BMC Genomics

SOFTWARE

Open Access

Spatially multimodal and multiscale network for representation learning from spatial multi-omics

Fengyu Zhang[1] and Cheng Peng[1*]

Abstract

Background Spatial multi-omics techniques provide powerful tools to decipher tissue architectures in multilayer perspectives, including gene expressions, gene regulations and microenvironments. The multimodal data integration is a critical step in spatial multi-omics data processing involved in many downstream analyses, such as spatial domain clustering. Thus, it is important to learn reliable latent representation by efficiently integrating the spatial multi-omics for downstream analyses.

Results In this work, we developed a Spatially Multimodal and Multiscale Network (SpaMM-Net) to learn the latent representations from spatial multi-omics data. Due to the intrinsic noises and the complex relationship among multiple omics, SpaMM-Net utilized spatially-guided multiscale graph attention networks to integrate the multimodal omics features at different spatial-scale levels for representation learning. We evaluated our method for downstream spatial clustering task on several spatial multi-omics datasets. The results show that SpaMM-Net achieves good performance in deciphering both detailed regions and tissue architectures. Our scale-wise weight analyses further reveal that the SpaMM-Net effectively leverages multiscale spatial information to enhance the robustness of the learned representations.

Conclusion SpaMM-Net is an efficient tool to capture and integrate latent representations from spatial multi-omics for spatial domain identification. This strategy can be extended to other multi-omics modalities with the rapid development of spatial multi-omics technologies.

Keywords Spatial multi-omics, Multimodal, Multiscale, Representation learning, Spatial domain

*Correspondence:
Cheng Peng
chengpeng@ynu.edu.cn
[1]Yunnan Key Laboratory of Cell Metabolism and Diseases, Center for Life Sciences, School of Life Sciences, Yunnan University, Kunming 650500, China

Introduction

With the rapid development of spatial transcriptomics [1–4] and spatial multi-omics techniques [5, 6], it has become feasible to simultaneously capture multiple kinds of omics on the same tissue section, such as spatial RNA-seq, spatial ATAC-seq, and spatial CUT&Tag [7, 8]. These advances have provided powerful tools to align spatial multi-omics with tissue morphology, facilitating more comprehensive understanding of biological processes and disease mechanisms.

Though spatial multi-omics data provide valuable multi-layer information, these data exhibit complex relationship among different kinds of omics, including the discrepancies across modalities observed in some tissue regions. In addition, the intrinsic noises in each modality further make the data analysis more complicated [9]. Thus, efficient and accurate representation learning from noisy and complex spatial multi-omics data is one critical step for various downstream analyses, including spatial domain clustering. Several methods were proposed to learn feature representations for spatial domain detection. For example, the Seurat package used the weighted nearest neighbor algorithm to build joint multimodal graphs to support community detection methods [10], such as Louvain and Leiden. Ashuach et al. proposed MultiVI [11], a variational autoencoder based model that utilizes variational inference to learn shared latent representations across different kinds of omics. Coleman et al. concatenated pairwise interactions of latent features across modalities to enhance multimodal representation [12]. Long et al. introduced a dual-branch graph convolutional autoencoder with shared weights to jointly learn multimodal embedding while incorporating spatial coordinates [13].

However, the existing approaches mainly rely on single-scale k-nearest neighbor (KNN) graph structures, limiting their ability to capture the multi-level semantic features inherent in the spatial multi-omics data due to high noise level and specific discrepancies across modalities in many tissue regions. Inspired by the success of multi-scale feature extraction in computer vision where receptive fields of different sizes contribute to improved structural robustness, we proposed SpaMM-Net, Spatial Multimodal and Multiscale Network, to integrate spatial multi-omics data by using multi-scale graph neural architectures to generate stable and informative latent representations.

Specifically, SpaMM-Net is built upon an attention-based graph convolutional network. It first employs a spatially guided cross-attention mechanism to fuse features from multiple modalities. Then, a set of K-hop graph convolutional modules are used to capture shared representations at multiple receptive field scales. The gating module is used to align semantic information among scales, and the representations derived from different scales are aggregated into a unified latent space by using a multi-scale fusion module. We evaluated SpaMM-Net on multiple spatial multi-omics datasets, and the results demonstrate that our method effectively enhances multimodal representation robustness and provides more reliable embedding for downstream spatial domain detection.

Methods

Data processing

Three spatial multi-omics datasets were used to evaluate our method, covering the RNA-seq and ATAC-seq co-profiling, and RNA-seq and CUT&Tag co-profiling [7, 8]. To align the multimodal inputs, we first used ChIPseeker [14] to annotate the ATAC and CUT&Tag sequencing data into gene-level count matrices. Then, both the RNA count data and the annotated chromatin data were preprocessed using the standard pipeline with Scanpy [15], including total-count normalization for each cell, log-transformation of count values, and selection of the top 3,000 highly variable genes. These data were further reduced to 50 dimensions by using principal component analysis (PCA), which were used as input in SpaMM-Net to reduce the spatial complexity of the model and mitigate the effects of high-dimensional sparse features on representation learning.

