conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics

Yongshuo Zong[1], Tingyang Yu[2, 3, 4], Xuesong Wang[2], Yixuan Wang[2, 5], Zhihang Hu[2] and Yu Li*[2, 6]

[1]School of Informatics, the University of Edinburgh, Edinburgh, United Kingdom
[2]Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong SAR, China
[3]Department of Mathematics, Chinese University of Hong Kong, Hong Kong SAR, China
[4]Department of Information Engineering, Chinese University of Hong Kong, Hong Kong SAR, China
[5]Department of Mathematics, Harbin Institute of Technology, Weihai, 264209, China
[6]The CUHK Shenzhen Research Institute, Hi-Tech Park, Nanshan, Shenzhen, 518057, China

Abstract

Motivation: Spatially resolved transcriptomics (SRT) shows its impressive power in yielding biological insights into neuroscience, disease study, and even plant biology. However, current methods do not sufficiently explore the expressiveness of the multi-modal SRT data, leaving a large room for improvement of performance. Moreover, the current deep learning based methods lack interpretability due to the "black box" nature, impeding its further applications in the areas that require explanation.

Results: We propose conST, a powerful and flexible SRT data analysis framework utilizing contrastive learning techniques. conST can learn low-dimensional embeddings by effectively integrating multi-modal SRT data, *i.e.* gene expression, spatial information, and morphology (if applicable). The learned embeddings can be then used for various downstream tasks, including clustering, trajectory and pseudotime inference, cell-to-cell interaction, *etc.* Extensive experiments in various datasets have been conducted to demonstrate the effectiveness and robustness of the proposed conST, achieving up to 10% improvement in clustering ARI in the commonly used benchmark dataset. We also show that the learned embedding can be used in complicated scenarios, such as predicting cancer progression by analyzing the tumour microenvironment and cell-to-cell interaction (CCI) of breast cancer. Our framework is interpretable in that it is able to find the correlated spots that support the clustering, which matches the CCI interaction pairs as well, providing more confidence to clinicians when making clinical decisions.

*Correspondence should be addressed to liyu@cse.cuhk.edu.hk
Code is available at https://github.com/ys-zong/conST

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have enabled the characterization of the transcriptome of individual cells, yielding cell sub-populations across organs by high-throughput profiling. However, the dissociation step erases the spatial context of cells from their original tissue, which is of vital importance for understanding cellular functions and organizations.

Spatially resolved transcriptomics (SRT) (1) addresses these limitations by measuring the gene expression matrix and its corresponding spatial information simultaneously. Current spatial transcriptomics technologies can be mainly divided into two categories: (1) High-plex RNA imaging (HPRI) methods use probes targeting specific genes to localize mRNA transcripts. This method includes fluorescence *in situ* hybridization (FISH) or *in situ* sequencing (ISS), such as MERFISH (2), seqFISH (3), seqFISH+ (4), and STARmap (5). (2) Spatial labeling methods utilize spatial barcodes to capture mRNA transcripts across tissue cross-sections, and then deep sequencing is performed after detachment. Typical methods include ST (6), slide-seq (7), Slide-seqV2 (8), and 10x Genomics. The data generated from spatial labeling platforms, such as ST, 10x Genomics, are often accompanied by histology images (morphology) besides gene expression and spatial information. These two platforms are usually complementary: HPRI methods can achieve single-cell resolution with greater depth, while spatial labeling methods have greater coverage and are more accessible (9). In both platforms, mRNA transcripts are captured by the individual coordinates or spatial barcoding locations. For simplicity, we refer them to *spots* in the following texts.

The rich information from different modalities of the SRT data sheds light on the biological insights. A multitude of works utilizes Markov random field to integrate gene expression and spatial information. (10) proposed a hidden Markov random field (HMRF) model for cell type and spatial domain identification according to spatial dependency. Giotto (11) analyzes the correlation of gene expression among its neighbors for spatial domain detection. BayesSpace (12) utilizes the information of spatial neighborhood to iteratively update the model in a Bayesian manner. While these methods can achieve state-of-the-art performance on some specific tasks, such as cell type and spatial domain identification, they are unable to learn a universal embedding for downstream tasks, thus being less flexible.

Recently, there are some analytical tools that can generate low-dimensional embeddings. Seurat (13) adopts the single-cell analysis pipeline that is mainly focused on gene expression data while overlooking the spatial link. SpaCell (14) uses a ResNet (15) pretrained on ImageNet (16) to extract features from images of each spot and uses autoencoders to learn embedding from the extracted features and gene expression, where the spatial information is completely ignored. And as it is only tested on old version SRT data, it is unclear whether this method is effective for SRT of a higher resolution. stLearn (17) develops a Spatial Morphological gene Expression (SME) normalization method to recompute gene expression values by averaging neighboring spots. It also uses an ImageNet-pretrained ResNet to extract morphology features to calculate weights for averaging. The normalized expression values are used for downstream tasks, but containing excessive noise. SpaGCN (18) applies a graph convolutional network (GCN) to a graph constructed by spatial coordinates and morphological features, where gene expression of each spot is regarded as a node attribute. SpaGCN only incorporates the simple statistics of morphology during graph construction, and therefore the morphology may not be fully explored. SEDR (19) learns embeddings by reconstructing gene expression and spatial information with an autoencoder and a variational graph autoencoder, respectively. Morphological information is ignored in SEDR pipeline.

To sum up, there are three major limitations that remain unsolved in the previous methods. (1)

The morphological features are extracted by pretrained CNN or simply ignored. It has not been involved in the training process for further integration. (2) The biological relationship between spots, sub-clusters, and the global structure of the SRT data has not been fully explored during the embedding learning process, remaining a large room for performance improvement. (3) While the learned embeddings can be used in various downstream tasks, the "black box" model stems the interpretation of the obtained results.

Here, we present an interpretable multi-modal contrastive learning framework for spatial transcriptomics, conST, to address the above-mentioned problems. For the first concern, conST can effectively integrate gene expression, spatial information, and morphology (if accessible) to learn low-dimensional embeddings. A state-of-the-art computer vision model, MAE (20), is used to extract informative features from the morphology of each spot. The extracted features are then regarded as an individual node attribute to participate in the training. Our framework is also very flexible. When morphology is not available, it can take gene expression and spatial information as input, thus can be applied for both HPRI and spatial labeling data.

As for the second concern, we argue that there are three natural underlying relationships in SRT data that can be used as supervision signals to guide the network to learn more meaningful embeddings: (1) local: a small portion of noises does not prevent a spot to be identified by its distinguishing features, (2) global: as a dataset is taken from the same slice (*e.g.*, same piece of tissue), the spots within it possess similar globally general features, (3) context: the node features are more related inside a sub-cluster, *e.g.*, more similar expression pattern and imaging texture. Based on the above assumptions, we propose to utilize contrastive learning (21; 22) to learn embeddings by maximizing the mutual information in local-local, local-global, and local-context levels.

In terms of interpretation, we monitor the training process and utilize GNNExplainer (23) to identify the important subgraphs and node attributes contributing most to the predictions according to the mutual information. Therefore, we can not only obtain the results, but also know what contributes to the prediction, *i.e.*, finding the correlated spots that support the clustering, which matches the CCI interaction pairs as well. The interpretability will give clinicians more confidence when using conST for the real clinical data analysis. We demonstrate how to analyze the tumour microenvironment and cell-to-cell interaction of breast cancer tissue with conST, leading to a clear cancer progression prediction, which will be helpful for making clinical treatment decisions.

The experimental results demonstrate that the learned embeddings can be successfully applied to many downstream tasks, such as clustering, trajectory inference, spatially variable genes (SVGs) detection, batch correction, and cell-to-cell interaction, *etc.* Quantitatively, conST outperforms other concurrent methods, *e.g.*, by up to 10% increase regarding ARI on clustering tasks. The qualitative results in both HPRI and spatial labeling platforms demonstrate the superiority of our method.

conST is user-friendly and flexible. With the standard SRT input, users can obtain the embeddings in the commonly-used formats in an end-to-end manner. We also provide detailed tutorials for downstream analysis with the generated embeddings.

