
[25] 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(1):6012.
[26] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–87.
[27] Chang Y, He F, Wang J, Chen S, Li J, et al. Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning. Comput Struct Biotechnol J 2022;20:4600–17.
[28] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint. arXiv 2016:02907.
[29] Beach TG, Adler CH, Sue LI, Serrano G, Shill HA, Walker DG, et al. Arizona study of aging and neurodegenerative disorders and brain and body donation program. Neuropathology 2015;35:354–89.
[30] Vonsattel JP, Del Amaya MP, Keller CE. Twenty-first century brain banking. Processing brains for research: the Columbia University methods. Acta Neuropathol 2008;115(5):509–32.
[31] Navarro JF, Croteau DL, Jurek A, Andrusivova Z, Yang B, et al. Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. iScience 2020;23(10): 101556.
[32] Ståhl PL, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016;353(6294):78–82.

[33] Garrido-Trigo A, Corraliza AM, Veny M, et al. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. Nat Commun 2023;14(1):4506.
[34] Bill R, Wirapati P, Messemaker M, et al. CXCL9: SPP1 macrophage polarity identifies a network of cellular programs that control human cancers. Science 2023;381(6657):515–24.
[35] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374(2065). 20150202-20150202.
[36] Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008; 9:2579–625.
[37] Mcinnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. J Open Source Softw 2018;3(29):861.
[38] MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab 1967;1(14):281–97.
[39] Hubert L, Arabie P. Comparing partitions. J Classif 1985;2(1):193–218.
[40] Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. Electron Lett 2008;44(13):800–1.
[41] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.
[42] Thung K.H., Raveendran P. A survey of image quality measures. Proceedings of the 2009 International Conference for Technical Postgraduates (TECHPOS) 2009.
[43] Zong Y., Yu T., Wang X., et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. bioRxiv 2022:2022.01.14.476408.
[44] Liu J, Liu H, Zhao Z, et al. Regulation of Actg1 and Gsta2 is possible mechanism by which capsaicin alleviates apoptosis in cell model of 6-OHDA-induced Parkinson's disease. Biosci Rep 2020;40(6). BSR20191796.
[45] Bertucci F, Borie N, Ginestier C, et al. Identification and validation of an ERBB2 gene expression signature in breast cancers. Oncogene 2004;23(14):2564–75.
[46] Tsuda B, Miyamoto A, Yokoyama K, et al. B-cell populations are expanded in breast cancer patients compared with healthy controls. Breast Cancer 2018;25:284–91.