Graph generation

In SpaMM-Net, we employed three KNN graphs in the shared-weight graph convolutional network: a spatial KNN graph based on spatial coordinates A^{sp}, the modality one (e.g. gene expression) based KNN graph A^{m1}, and the modality two (ATAC, CUT&Tag, or others) based KNN graph A^{m2}. The graph A^{sp} was constructed using Euclidean distance from spatial coordinates. We did not use histological image as modality in this work since some datasets are absent of high-quality tissue images. For A^{m1} and A^{m2}, we used a fused strategy that combines Euclidean distance and cosine similarity by considering the architectural design of SpaMM-Net. Specifically, let S^{eu} denote the Euclidean similarity matrix consisting of all spot pairs, and then we calculated the pairwise Euclidean distance between two given nodes and converted them into similarity scores using a Gaussian kernel:

$$S_{i,j}^{eu} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \quad (1)$$

where $S_{i,j}^{eu}$ denotes the Euclidean similarity between spot i and j, x_i and x_j denote the aforementioned dimension-reduced features derived from PCA in spot i and j

respectively, and the kernel bandwidth σ was defined as the median of all pairwise Euclidean distances:

$$\sigma = median\left(\|x_i - x_j\|_2\right) \tag{2}$$

Meanwhile, we also calculated the cosine similarity between all dimension-reduced features to capture their angular relationships. To ensure numerical consistency and comparability across modalities, the cosine similarity matrix was normalized using min-max method to enable more effective fusion in the subsequent similarity integration step:

$$\tilde{S}_{i,j}^{cos} = \frac{S_{i,j}^{cos} - \min\left(S_{i,j}^{cos}\right)}{\max\left(S_{i,j}^{cos}\right) - \min\left(S_{i,j}^{cos}\right)} \tag{3}$$

where $S_{i,j}^{cos}$ denotes the cosine similarity between spot i and j. Let $\tilde{S}^{cos}$ denote the normalized cosine similarity matrix consisting of all spot pairs. To combine the Gaussian-based similarity S^{eu} and the normalized cosine similarity $\tilde{S}^{cos}$, we calculated a weighted fusion similarity matrix S^m and generated the adjacent matrix of KNN network A^m as follows:

$$S^m = \alpha * S^{eu} + (1 - \alpha) * \tilde{S}^{cos} \tag{4}$$

$$A_{i,j}^m = \begin{cases} 1, & j \in KNN\left(S_i^m\right) \\ 0, & otherwise \end{cases} \tag{5}$$

where α denotes tunable parameter that controls the fusion weights of the similarity components, and are typically set to 0.5 by default. The i and j denote the row and column index in matrices S^m and A^m.

Spatially-guided cross attention

SpaMM-Net utilized an attention-based graph convolutional network with shared weights across three KNN graphs. To effectively fuse the multimodal embeddings derived from the three KNN graphs, we proposed a Spatially-guided Cross-Attention (SC-Attn) mechanism to integrate the modality-specific features, which was defined as following formula:

$$\begin{cases} Attn^{m1} = softmax\left(\frac{FC(sp)\cdot FC(m1)^T}{\sqrt{d_k}}\right) FC\left([sp, m1, m2]\right) \\ Attn^{m2} = softmax\left(\frac{FC(sp)\cdot FC(m2)^T}{\sqrt{d_k}}\right) FC\left([sp,\ m1,\ m2]\right) \end{cases} \tag{6}$$

where $m1$, $m2$, and sp denote the outputs of the shared-weight graph attention networks in the modality one, modality two and spatial KNN graphs, respectively. The scaling factor d_k is used to prevent excessively large attention logits and improve numerical stability, which is fixed by the hidden dimension of the attention module. The operator $[\cdot, \cdot]$ denotes concatenation along the feature dimension, and FC represents the fully connected network.

Given the varying quality of different modalities, we incorporated a confidence-aware mechanism into the attention module. Specifically, each modality was assigned a confidence score that was adaptively learned during training to reflect its feature reliability. These confidence scores were guided by spatial information and normalized via Softmax to form weighted coefficients, which were then embedded into the attention calculation:

$$\begin{cases} c^{m1} = Sigmoid\left(FC\left([sp,\ m1]\right)\right) \\ c^{m2} = Sigmoid\left(FC\left([sp,\ m2]\right)\right) \end{cases} \tag{7}$$

$$\left(Conf^{m1} | Conf^{m2}\right) = Softmax_s\left(c^{m1} | c^{m2}\right) \tag{8}$$

where the $Softmax_s$ denotes the Softmax operation along the sample dimension, and $[\cdot\,|\,\cdot]$ represents concatenation along the sample (or node) dimension. The above design allows the model to automatically reduce noisy or less informative modalities during feature fusion, thereby enhancing the robustness of the integrated representation. Finally, the SC-Attn module calculated the fused output feature SA by multiplying the attention features of each modality with their confidence scores, followed by a weighted summation:

$$SA = Conf^{m1} * Attn^{m1} + Conf^{m2} * Attn^{m2} \tag{9}$$

Cross-scale gating unit

To effectively integrate multi-scale features, we designed a Cross-Scale Gating Unit (SGU) that leverages high-resolution embeddings to guide the learning of lower-resolution features:

$$S_s = SA_s + \beta * Sigmoid\left(FC\left([SA_s, FC\left(SA_{s-1}\right)]\right)\right) \tag{10}$$

where β is a learnable parameter, and the subscript **s** denotes the corresponding feature scale. This fine-to-coarse interaction enhances semantic consistency and enables coordinated representation learning across different receptive fields.

Multi-scale feature fusion

After obtaining the multiscale feature S, we used a Multi-Scale Feature fusion (MSF) module to integrate them into a unified embedding. The MSF module employed a Softmax-based weighted aggregation mechanism that adaptively assigned importance to each scale. This fusion strategy preserved the different levels of informative content and enhanced the stability and robustness of the final representation. The fusion process was formally defined as follows:

$$W = Softmax_s\left([Sigmoid\left(FC\left(S_1\right)\right)|\cdots|Sigmoid\left(FC\left(S_s\right)\right)]\right) \quad (11)$$

$$\widetilde{S}_i = LayerNorm\left(FC\left(S_i\right)\right) \quad (12)$$

$$Z = W * \widetilde{S} \quad (13)$$

Since the feature derived from each network in this Multi-Scale Fusion module could contribute to the final spatial clusters, the scale-wise weights in W were used to represent the contributions of each scale.