Our main contribution can be summarized as follows:

- We propose conST, a powerful and flexible multi-modal framework that can effectively incorporate gene expression, spatial information, and morphology to learn low-dimensional yet expressive embeddings for downstream tasks.

- To the best of our knowledge, we are the first to introduce contrastive learning in the area of

spatial transcriptomics analysis, demonstrating its effectiveness in exploring the pattern from the data itself.

- The use of GNNExplainer enhances the interpretability of conST, giving users more confidence especially when used in clinical situations by not only telling what, but also why.
- Extensive experiments are conducted using the learned embedding in various datasets. The experimental results demonstrate the superiority and robustness of the proposed method in both HPRI and spatial labeling platforms.

2 Methods

2.1 Problem Definition and Framework

The whole framework is presented in Figure 1. Gene expression is undoubtedly the deterministic factor of the biological property. Besides, the rich information contained in histology images has been proven to be successful in revealing other important characterizations (24; 25). The spatial information provided by SRT acts as a bridge linking them together so that gene expression and morphology can complement each other as well as be more aware of the neighborhood.

Therefore, naturally, their underlying relationship can be modeled by graph structure and processed by graph neural networks (GNNs) (26). Hence, we treat the embedding generation process as a multi-modal self-supervised graph representation learning problem. In our setting, a graph is constructed based on K-Nearest Neighbor (KNN) distance of the coordinates of each spot, where the spots are regarded as nodes whose attributes are gene expression and morphology features (if accessible).

Denote $\mathcal{G} = (\mathcal{V}, E)$ as a graph with N nodes, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ represents the node set and $E \subseteq \mathcal{V} \times \mathcal{V}$ represents the edge set. The feature matrix of all the nodes is denoted as $\mathbf{X} = \{x_1, x_2, ..., x_N\} \subseteq \mathbb{R}^{N \times F}$, where F is the dimension of node features. If nodes possess multiple attributes, the t-th feature matrix is denoted as $\mathbf{X}_t = \{x_1, x_2, ..., x_N\} \subseteq \mathbb{R}^{N \times F_t}$. The adjacency matrix is given by $\mathbf{A} = [0, 1]^{N \times N}$, of which $\mathbf{A}_{ij} = 1$ if $\{v_i, v_j\} \subseteq \mathcal{E}$ or 0 otherwise.

Our objective is to learn a general encoder $\mathcal{E}(\mathbf{X}, \mathbf{A})$ in a self-supervised manner that can produce node embeddings in the low-dimensional space. Denote the learned embedding $\mathbf{H} = \mathcal{E}(\mathbf{X}, \mathbf{A}) \subseteq \mathbb{R}^{N \times F^{'}}$, where $F^{'} \ll F$, and $\mathbf{h_i}$ is the embedding of node v_i. We utilize Graph Convolutional Networks (GCNs), a powerful variants of GNN. Let $\hat{\mathbf{A}}$ denote the normalized adjacency matrix and $\mathbf{H}^{(l-1)}$ denote the embedding of layer $l-1$. The propagation of the GCNs is defined as $\mathbf{H}^{(l)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^l)$, where $\mathbf{W}^l$ is a learnable weight matrix and σ is a non-linear activation function. The normalized adjacency matrix is defined as $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix of $\mathbf{A}$, *i.e.*, $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, and $\mathbf{I}$ is an identity matrix for self-loop.

The training can be divided into two stages. In the pretraining stage, gene expression and morphology features are input into an autoencoder-based network to further reduce dimensions and initialize the weights of the encoder $\mathcal{E}$. In the major training stage, we focus on contrastive learning to guide the network to learn more informative and robust embeddings.

2.2 Pretraining stage

Based on the euclidean distance, the adjacency matrix $\mathbf{A}$ is constructed, where $\mathbf{A}_{ij} = 1$ if $\{v_j\}$ is in the KNN of $\{v_i\}$ and 0 else. We filter the genes that have very low expression and use principle

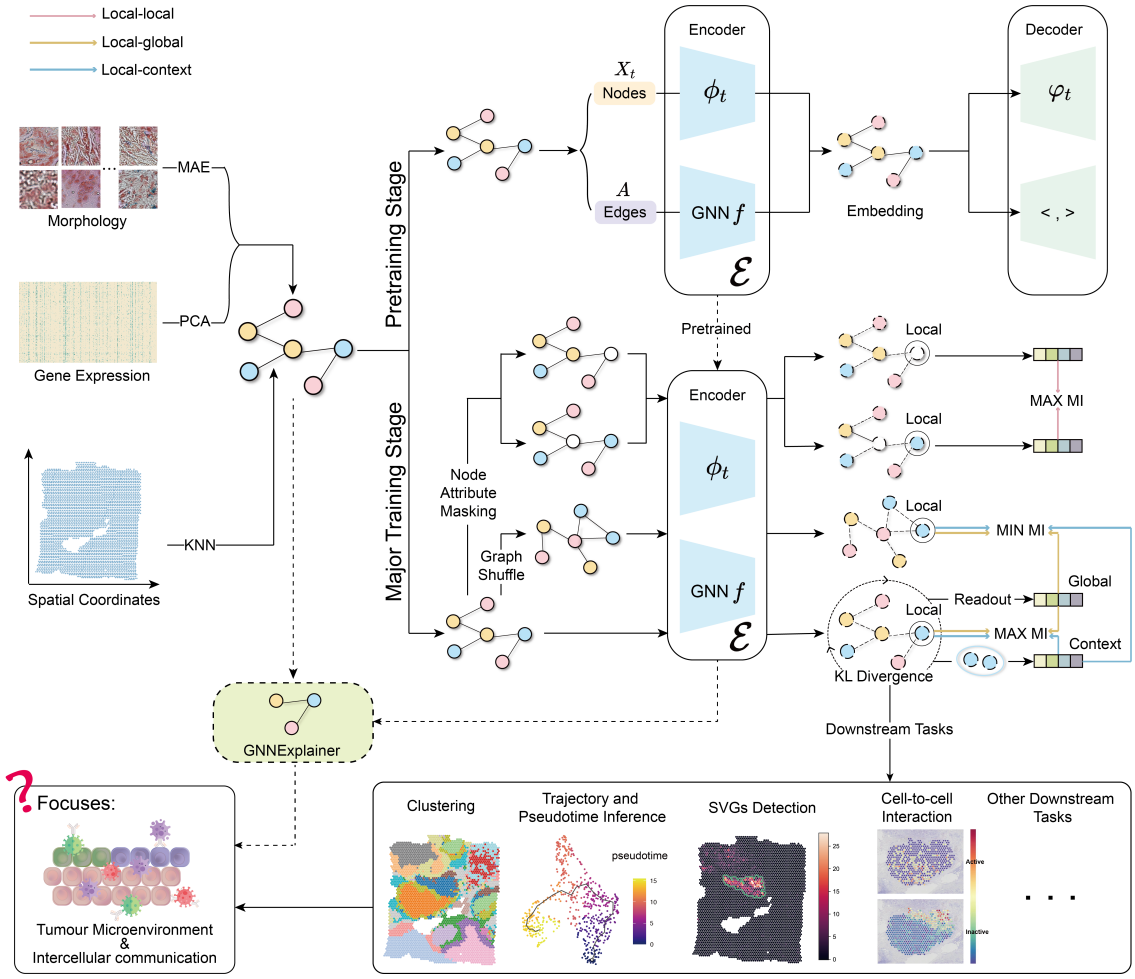

Figure 1: Framework of conST. conST models the ST data as a graph by treating gene expression and morphology as node attributes and constructing edges by spatial coordinates. The training is divided into two stages: pretraining and major training stage. Pretraining stage initializes the weights of the encoder $\mathcal{E}$ by reconstruction loss. In major training stage, data augmentation is applied and then contrastive learning in three levels, *i.e.*, local-local, local-global, local-context, are used to learn a low-dimensional embedding by minimize or maximize the mutual information (MI) between different embeddings. The learned embedding can be used for various downstream tasks, which, when analyzed together, can shed light on the widely concerned tumour microenvironment and cell-to-cell interaction. GNNExplainer helps to provide more convincing predictions with interpretability.

component analysis (PCA) to reduce the expression matrix to dimention F_1, *i.e.*, $\mathbf{X}_1 \subseteq \mathbb{R}^{N \times F_1}$.

The morphological features of each spots are extracted by a pretrained MAE model (20), which achieves SOTA performance and transferability on various datasets. Due to the small number of the available Hematoxylin and Eosin (H&E) stain images from ST, MAE is pretrained using ImageNet in a self-supervised manner. The recent development in computer vision communities has proven that features extracted by self-supervised models are more robust and transferable (27). MAE contains an asymmetric encoder-decoder architecture, with a Vision Transformer (ViT) (28) encoder. The decoder is used for pre-training, and only its encoder is used to extract the morphology features.

The pre-training process masked a large proportion (75%) of each image for reconstruction with positional encoding to enable the model to be aware of both the global and local information, while the large ViT makes the extractor more powerful. The detailed explanation of MAE can be found in the supplementary material. Compared with supervised pretrained ResNet used in previous methods (14; 17), we argue that the MAE is more powerful for extracting features of the H&E stain images. The extracted feature dimension of each spot is F_2, and the feature matrix is regarded as another node attribute $\mathbf{X}_2 \subseteq \mathbb{R}^{N \times F_2}$.

In the pretraining stage, we take advantage of proximity-based learning to initialize the parameters and further reduce the dimension of node attributes. For node attribute $\mathbf{X}_t \subseteq \mathbb{R}^{N \times F_t}$, a deep autoencoder Φ_t with an encoder ϕ_t and decoder of φ_t is used. The encoder produce a latent embedding $\mathbf{U_t} = \phi_t(\mathbf{X}_t) \subseteq \mathbb{R}^{N \times D_{ft}}$, where D_{ft} denotes the dimension of the learned low-dimensional embedding. After that, a variational graph autoencoder (VGAE) (29) with a encoder f is used to encode the spatial information into the node features. Depending on whether morphology is available, denote $\mathbf{U} = \|_t \mathbf{U}_t$, where $\|$ is the concatenation operation. The latent embedding from the VGAE is obtained by $\mathbf{V} = f(\mathbf{A}, \mathbf{U}) \subseteq \mathbb{R}^{N \times D_g}$. The final embedding $\mathbf{H}$ is the concatenation of $\mathbf{U}$ and $\mathbf{V} \subseteq \mathbb{R}^{N \times F^{'}}$, where $F^{'} = \sum_t D_{ft} + D_g$.

To be more specific, the encoder f of the VGAE is a two-layer GCN whose inference process is defined as

$$f(\mathbf{V}|\mathbf{A}, \mathbf{U}) = \prod f(\mathbf{V}_i|\mathbf{A}, \mathbf{U}), \tag{1}$$

with

$$f(\mathbf{V}_i|\mathbf{A}, \mathbf{U}) = \mathcal{N}(\mathbf{V}_i|\mu_i, diag(\sigma^2)), \tag{2}$$

where the $\boldsymbol{\mu} = \mathrm{GCN}_\mu(\mathbf{A}, \mathbf{U})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$, $\log \sigma = \mathrm{GCN}_\sigma(\mathbf{A}, \mathbf{U})$, and $\mathcal{N}$ is the normal distribution.

The decoding part of the autoencoder Φ_t and VGAE all take the concatenated embedding $\mathbf{H}$ as input, which possesses both spatial information and node attributes. For the autoencoder Φ_t, the decoder φ_t tries to reconstruct the input feature matrix as $\mathbf{X}_t^{'} = \varphi_t(\mathbf{H}) \subseteq \mathbb{R}^{N \times F_t}$. The decoder of VGAE is given by a simple inner product to generate the reconstructed $\mathbf{A}^{'} = \sigma(\mathbf{H} \cdot \mathbf{H}^T)$.

The autoencoder Φ_t is optimized by Mean Squared Error (MSE) between $\mathbf{X}_t$ and $\mathbf{X}_t^{'}$. The VGAE is optimized by minimizing the standard cross entropy between the input $\mathbf{A}$ and reconstructed $\mathbf{A}^{'}$, as well as Kullback-Leibler (KL) divergence between $f(\mathbf{H}|\mathbf{A}, \mathbf{U})$ and the Gaussian prior $p(\mathbf{H}) = \prod_i p(\mathbf{h}_i) = \prod_i \mathcal{N}(\mathbf{h}_i|0, \mathbf{I})$.

2.3 Major training stage

In the major training stage, the encoder part of the autoencoder and the VGAE are preserved, and the decoder part is dropped. We hope to learn a general encoder $\mathcal{E} = \{\phi_t, f\}$ that produce the embeddings, *i.e.*, $\mathbf{H} = \mathcal{E}(\mathbf{X}, \mathbf{A})$. Contrastive learning is then used to supervise the network for better representation learning at three levels: local-local, local-context, and local-global.

Local-local. In the process of sequencing, noises to the gene reads are commonly introduced by device deficiencies or manual operations. However, experts are still able to identify important characteristics of a specific spot based on the overview of its gene expression and the marker genes. Similarly, we hope the network to be able to distinguish a specific node from other nodes even if there are unwanted noises. Motivated by this and the recent work in node-level graph contrastive learning (30), the local-local level contrastive learning is utilized to force the network to focus on

more important features of the node attributes. Two views of the original graphs are created by masking node attributes. Then, the mutual information of the same node between these two views is maximized.

Specifically, we first randomly mask the node attributes of each spot at a given ratio to construct two augmented graphs $\widetilde{\mathcal{G}}_1$ and $\widetilde{\mathcal{G}}_2$, with an unchanged adjacency matrix $\mathbf{A}$ and masked feature matrices $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$. The masked feature matrix $\widetilde{\mathbf{X}}$ is computed by

$$\widetilde{\mathbf{X}} = [\mathbf{x}_1 \circ \widetilde{\mathbf{m}} \parallel \mathbf{x}_2 \circ \widetilde{\mathbf{m}} \parallel \cdots \parallel \mathbf{x}_N \circ \widetilde{\mathbf{m}}]^T, \tag{3}$$

where $\parallel$ is concatenation operation and $\circ$ is the element-wise product. $\widetilde{\mathbf{m}} \subseteq [0,1]^F$ is a random sampling vector. Given a node dropping probability p_m, each dimension of $\widetilde{\mathbf{m}}$ is independently drawn from a Bernoulli distribution, namely, $\widetilde{m}_i \sim \mathcal{B}(1-p_m)$.

Denote generated embeddings from the two augmented graph as $\mathbf{H_U}$ and $\mathbf{H_V}$. The contrastive objective is to distinguish the embedding of the same node in two augmented views from that of the other nodes. For node v_i, we set its embedding $\mathbf{H_U}_i$ generated in one view as an anchor. The embedding $\mathbf{H_V}_i$ generated in the other view is treated as a positive sample, and the rest of nodes are treated as negative samples. Define a similarity measurement as $\text{sim}(\mathbf{H_U}, \mathbf{H_V}) = \theta(p(\mathbf{H_U}), p(\mathbf{H_V}))$, where θ is the cosine similarity and $p(\cdot)$ is a projection head constructed by a two-layer multiple layer perceptron (MLP). Formally, for a positive pair ($\mathbf{H_U}_i$, $\mathbf{H_V}_i$) of node v_i, the objective is formulized similar to NT-Xent (31). For positive pair, we have $S_{\text{pos}} = \exp(\text{sim}(\mathbf{H_U}_i, \mathbf{H_V}_j)/\tau)$, where τ is the a temperature hyperparameter. For negative pairs within a same view, *i.e.*, intra-view, we have $S_{\text{intra-neg}} = \sum_{l=1}^{N} \mathbb{1}_{[l\neq i]} \exp(\text{sim}(\mathbf{H_U}_i, \mathbf{H_U}_l))$, where $\mathbb{1}_{[l\neq i]} \subseteq \{0,1\}$ is the indicator function. And for negative pairs between two views, *i.e.*, inter-view, we have $S_{\text{inter-neg}} = \sum_{l=1}^{N} \mathbb{1}_{[l\neq i]} \exp(\text{sim}(\mathbf{H_U}_i, \mathbf{H_V}_l))$. Thus, the objective can be given by

$$L_{\text{pair}}(\mathbf{H_U}_i, \mathbf{H_V}_i) = \log \frac{S_{\text{pos}}}{S_{\text{pos}} + S_{\text{intra-neg}} + S_{\text{inter-neg}}}. \tag{4}$$

Since the two views are symmetric, the whole objective for local-local contrastive learning is calculated by averaging all positive pairs as

$$\mathcal{L}_{\text{l2l}} = \frac{1}{2N} \sum_{i=1}^{N} [L_{\text{pair}}(\mathbf{H_U}_i, \mathbf{H_V}_i), L_{\text{pair}}(\mathbf{H_V}_i, \mathbf{H_U}_i)]. \tag{5}$$

Local-global. For a specific SRT dataset (*e.g.*, a slice), the gene expression or morphology of spots are intrinsically coherent with the global property, as they are captured from the same section of the same species. Therefore, it is beneficial to maximize the mutual information between the embeddings of each individual node and the whole graph summary to endow learned embeddings with the global structure and more robustness to the neighboring noises.