SpaMM-Net training

We defined the reconstruction loss as the weighted Mean Squared Error (MSE) between the original and reconstructed features of all omic modalities:

$$L_{recon} = w_1 * MSE\left(X_1,\ X_1'\right) + w_2 * MSE\left(X_2, X_2'\right) \quad (14)$$

where X_1 and X_2, are the input features from omics modality one and two, and X_1' and X_2' are the reconstructed features corresponding to each modality. In the above equation, w_1 and w_2 are hyperparameters that control the contributions of each modality-specific reconstruction loss. To effectively manage multiple hyperparameters and balance heterogeneous loss terms, we adopted an uncertainty-based loss weighting strategy, which dynamically adjusts the weights based on the predictive uncertainty of each task [16].

For the correspondence loss, we adopted a training strategy inspired by the SpatialGlue method, in which the reconstructed features X_1' and X_2' were concatenated along the feature dimension and fed back into the encoder. By comparing the multiscale embeddings and the integrated representation, we defined two mean squared error losses, L_{corr_s} and L_{corr_e},to enhance reconstruction consistency at both the scale-specific and global levels:

$$L_{corr\_s} = \sum_{i \in s} w_i * MSE\left(S_i, S_i'\right) \quad (15)$$

$$L_{corr\_e} = MSE\left(Z, Z'\right) \quad (16)$$

where $\mathrm{i} \in \mathrm{s}$ denotes the index over different scales, and w_i are hyperparameters controlling the weight of the loss term.

The inner-layer alignment loss was calculated as following steps. While multiscale k-hop graph convolutions enable feature extraction at different levels of granularity, larger receptive fields can lead to over-smoothing, which may cause the loss of fine-grained information and negatively affect downstream performance. To make the large-scale features remain semantically consistent with the more detailed representations captured at intermediate scales, SpaMM-Net introduced an inter-scale constraint by computing a mean squared error loss L_{inn} between the embeddings obtained at the medium and large scales:

$$L_{inn} = MSE\left(S_{n-1},\ S_n\right) \quad (17)$$

where n denotes the index of the largest scale. To ensure that the medium-scale features guide the learning of the large-scale features without being influenced in return, the embedding $S_{\mathrm{n}-1}$ was excluded from parameter update during this step.

The final optimization objective combined all aforementioned loss terms as follows:

$$L = \gamma_1 * L_{recon} + \gamma_2 * L_{corr_s} + \gamma_3 * L_{corr_e} + \delta * L_{inn}$$

where the γ_1, γ_2 and γ_3 are dynamically adjusted using an uncertainty-based loss weight strategy. In contrast, δ is empirically set to a small constant (default: 0.1), as L_{inn} is particularly sensitive to the behavior of the large-scale feature extraction module.

Method comparison

Four multimodal methods: pyWNN (https://github.com/dylkot/pyWNN), MultiVI [11], MISO [12], and SpatialGlue [13] were selected to perform comparison on the latent representation derived from SpaMM-Net. The recommended data preprocessing pipelines were used in MultiVI and MISO to improve their performance by following their instructions. The same input features were used in pyWNN, SpatialGlue and SpaMM-Net to fairly compare the learned latent representations. Then the spatial domain clustering was used as the downstream analysis for the learned latent representations, in which the Leiden algorithm [17] was used in clustering and the cluster number was same in each dataset for all methods. To compare the learned representations to the original omics modalities, we also used the single omic modality (m1 or m2) to independently perform spatial domain clustering for comparison by using Leiden algorithm. For the datasets with ground truth or annotated regions, the adjusted rand index (ARI) was used to quantify the similarity between spatial clusters and annotated regions, and 10 random seeds were used to estimate the stability and/or uncertainty of the models. The pyWNN was excluded from random-seed estimation due to lack of this setting in this method. Since spatial domains have the trend to exhibit spatial coherence for some neighbor spots, the Moran's I [18] was also used to evaluate the spatial autocorrelation of clustering assignments. For the tissue with no available and detailed annotations, only the Moran's I were utilized to quantify the spatial clustering performance. Additionally, to investigate the contribution of features derived from different attention scales in

SpaMM-Net, we visualized the learned fusion weights from the MSF module using radar plots to reveal the varying contributions of attention scales in different clusters, indicating the effective adaption to dataset-specific structures in SpaMM-Net.

We used scMultiSim (v1.6.0) [19] to generate the simulated spatial multi-omics datasets, e.g. paired RNA-seq and ATAC-seq, for method comparisons. Specifically, the gene regulatory network was initialized using the built-in GRN_params_100, and additional ligand-receptor interactions were introduced with predefined target-regulator pairs and effect sizes. Each dataset consisted of 200 genes and 300 cells, with 50 cell identity factors to model transcriptional heterogeneity. Cell populations were structured using a Phyla3 lineage tree, enabling continuous differentiation patterns. To incorporate spatial information, we enabled the cell-cell interaction module with a layered spatial layout (layout = "layers"), a grid size of 28, and a maximum of 4 neighbors per cell. Cell-type interactions were simulated randomly. Eleven independent datasets were generated by varying the parameter rand.seed from 0 to 10. Other parameters were set to the default values.