We adopt a similar method with Deep Graph Infomax (DGI) (32), a commonly used graph self-supervised method that has shown superior performance. First, a *readout function* $\mathcal{R}$ is used to obtain the overall summary $\mathbf{s}$ of the graph, where $\mathcal{R}: \mathbb{R}^{N\times F'} \rightarrow \mathbb{R}^{F'}$ and $\mathbf{s} = \mathcal{R}(\mathcal{E}(\mathbf{X}, \mathbf{A}))$. A corrupted graph $\widetilde{\mathcal{G}} = \mathcal{C}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}})$ is obtained by a corruption function $\mathcal{C}$, which randomly shuffles the rows of the feature matrix and randomly adds/drops edges. The embeddings obtained from original graph and corrupted graph are denoted as $\mathbf{H}$ and $\widetilde{\mathbf{H}}$.

A discriminator $\mathcal{D}_G : \mathbb{R}^{F'} \times \mathbb{R}^{F'}$ is employed to assign higher probability scores to positive pairs $(\mathbf{s}, \mathbf{h}_i)$ than the negative pairs $(\mathbf{s}, \widetilde{\mathbf{h}}_i)$. The objective is to maximize the Jensen-Shannon-based BCE as

$$\mathcal{L}_{\text{l2g}} = \sum_{i=1}^{N} \mathbb{E}_{(\mathbf{X},\mathbf{A})}[\log \mathcal{D}_G(\mathbf{s}, \mathbf{h}_i)] + \sum_{i=1}^{N} \mathbb{E}_{(\widetilde{\mathbf{X}},\widetilde{\mathbf{A}})}[\log(1 - \mathcal{D}_G(\mathbf{s}, \widetilde{\mathbf{h}}_i)]. \tag{6}$$

Local-context. Besides the global property, spots tend to be more similar at the cluster level. For example, spots may share similar marker genes and morphological texture within the same tissue type. Hence, we also try to maximize the mutual information between the node attributes and the cluster-level summary, which is beneficial for learning the embedding with hierarchical structure.

Inspired by Graph InfoClust (33), first, we use a differentiable K-Means clustering algorithm (34) to obtain M clusters. The centroid of each clusters $\boldsymbol{\mu}_m \subseteq \mathbb{R}^{1\times F'}$ is updated iteratively through a cluster layer

$$\boldsymbol{\mu}_m = \frac{\sum_i r_{im}\mathbf{h}_i}{\sum_i r_{im}}, \text{with } r_{im} = \frac{\exp(-\gamma \text{sim}(\boldsymbol{\mu}_m), \mathbf{h}_i))}{\sum_m \exp(-\gamma \text{sim}(\boldsymbol{\mu}_m), \mathbf{h}_i))}, \tag{7}$$

where $m = 1, 2, \ldots, M$, and γ is a inverse-temperature hyperparameter. For each node, we maximize the mutual information between $\mathbf{h}_i$ and its corresponding cluster summary $\mathbf{z}_i$. $\mathbf{z}_i$ is computed as

$$\mathbf{z}_i = \sigma(\sum_{m=1}^{M} r_{im}\boldsymbol{\mu}_m), \tag{8}$$

where r_{im} is the probability of node v_i assigned to cluster m and $\sum_{m=1}^{M} r_{im} = 1$.

Similar to local-global level, another discrimintor $\mathcal{D}_C$ is used to assign higher scores to the positive pairs $(\mathbf{h}_i, \mathbf{z}_i)$ than the negative pairs $(\widetilde{\mathbf{h}}_i, \mathbf{z}_i)$. Note that we use the same corruption function $\mathcal{C}$ as that in the local-global level. The objective is also to maximize the Jensen-Shannon-based BCE as:

$$\mathcal{L}_{\text{l2c}} = \sum_{i=1}^{N} \mathbb{E}_{(\mathbf{X},\mathbf{A})}[\log \mathcal{D}_C(\mathbf{z}_i, \mathbf{h}_i)] + \sum_{i=1}^{N} \mathbb{E}_{(\widetilde{\mathbf{X}},\widetilde{\mathbf{A}})}[\log(1 - \mathcal{D}_C(\mathbf{z}_i, \widetilde{\mathbf{h}}_i)]. \tag{9}$$

Weighted by λ_i, the overall contrastive learning objective function is given by

$$\mathcal{L}_{\text{cont}} = \lambda_1\mathcal{L}_{\text{l2l}} + \lambda_2\mathcal{L}_{\text{l2g}} + \lambda_3\mathcal{L}_{\text{l2c}}. \tag{10}$$

KL Divergence. In addition to contrastive learning, we use a deep clustering method (35) to further refine the embedding for more compactness. Simply, we use the normal K-means to cluster the embedding to K clusters. First, a Student's t-distribution kernel is used to calculate the soft assignment probability q_{ij} of the embedding $\mathbf{h}_i$ to the cluster centroid $\boldsymbol{\nu}_j$

$$q_{ij} = \frac{(1 + ||\mathbf{h}_i - \boldsymbol{\nu}_j||^2)^{-1}}{\sum_{j'}(1 + ||\mathbf{h}_i - \boldsymbol{\nu}_{j'}||^2)^{-1})}. \tag{11}$$

Next, based on q_{ij}, a target distribution P is calculated to help learn from the assignments with higher scores

$$p_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_j^{'}(q_{ij'}^2/)\sum_i q_{ij'}}. \tag{12}$$

Finally, an auxiliary KL Divergence objective is defined as

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{13}$$

Finally, the loss function in the major training stage is

$$\mathcal{L}_{\mathrm{major}} = \alpha \mathcal{L}_{\mathrm{cont}} + \beta \mathcal{L}_{\mathrm{KL}}, \tag{14}$$

where α and β are weight factors.

2.4 GNNExplainer

GNNExplainer (23) is a method for providing explanations for any GNN-based models in a model-agnostic manner. Here, we use the trained conST model and specific node index we would like to inspect as inputs. The GNNExplainer can identify an important subgraph and node features that are most influential to the predictions as an explanation by maximizing the mutual information between the subgraph and the predictions.

Formally, for a given node v_i, GNNExplainer aims to identify a subgraph $\mathcal{G}_S \subseteq \mathcal{G}$ and the corresponding feature matrix $\mathbf{X}_S = \{\mathbf{x}_j | v_j \subseteq \mathcal{G}_S\}$. The GNNExplainer can be optimized according to the mutual information (MI) as:

$$\underset{\mathcal{G}_S}{\mathrm{MAX}} MI(Y, (\mathcal{G}_S, \mathbf{X}_S)) = H(Y) - H(Y|\mathcal{G} = \mathcal{G}_S, \mathbf{X} = \mathbf{X}_S), \tag{15}$$

where Y is the prediction by conST, and $H(), H(|)$ are the marginal entropy and conditional entropy, respectively. The mutual information measures whether removing an edge link or a node feature is determinant to the final prediction results. Thus, by maximizing the MI, the GNNExplainer returns a subgraph and node features that have the highest mutual information with the original input, *i.e.*, the returned subgraph and node features can lead to the most similar prediction as the original input.

For conST, we are not only interested in what predictions it makes, but also why it makes such predictions. The GNNExplainer can identify which spots have the most important correlation with the given spot and their correlation can further explain gene expression and cell-to-cell interaction activities.

3 Experiments

3.1 Datasets and Implementation Details

We use various datasets from both HPRI and spatial labeling platforms to verify the effectiveness of our method. For HPRI data, we use mouse hypothalamus MERFISH (36), mouse visual cortex seqFISH (10). For spatial labeling data, we use human dorsolateral prefrontal cortex spatialLIBD (37) generated by 10x Visium, Human Breast Cancer (Block A Section 1) generated by 10x genomics (38), and mouse olfactory bulb Stereo-seq (39).

We set $K = 20$ for KNN graph construction. PCA is used to reduce the dimension of gene expression to $F_1 = 300$. The dimension of morphological features is $F_2 = 768$. Users can also easily adjust these parameters accordingly to meet the requirements of specific datasets. The two stages are trained 200 epochs respectively.

For downstream tasks, Leiden algorithm (40) is used for clustering; Monocle3 (41) and PAGA (42) is utilized for trajectory inference and pseudo-time inference; we adopt a similar SVGs detection method as SpaGCN (18) with our own embedding; Cell-to-cell interaction is performed with TraSig (43) and Seurat (13) for label transfer. The detailed descriptions of datasets and implementation can be found in the supplementary material.

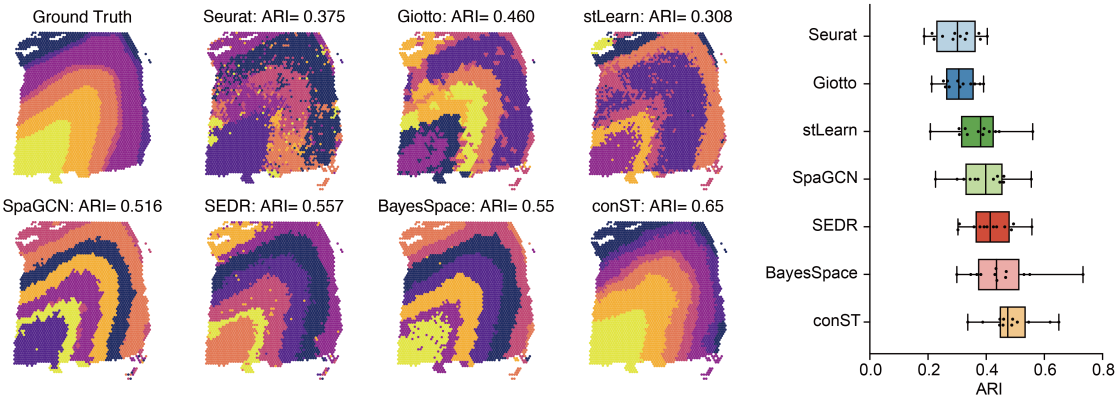

Figure 2: Comparison of different methods on slice 151673 of spatialLIBD dataset evaluated by ARI. conST achieves an ARI of 0.65, around 10% increase compared to other state-of-the-art methods. conST also achieves the highest mean and median ARI over all the 12 slices.

3.2 Clustering

Accurate clustering of spatial domains or cell types reveals the structure of the SRT data. We experimented on SRT data produced in different platforms with different resolutions, coverages, and depths. The results prove that conST is a powerful and generalizable framework that can be used to cluster both HPRI and spatial labeling data.

As shown in Figure 2, we compare the clustering results of conST with other methods in spatialLIBD dataset and illustrate slice 151673 which has clear layer boundaries. Visually, the results of conST do not have distinct outliers compared to other methods. Also, although the results of SpaGCN, SEDR, and BayesSpace all have seven layers that are relatively clear, the white matter (WM) layer and Layer 1 together (the first two layers in left bottom of the ground truth) they clustered are of the similar size of the WM layer in the ground truth. That will lead to the mismatch of the following layers, while each of the cluster boundaries of conST aligns well with the ground truth. conST achieves an ARI of 65%, about 10% higher than the state-of-the-art method. Also, conST achieves the highest mean and median ARI among all the methods.

For HPRI data, we experimented with MERFISH and seqFISH data. Different from spatial labeling data, they are captured in a higher resolution but with fewer genes in each spot and no morphology is available. Even so, conST still demonstrates strong performance. In Figure **??**, we compare the clustering performance of mouse hypothalamus MERFISH data with SpaGCN and SEDR. Visually, conST produces more consistent clusters. Quantitatively, as there is no ground truth for MERFISH data, we use three unsupervised metrics to evaluate the quality of the clusterings, *i.e.*, Silhouette Coefficient (SC), Calinski Harabasz Score (CHS), and Davies Bouldin Index (DBI) (Supplementary material Eq S1, S2, S6). For SC and CHS, the higher the better, while for

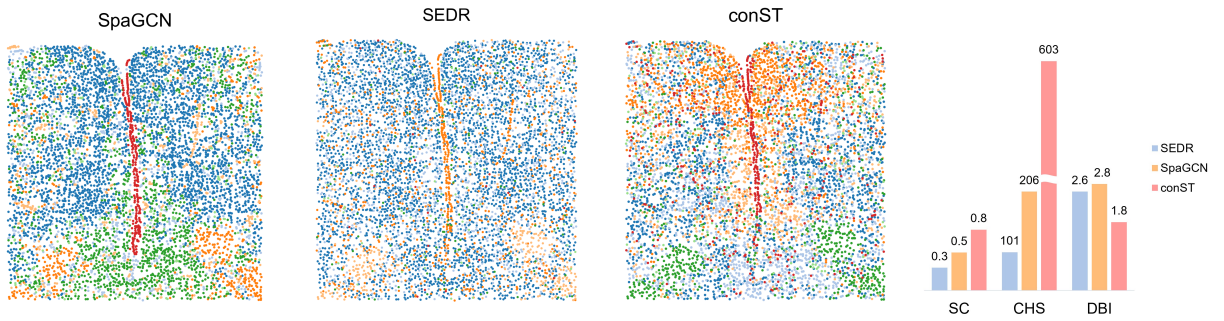

Figure 3: Comparison of the clustering results on MERFISH data. conST produces visually more consistent clusters and outperforms SpaGCN and SEDR in Silhouette Coefficient (SC), Calinski Harabasz Score (CHS), and DaviesBouldin Index (DBI). SC and CHS are the higher the better, while DBI is the lower the better.

Case	Slice 151673	Mean	Median
w/o local-local	0.559	0.410	0.421
w/o local-global	0.615	0.457	0.438
w/o local-context	0.586	0.423	0.419
All	**0.650**	**0.487**	**0.474**

Table 1: Ablation study of different contrastive learning components measured by ARI.

DBI, the lower the better. conST significantly outperforms SpaGCN and SEDR on all the three metrics.

We also evaluated conST on Stereo-seq and seqFISH data (Supplementary Figure S2, S3), where conST produces clear boundaries and biologically meaningful clusters. The statistical results also prove the superiority of our method.

Ablation Study. Clustering is arguably the most intuitive way to justify the quality of the learned embeddings. We perform an ablation study to verify the necessity and effectiveness of each of the contrastive learning components. Results on the spatialLIBD dataset demonstrate that all of the three levels of contrastive learning are beneficial to embedding learning. It can be seen from Table 1 that the local-local level has the most important influence on the performance, of which the data augmentation randomly masks some features of the node attributes. It produces a similar effect as dropout (44) that widely occurs in SRT data as well. Then, by using contrastive learning, conST can effectively learn to distinguish the more important invariant features among the noising data, thus obtaining a stronger ability. It achieves the best performance when all of the three contrastive losses are used together, as shown in Table 1.

3.3 Trajectory and Pseudotime Inference

Trajectory inference infers the pattern of cells in the dynamically developmental progress, and pseudotime represents the progression through this progress. Monocle3 (41) and PAGA (42) are used to produce trajectory and pseudotime inference where we replace the default PCs with the generated embeddings to validate the quality of the learned embeddings. For pseudotime in the spatialLIBD dataset, white matter (WM) is selected as the starting point.