Results

SpaMM-Net overview

SpaMM-Net utilized an encoder-decoder architecture for multimodal feature fusion and representation learning (Fig. 1). The encoder was built upon the Graph Attention Network (GAT) [20] for feature embedding. Initially, GATs with shared weights were used to extract base representations from the three input modalities: omics modality one (m1), omics modality two (m2), and spatial coordinates (sp). These representations were then fused using a Spatial-guided Cross-Attention (SC-Attn) module, which integrates information across modalities under spatial guidance. To capture contextual features across different receptive fields, we constructed a series of stacked K-hop GAT layer for multiscale spatial feature extraction. At each scale, SC-Attn was used to fuse the K-hop spatial features with the original multimodal embeddings derived from m1 and m2, resulting in scale-specific multimodal representations. To ensure

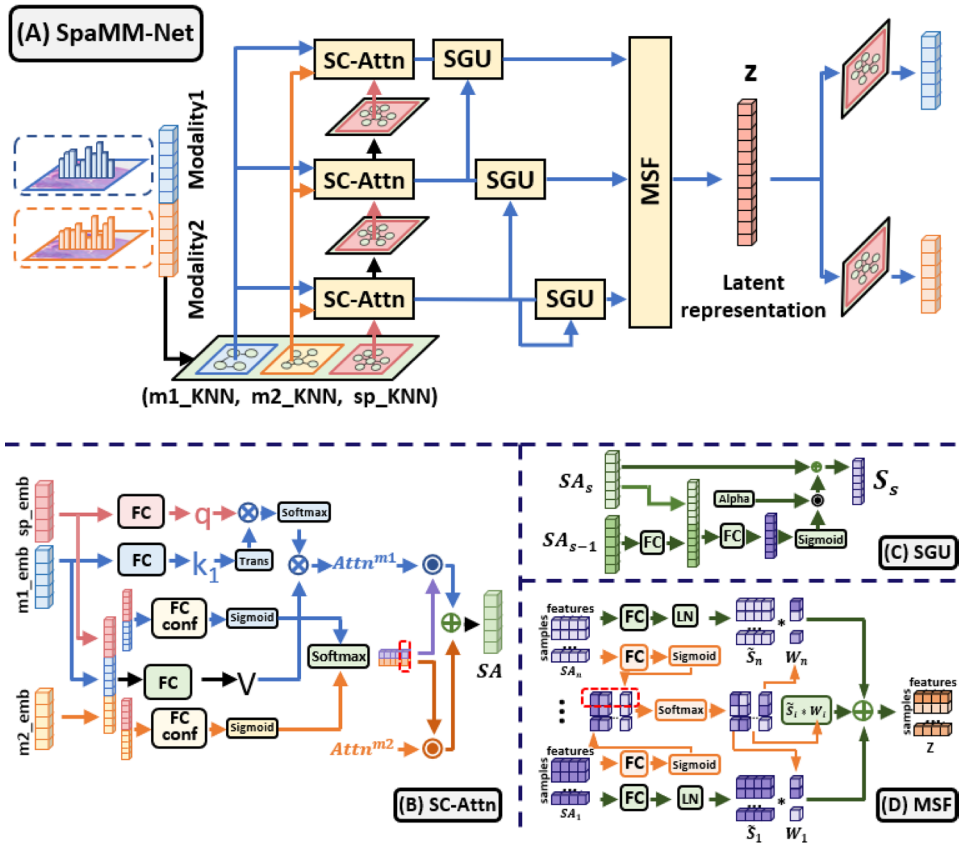

Fig. 1 The model architecture of SpaMM-Net. **A** The overall workflow of SpaMM-Net. m1_KNN and m2_KNN represent the graphs constructed from modality one and modality two respectively, and sp_KNN denotes the KNN graph constructed from spatial coordinates. The 50 dimensions were used for modality one and two in this work. **B** The workflow of SC-Attn module in SpaMM-Net. The $Attn^{m2}$ was calculated in the same way as $Attn^{m1}$ by using m2_emb to replace m1_emb. **C** The workflow of SGU module in SpaMM-Net. **D** The workflow of MSF module in SpaMM-Net. FC: Fully connected layer; LN: Layer normalization; SA: the fused output feature derived from spatially-guided cross attention

semantic alignment across scales, we introduced a Scale Gated Unit (SGU) that enhances cross-scale consistency. These representations were then aggregated by the Multi-Scale Feature fusion (MSF) module to obtain a robustly unified latent representation. The decoder reconstructed each modality using GATs built upon spatial KNN graph. Then the latent representations were used to downstream analysis such as spatial domain detection.

SpaMM-Net evaluation

We first used the scMultiSim [19] with its default parameters to generate the simulated spatial multi-omics data to evaluate the performance of our method since there are known spatial domains for the simulated multi-omics data (e.g. RNA-seq and ATAC-seq) in this kind of dataset (Fig. 2A). We then compared SpaMM-Net to other methods, including the spatial domains derived from only the RNA-seq or ATAC-seq (Fig. 2B). In overall, the spatial domains derived from all five selected methods resemble the ground truth in this simulated dataset. Further quantifications from ARI indicated that SpaMM-Net achieved highest similarity to ground truth, followed by SpatialGlue and pyWNN (Fig. 2C). The SpaMM-Net and SpatialGlue also exhibited highest Moran's I values (Fig. 2D), consistent with the smoothed spatial domains shown in Fig. 2B. These results indicated that SpaMM-Net and SpatialGlue learned the better representations for spatial domain clustering in this simulated dataset. We visualized the scale-wise weights of the features derived from different scales in the SpaMM-Net MSF module (Fig. 2E). In overall, scale 1, scale 2 and scale 3 contributed to the total weights in MSF module around 45.0%, 29.3% and 25.7% respectively. However, the scale-wise weights in different domains varied across different scale network,