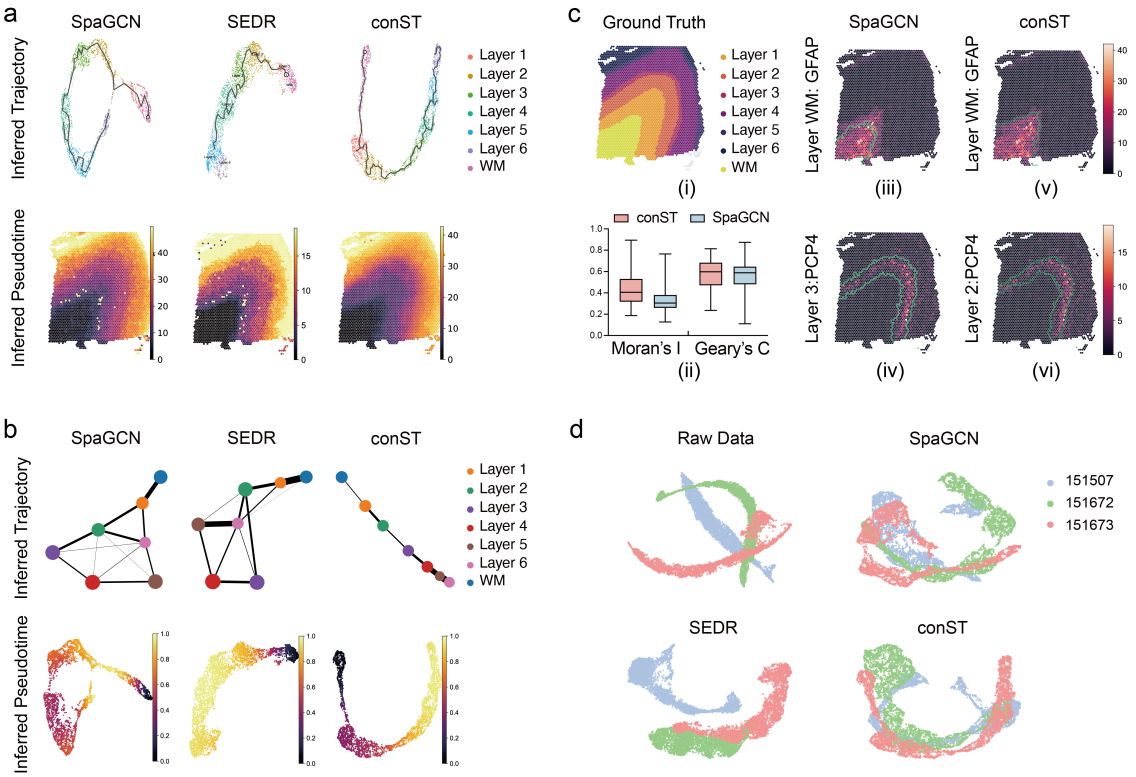

Figure 4: **a**. Comparison of trajectory and pseudotime inference using Monocle3 based on the learned embeddings of SpaGCN, SEDR, and conST. The trajectory inferred from conST goes through all the layers consistently and the pseudotime is much smoother with fewer outliers. **b**. Comparison of PAGA trajectory and pseudotime inference based on the learned embeddings of SpaGCN, SEDR, and conST. The trajectory inferred from conST is nearly linear, showing a clear progression process. The pseudotime of conST is also more consistent compared to SpaGCN and SEDR. **c**. SVGs detection comparison of SpaGCN and conST. conST outperforms SpaGCN in both Moran's I and Geary's C. The expression pattern of SVGs detected by conST aligns better with the actual anatomical layers, where the green lines indicate the predicted clustering boundaries. **d**. conST can alleviate batch effects. Raw data have apparent batch effects. Compared with other methods, conST can correct batch effects while preserving the semantic meanings.

As shown in Figure 4a, the trajectory generated from our embeddings are more consistent along with the clusters, covering almost all the spots, while the embeddings of SpaGCN are rewinded like a circle and the trajectory of SEDR does not reach Layer 6. Also, our pseudotime is smoother with much fewer outliers than the other two methods.

In Figure 4b, the trajectory of conST inferred from PAGA is nearly linear, indicating a clear progression process from the white matter to the layer 6. It is also worth noticing that there are connections between layer 4 and layer 6 in the trajectory infered from conST, where the layer 4 to layer 6 are thin and close to each other in the ground truth. It is possible that they are not well diverged, indicating the biological consistency of our prediction. The trajectories from SpaGCN and SEDR are disordered with crossing lines between different layers. We also visualized the pseudotime on embeddings projected by UMAP. conST shows more consistent progression than SpaGCN and SEDR.

Furthermore, the accurate trajectory and pseudotime inference by conST can be also beneficial

for more in-depth analysis such as cell-to-cell interaction, as detailed in Section 3.6.

3.4 SVGs Detection

Spatially variable genes (SVGs) have different expression patterns in different spatial locations. Detection of SVGs is helpful for identifying the tissue structure and clinical phenotypes. The better clustering performance enables conST to detect SVGs more accurately. For a fair comparison, we use the same detection method and default parameters as that in SpaGCN.

For slice 151673 in spatialLIBD, we have better performance on two commonly used metrics Moran's I and Geary's C (Figure 4c (ii)), which demonstrates SVGs detected by conST have a higher spatial autocorrelation. Furthermore, we select gene GFAP and PCP4 for comparison, as they are all highly expressed and detected by both SpaGCN and conST. While GFAP is detected in Layer WM by both methods, it can be seen that the boundary of Layer WM of conST aligns well with the highly expressed area of GFAP than SpaGCN (Figure 4c (iii), 4c (iv)), compared with the ground truth (Figure 4c (i)). For PCP4 (Figure 4c (v), 4c (vi)), although SpaGCN and conST all have cluster boundaries aligned well with the highly expressed area, it is however detected at Layer 3 by SpaGCN, which should be in Layer 2 as shown in ground truth and conST. The wrong layer could even lead to worse understanding if we want to know the relationship between the layers and SVGs. Hence, conST can be a more accurate tool for SVGs detection, and it confirms the superiority of our clustering results in turn.

3.5 Batch Correction

The batch effect refers to the changes to the result data in different experiments caused by non-biological factors, which can lead to imprecise conclusions. Batch effects are commonly observed in high-throughput sequencing experiments including SRT data. conST is able to alleviate the batch effects through contrastive learning. The local-global level contrastive learning enables spots to be aware of the global property. In this way, the learned embeddings of different slices of the same species will share similar properties. Furthermore, the local-local level contrastive learning makes the network more robust to the technical noises introduced by devices or manual operations (45; 46). Thus, the learned embedding can be projected into a shared latent space, correcting batch effects.

We select three slices showing substantial batch effect (19) in spatialLIBD dataset for illustration. We project the learned embeddings to 2-D space by UMAP (47). As shown in Figure 4d, the raw data demonstrate distinct batch effects. conST effectively corrects the batch effects with more compact and mixed embeddings, compared with SpaGCN and SEDR.

3.6 Cell-to-cell Interaction

For a tumour tissue, normally it can be grouped into 4 main morphotypes: Ductal Carcinoma in Situ (DCIS), healthy tissue (Healthy), Invasive Ductal Carcinoma (IDC), and tumor surrounding regions with low features of malignancy (Tumor edge) (48). DCIS consists of the proliferation of malignant cells which do not invade the basement membrane of the breast ducts. It is a nonobligate precursor to IDC and a large amount of DCIS lesions remain indolent in practice. However, currently, almost all the DCIS lesions are treated in practice to prevent further invasion, and the treatment comprises either mastectomy or breast-conserving radiotherapy surgery that will be harmful to the patients (49).

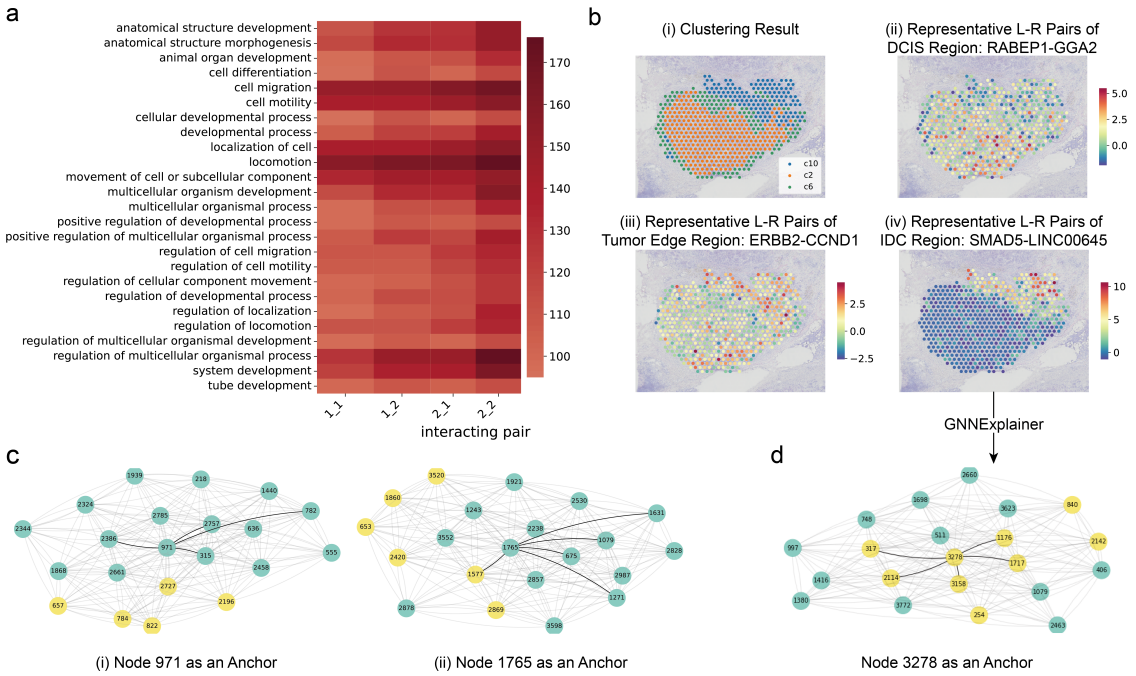

Figure 5: Cell-to-cell interaction and GNNExplainer. **a.** GO term analysis for cluster c2, c6 and c10 using Trasig based on embedding from conST. **b.** Representative L-R pairs of cell-to-cell interaction analysis. (i) Three clusters obtained by conST. (ii) (iii) (iv) representative L-R pairs of DCIS, Tumor edge, and IDC region respectively. **c.** GNNExplainer can identify subgraphs that contribute most to the predictions. Node 971 and 1765 in cluster boundary are selected as anchor nodes for illustration, whose identified subgraphs are plotted with bold lines. The color of nodes represents the clustering label. Most of the nodes in the subgraph belong to the same cluster as the anchor node, indicating conST can aggregate information from spots with similar attributes. **d.** GNNExplainer can further support the identified active L-R pairs in cell-to-cell interaction. All the spots are selected from the same cluster c10 of BRCA in **b**. We set a threshold of interaction value to define spots as active in yellow, and inactive in green. The nodes in the identified subgraph of node 3278 all have active L-R pairs interaction as node 3278.

To counter the overtreatment, we predict the potential cancer progression by conST, hoping to provide more insights for clinical decision-making. First, the clinically potential target receptors on breast cancer cells are detected. Then, we analyze the neighbor-spreading trend for the IDC region and evaluate the risk for developing invasive breast cancer with a tool of cell-to-cell interaction (CCI).

We obtained the embeddings of the dataset of Human Breast Cancer (Block A Section I) (38) with conST, in short as BRCA. With the learned embeddings, we cluster them into 20 regions (Supplementary Figure S4). Given the heterogeneity of the tissue, we specifically focus on the obvious lesion area, *i.e.*, cluster c2, c6, and c10 (Figure 5b (i)). Trajectory inference is applied to these three clusters to gain pseudotime ordering (Supplementary Figure S10) as we did in section 3.2 and 3.3. Then, by detecting SVGs, we find some highly expressed genes, including IGFBP2, PVALB, RABEP1, RAB11FIP1, LINC00645, *etc.* (Supplementary Figure S8, S9).

We demonstrate how to analyze the tumour microenvironment for evaluating neighbor-spreading risk with the above-mentioned downstream tasks, which is divided into cross-clusters CCI analysis and within-cluster CCI analysis.

3.6.1 Cross-clusters: Morphotype Prediction

By inputting the cluster prediction, trajectory and pseudotime inferred from our embeddings to TraSig (43), an analyzing tool for cell-to-cell interaction, we identify the interacting cell-type pairs and active ligand-receptor (L-R) pairs for the tumour tissue.

Cluster c2 and c10 demonstrate more active breast-cancer-related genes interactions, such as BRCA1, BRCA2, BRCC3, and TP53 (50). Cluster 6, as an inactive region, represents the tumour edge. Also, according to GO enrichment analysis (Figure 5a) obtained from TraSig, c10 performs the highest potential on cell migration, cell motility, and regulation of the multicellular organismal process. Strong locomotion ability is one of the most representative phenomena for the IDC region to be distinguished from DCIS regions (48). Therefore, we deduce that clusters 2, 6, and 10 are IDC region, tumor edge, and DCIS region, respectively.

3.6.2 Within-cluster: Neighbor-spreading Risk Evaluation

With the property of breast cancer invasive progression, the within-cluster analysis focuses on the interaction-specific function on cell migration, locomotion, and cell proliferation as shown in the GO term analysis (Figure 5a). We used label transfer from Seurat (13) to annotate spots of the given tissue and obtain the probability distribution of spots related to inferred clusters. Then, we scored the spatial L-R co-expression at every spot and its neighbors.

IDC Region (c2): For the IDC region, cell migration and invasion related to miR-205-3p have been recently reported (51). Upregulated LINC00645 significantly influences the progression of cells *in vivo* as miR-205-3p was a target of LINC00645 and LINC00645, modulating TGF-β-induced cells locomotion via miR-205-3p (52). Here we detect the interaction between SMAD5, which belongs to the TGF-β superfamily of modulators, and LINC00645. As Figure 5b (iv) shows, the interaction is much more active in IDC regions.

Also, it is reported that at least 50% of breast tumours have an activated type 1 insulin-like growth factor-1 receptor (IGF-1R). The active interaction of IGF-1R with its two natural ligands, insulin-like growth factor-1 (IGF-1) and IGF-2, has been associated by many investigations, as one primary risk factor in breast cancer (53). The receptor system is complex since IR and IGF-1R genes can form several types of hybrid receptors. Here we detect the activities of some highly expressed IGFs, such as IGFBP2-IGF1, IGFBP2-IGF1R, and IGFBP2-TUBGCP5 (Supplementary Figure S11). It also performs not only an upregulated expression within IDC spots but also a more active interaction than DCIS and tumor edge regions.

With a higher chance, LINC00645 and IGF families are potentially well-applied targets in clinical trials on breast cancer, especially for IDC region.

Tumor Edge Region (c6): Genes are considered as "amplified" when the ratio of their copy number in tumour and normal samples is greater than two. Estrogen receptor gene amplification is very frequent in breast cancer. It is reported that more than 20% of breast cancers harbor genomic amplification of the ESR1 gene1 (54). We detect the interaction of some top amplified genes such as ERBB2 (Figure 5b (iii)), ESR1, and CCND1 (Supplementary Figure S12). The result shows that these genes, such as ERBB2-CCND1, have a very active interaction on the border of tumor edge (c2) and IDC region (c6). The overexpression of these amplified gene enhances metastasis-related properties (invasion, angiogenesis, increased survival) of cancer cells that might lead to increased cancer metastases (55). It is very possible that cancer cells from the IDC region are invading the tumor edge, and the tumor edge area will decrease after several cell cycles.

DCIS Region (c10): We further evaluate the following stage of the DCIS region. RAB4A is an essential regulator as it is functional for the fast recycling of integrin $\beta 3$. Integrin $\beta 3$ regulates cell polarity and migration when localized appropriately to the plasma membrane, thereby having an essential role in cancer metastasis (56). Also, GGA2 functions in recycling endosomes to retrieve endocytosed EGFR, thereby sustaining its expression on cell surface, and consequently, cancer cell growth (48). We detect the interaction of LR-pairs RABEP1-GGA2 (Figure 5b (ii)), RABEP1-GGA1, RABEP1-RAB4A, RABEP1-RAB5A, *etc* (Supplementary Figure S13). These pairs show a more obvious interaction, with a strong possibility indicating that the DCIS region is deteriorating and heading to the next stage: Invasive Ductal Carcinoma.