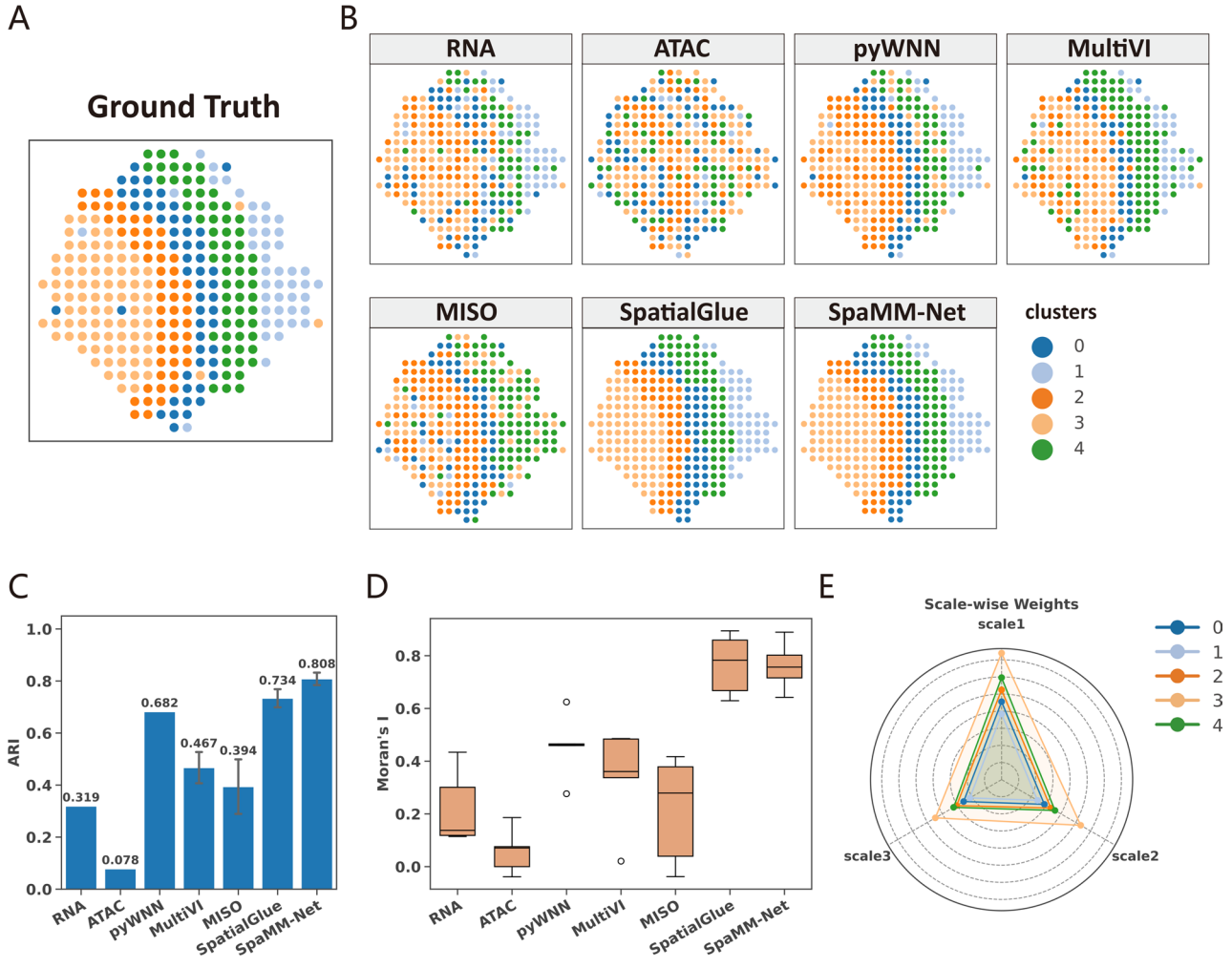

Fig. 2 Evaluation on the simulated dataset. **A** The simulated spatial multi-omics data and ground truth derived from the software scMultiSim. **B** The spatial clustering from RNA, ATAC and the representations generated by pyWNN, MultiVI, MISO, SpatialGlue and SpaMM-Net. **C** The ARI measuring the similarities between annotated domains and spatial clusters in all selected methods. Ten random seeds were used to estimate the uncertainty. **D** The statistics on Moran's I of all selected methods. In each method, the Moran's I was calculated on each cluster, and the statistical result was presented in bar plot. **E** The visualization of scale-wise weights. The weight of each cluster in each scale was presented with its relative value

indicating that SpaMM-Net captured the information from multiple spatial resolutions. We also performed the statistical comparisons by generating different simulated datasets, and the result showed that SpaMM-Net and SpatialGlue still achieved better performance than other methods (Supplementary Fig. 1).

We next conducted comparisons in the real datasets. We first used the spatial RNA and CUT&Tag co-profiling dataset derived from mouse brain coronal Sect [8]. since it was widely used in spatial multi-omics studies. To better evaluate the spatial clustering, we downloaded the reference anatomical map from the Allen Brain Atlas [21] and manually annotated the brain regions according to this reference map (Fig. 3A). After data preprocessing, the gene expression modality contained 9,750 spatial spots and 13,313 genes, while the CUT&Tag modality (H3K4me3 histone modification) included 9,750 spatial spots and 20,970 annotated signals. We compared SpaMM-Net to other methods, and the results showed that the pyWNN, MultiVI and MISO output relatively less organized clusters in this dataset, whereas SpatialGlue and SpaMM-Net output well organized clusters (Fig. 3B). We further quantified the clustering performance using the metrics ARI and Moran's I. The SpaMM-Net and SpatialGlue achieved highest ARI values (Fig. 3C). Both SpaMM-Net and SpatialGlue showed higher values for Moran's I (Fig. 3D). These two quantitative comparisons further confirmed that SpaMM-Net learned the better or comparable representations for the spatial domain clustering in this dataset. We also visualized the scale-wise weights of the features derived from different scales in the MSF module (Fig. 3E). In overall, scale 1, scale 2 and scale 3 contributed to the total

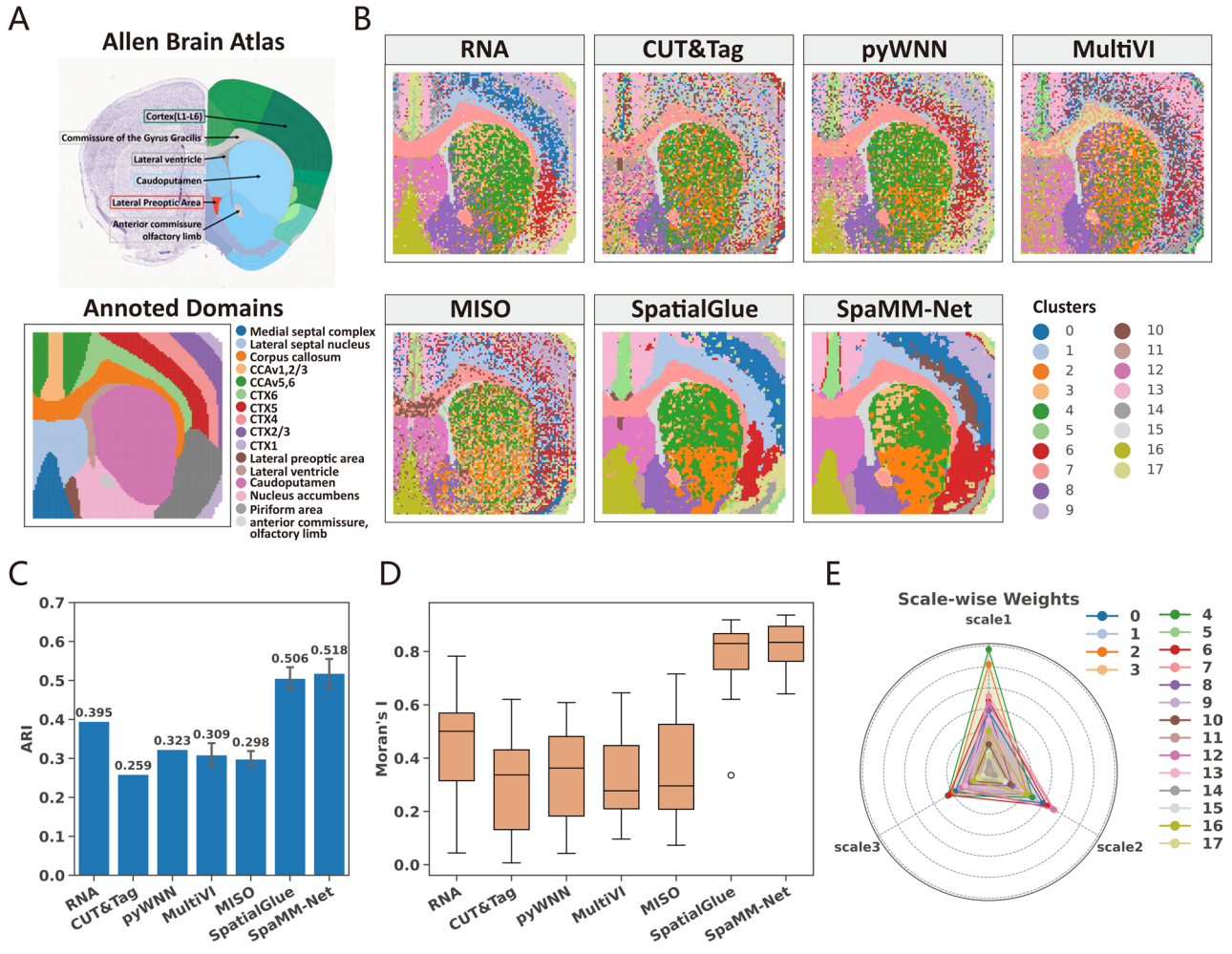

Fig. 3 Evaluation on coronal mouse brain section dataset. **A** The manually annotated domains according to the reference mouse brain map downloaded from Allen Brain Atlas. **B** The spatial clustering from RNA, CUT&Tag and the representations generated by pyWNN, MultiVI, MISO, SpatialGlue and SpaMM-Net. **C** The ARI measuring the similarities between annotated domains and spatial clusters in all selected methods. Ten random seeds were used to estimate the uncertainty. **D** The statistics on Moran's I of all selected methods. In each method, the Moran's I was calculated on each cluster, and the statistical result was presented in bar plot. **E** The visualization of scale-wise weights. The weight of each cluster in each scale was presented with its relative value

weights in MSF module around 44.5%, 34.2% and 21.3% respectively.