Knowing when a lesion will or will not be life-threatening is essential and requires a thorough understanding of the tumour microenvironment and cancer progression. Here, with the learned embeddings from conST, we can perform more accurate evaluations of the cell-to-cell interactions about IGF-related genes, top amplified genes, and cell metastasis. We conclude the high risk on neighbor-spreading for the IDC region in this special case and ensure that the outcome knowledge will contribute to the decision-making for clinicians in general.

3.7 Interpretability

3.7.1 Clustering Explanation

While conST demonstrates strong performance on clustering tasks, we are also interested in why it clusters a specific spot into a specific layer, *i.e.*, which neighboring spots contribute most to the prediction. Thus, given a spot, we utilize GNNExplainer to identify a subgraph that contributes most to the prediction results.

In Figure 5c, we select nodes 971 and 1765 near the cluster boundary of slice 151673 in spatialLIBD for illustration, where the subgraphs are plotted in bold lines and node color represents the predicted clustering labels. In Figure 5c (i), all the nodes in the subgraph are in the same cluster with node 971, suggesting node 971 is predicted based on the node attributes of those nodes. In Figure 5c (ii), although most of the nodes in the subgraph are in the same cluster with node 1765, node 1577 is in the other cluster. However, the ground truth label shows node 1577 and node 1765 are in the same layer, which means the subgraph we generate is reasonable. Thus, when the subgraph does not match the predictions, it provides further information and helps to identify the reason, which would be helpful in practice. Also, it demonstrates that conST is still able to learn how to aggregate the information and recognize the underlying important relationship among spots in difficult decision boundaries. The use of GNNExplainer points out a way to better refine the network.

3.7.2 CCI Explanation

GNNExplainer can also provide interpretability to CCI on a more fine-grained scale. We select spots all from the same cluster c10 of BRCA (Figure 5b (i)) and inspect the interaction in IDC of pair SMAD5-LINC00645 (Figure 5b (iv)). We set a threshold of interaction value to define spots as active in yellow, and inactive in green (Figure 5d). Setting node 3278 as input, the nodes involved in the subgraph all have active L-R interactions. It demonstrates that conST can not only identify clusters, but also the predictions made by conST are supported by the realistic biological process.

4 Conclusion

We propose conST, a powerful and flexible framework for SRT data analysis with contrastive learning. conST can learn low-dimensional embeddings by effectively exploring the multiple modalities of SRT data, including gene expression, spatial information, and morphology (if accessible). The learned embeddings can be utilized in various downstream tasks and the performance surpasses other concurrent methods. Furthermore, we utilize conST to study the tumour microenvironment and cell-to-cell interaction of a breast cancer dataset in detail. conST reveals the cancer development stages and ligand-receptor pairs that reflect the cancer progression. It proves that conST can be used in more complex scenarios, providing more insights for future treatment.

The GNNExplainer explains which neighboring spots contribute to the prediction that conST makes, which is also biologically consistent with the interaction of the L-R pair identified in CCI. The interpretability will enable conST to be used in complex clinical situations with more convincing predictions.

5 Acknowledgements

This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics.

References

[1] Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays* **42**, 1900221 (2020).

[2] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed rna profiling in single cells. *Science* **348**, aaa6090 (2015).

[3] Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ rna profiling by sequential hybridization. *Nature methods* **11**, 360–361 (2014).

[4] Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature* **568**, 235–239 (2019).

[5] Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018).

[6] Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

[7] Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

[8] Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology* **39**, 313–319 (2021).

[9] Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics* 1–18 (2021).

[10] Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.-C. Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nature biotechnology* **36**, 1183–1190 (2018).

[11] Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology* **22**, 1–31 (2021).

[12] Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology* 1–10 (2021).

[13] Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021).

[14] Tan, X., Su, A., Tran, M. & Nguyen, Q. Spacell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* **36**, 2293–2294 (2019).

[15] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

[16] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (IEEE, 2009).

[17] Pham, D. *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).

[18] Hu, J. *et al.* Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods* 1–10 (2021).

[19] Fu, H., Hang, X. & Chen, J. Unsupervised spatial embedded deep representation of spatial transcriptomics. *bioRxiv* (2021).

[20] He, K. *et al.* Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[21] Wu, L., Lin, H., Tan, C., Gao, Z. & Li, S. Z. Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[22] Han, W. *et al.* Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *bioRxiv* (2021).

[23] Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* **32**, 9240 (2019).

[24] Song, Z. *et al.* Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature communications* **11**, 1–9 (2020).

[25] Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24**, 1559–1567 (2018).

[26] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE transactions on neural networks* **20**, 61–80 (2008).

[27] Hendrycks, D., Mazeika, M., Kadavath, S. & Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems* **32**, 15663–15674 (2019).

[28] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[29] Kipf, T. N. & Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[30] Zhu, Y. *et al.* Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131* (2020).

[31] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).

[32] Velickovic, P. *et al.* Deep graph infomax. *ICLR (Poster)* **2**, 4 (2019).

[33] Mavromatis, C. & Karypis, G. Graph infoclust: Leveraging cluster-level node information for unsupervised graph representation learning. *arXiv preprint arXiv:2009.06946* (2020).

[34] LaLonde, R., Zhang, D. & Shah, M. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4003–4012 (2018).

[35] Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487 (PMLR, 2016).

[36] Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018). URL `https://www.science.org/doi/abs/10.1126/science.aau5324`.

[37] Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24**, 425–436 (2021).

[38] 10x Genomics. Human breast cancer (block a section 1), spatial gene expression dataset, 10x genomics (2020).

[39] Chen, A. *et al.* Large field of view-spatially resolved transcriptomics at nanoscale resolution. *bioRxiv* (2021).

[40] Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).

[41] Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

[42] Wolf, F. A. *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 1–9 (2019).

[43] Weiler, P., Van den Berge, K., Street, K. & Tiberi, S. A guide to trajectory inference and rna velocity. *bioRxiv* (2021).

[44] Qiu, P. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications* **11**, 1–9 (2020).

[45] Zhang, P., Jiang, Z., Wang, Y. & Li, Y. Clmb: deep contrastive learning for robust metagenomic binning. *bioRxiv* (2021).

[46] Wang, X. *et al.* Contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration. *arXiv preprint arXiv:2112.03266* (2021).

[47] Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology* **37**, 38–44 (2019).

[48] Nagasawa, S. *et al.* Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. *Communications biology* **4**, 1–13 (2021).

[49] van Seijen, M. *et al.* Ductal carcinoma in situ: to treat or not to treat, that is the question. *British journal of cancer* **121**, 285–292 (2019).

[50] Chai, K. M. *et al.* Downregulation of brca1-brca2-containing complex subunit 3 sensitizes glioma cells to temozolomide. *Oncotarget* **5**, 10901 (2014).

[51] Gregory, P. A. *et al.* The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting zeb1 and sip1. *Nature cell biology* **10**, 593–601 (2008).

[52] Li, C. *et al.* Long non-coding rna linc00645 promotes tgf-β-induced epithelial–mesenchymal transition by regulating mir-205-3p-zeb1 axis in glioma. *Cell death & disease* **10**, 1–17 (2019).

[53] Ekyalongo, R. C. & Yee, D. Revisiting the igf-1r as a breast cancer target. *NPJ precision oncology* **1**, 1–7 (2017).

[54] Reis-Filho, J. S. *et al.* Esr1 gene amplification in breast cancer: a common phenomenon? *Nature genetics* **40**, 809–810 (2008).

[55] Yu, D. & Hung, M.-C. Overexpression of erbb2 in cancer and erbb2-targeting strategies. *Oncogene* **19**, 6115–6121 (2000).

[56] Do, M., Chai, T., Casey, P. & Wang, M. Isoprenylcysteine carboxylmethyltransferase function is essential for rab4a-mediated integrin β3 recycling, cell migration and cancer metastasis. *Oncogene* **36**, 5757–5767 (2017).