We then conducted comparison on the mouse brain sagittal section dataset generated through the MISAR-seq technique (sample: E18.5_S2) [7]. Compared to coronal sections, sagittal sections show different challenges in representation learning due to the complex spatial structures. We also manually annotated this tissue by considering the reference map from Allen Brain Atlas and the annotations from the original work [7] (Fig. 4A). After data preprocessing, the RNA modality consisted of 2,248 spatial spots and 11,179 genes, while the ATAC modality included 2,248 spots and 20,938 annotated genomic regions. As shown in Fig. 4B, the spatial domains derived from SpaMM-Net representation also captured the major regions annotated from histological images and reference map (Fig. 4A and B). We further quantified the clustering performance using the ARI and Moran's I (Fig. 4C and D). SpaMM-Net achieved higher ARI and Moran's I than the best competing method SpatialGlue, indicating substantially improved spatial coherence. Additionally, we analyzed the contributions of different scales in the MSF module (Fig. 4E). The averaged fusion weights for scale 1, scale 2, and scale 3 are 38.3%, 35.1%, and 26.6%, respectively with varying weights in each clusters. These results suggested that SpaMM-Net effectively leveraged multi-scale information to enhance the diversity and flexibility of the learned representation.

To further evaluate the applicability of SpaMM-Net, we next conducted comparison on the spatial RNA and ATAC co-profiling dataset generated from E13 mouse embryo [8], the tissue type distinguishing from mouse brain (Fig. 5). Due to the complexity of this tissue, we only annotated part of regions according to histological image and previous work [8] (Fig. 5A). As shown in Fig. 5B, most methods exhibited fuzzy boundaries and

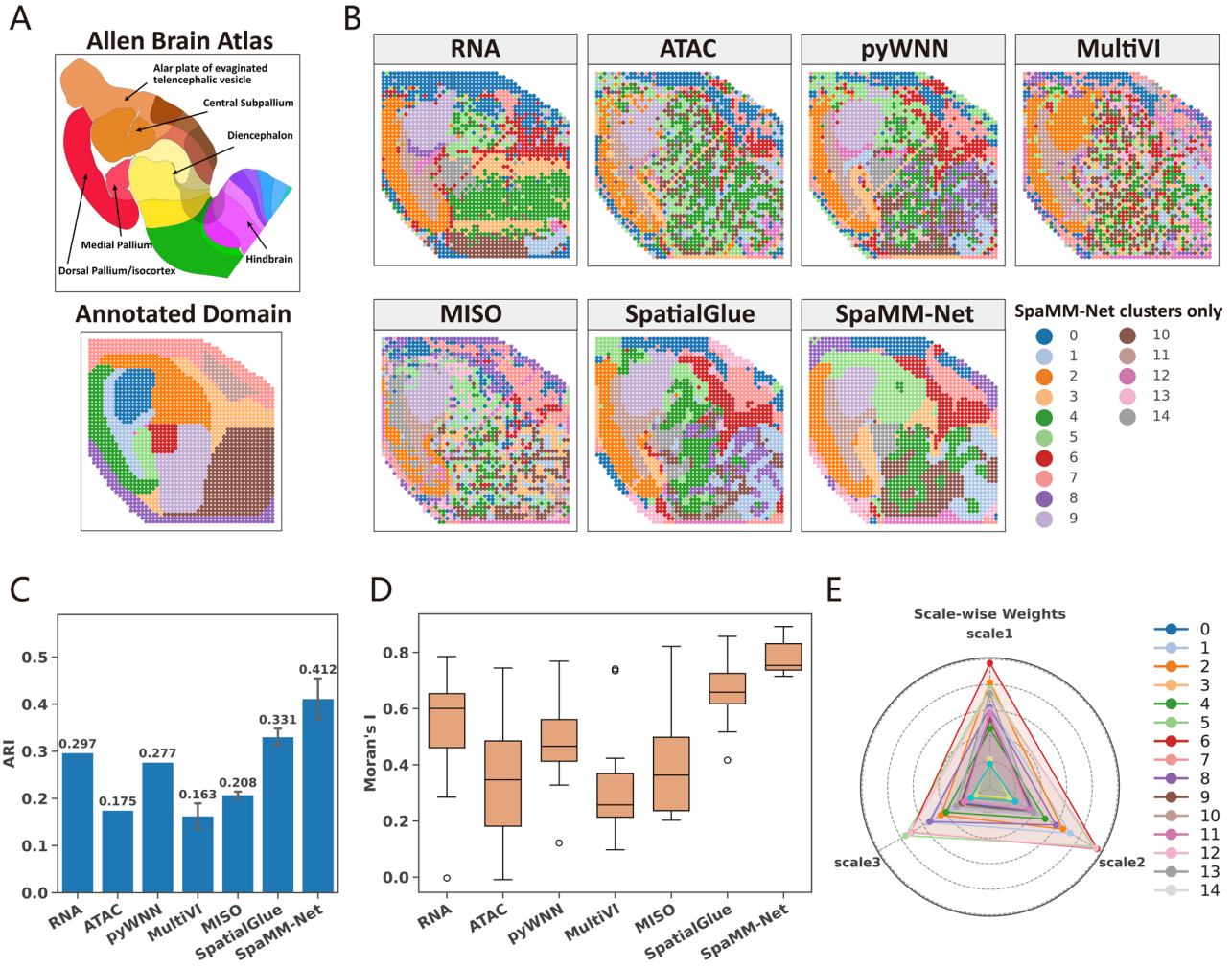

Fig. 4 Evaluation on sagittal mouse brain section dataset. **A** The manually annotated domains according to the reference mouse brain map from Allen Brain Atlas and original work. **B** The spatial clustering from RNA, ATAC and the representations generated by pyWNN, MultiVI, MISO, SpatialGlue and SpaMM-Net. **C** The ARI measuring the similarities between annotated domains and spatial clusters in all selected methods. Ten random seeds were used to estimate the uncertainty. **D** The statistics on Moran's I of all selected methods. **E** The visualization of scale-wise weights

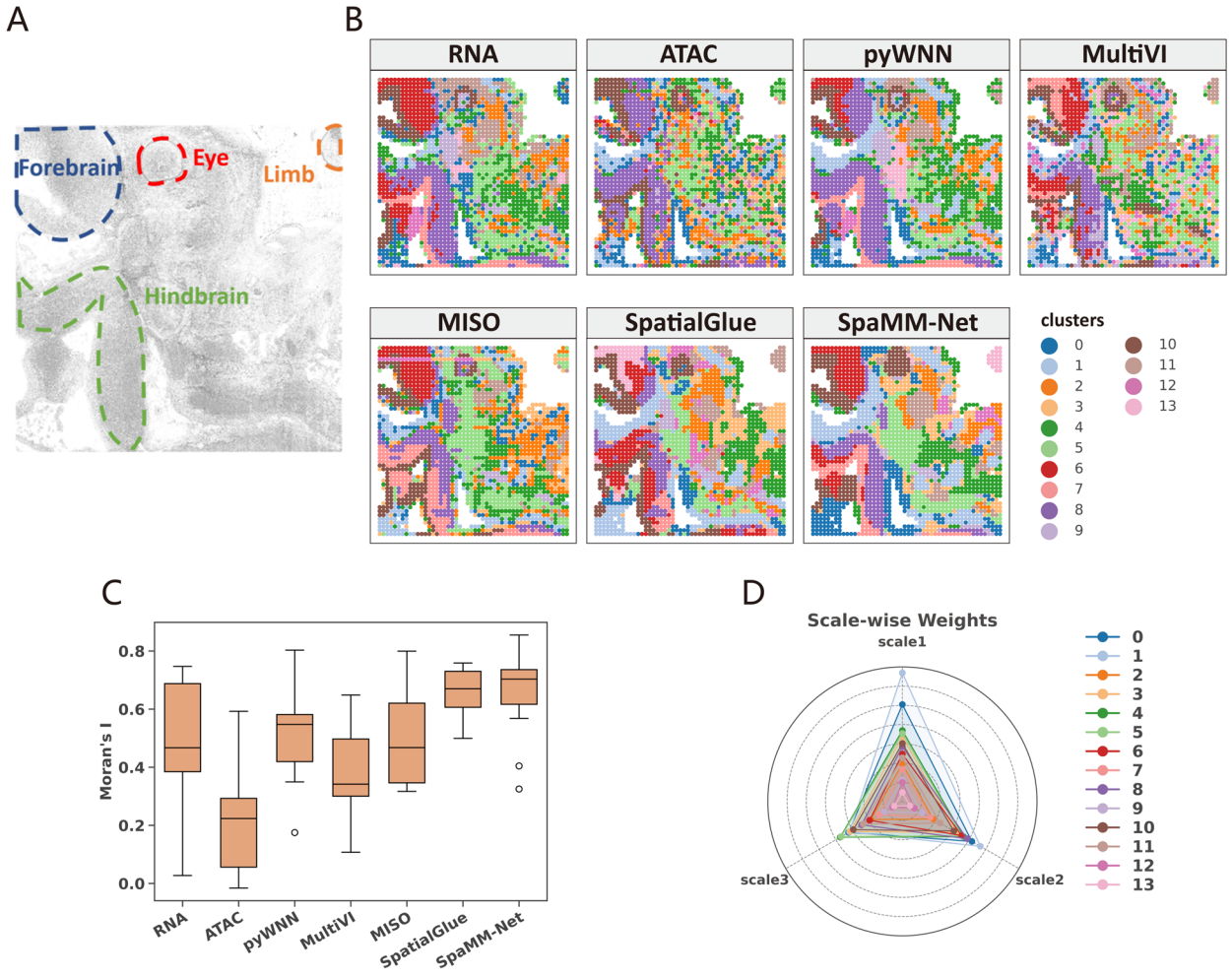

Fig. 5 Evaluation on mouse embryo dataset. **A** The histological image derived from the spatial multi-omics experiment. Some regions were highlighted according to the annotations derived from the original work. **B** The spatial clustering from RNA, ATAC and the representations generated by pyWNN, MultiVI, MISO, SpatialGlue and SpaMM-Net. **C** The statistics on Moran's I of all selected methods. **D** The visualization of scale-wise weights

fragmented clusters except for SpatialGlue and SpaMM-Net. In general, SpatialGlue and SpaMM-Net achieved relatively similar clusters and Moran's I (Fig. 5C). In details, SpaMM-Net showed better clustering results in some regions. For example, SpatialGlue showed discontinuities in the hindbrain region, whereas SpaMM-Net output relatively intact hindbrain. However, it should be noted that there is lack of objective comparison in this dataset. We then examined the contributions of different scales to the final representation during the MSF stage (Fig. 5D). The results show that the scale-wise weights also varies in different clusters, similar to previous datasets.

To evaluate the sensitivities of parameters used in SpaMM-Net, we examined the effects of the weighting coefficient α in Eq. (4) when calculating the fused similarity matrix, the neighbor number k in KNN for constructing modality-specific networks, and the number of principal components in dimension reduction. The results showed that the major spatial domains were preserved and the Moran's I values remained stable when weighting coefficient α varied (Supplementary Fig. 2 and Supplementary Fig. 3). As for the k nearest neighbor, the major spatial domains were stable with different k values (Supplementary Fig. 4). The local structures were slightly more coherent when $k_1 = 8$ and $k_2 = 15$. Therefore, we used the default parameters in the selected datasets. The spatial domains also remained stable when the number of principal components slightly varied (Supplementary Fig. 5 and Supplementary Fig. 6). Then we performed the ablation study on two important modules in SpaMM-Net, spatially-guided cross attention and multi-scale feature fusion. The results showed that the accuracy of spatial domain clustering decreased greatly if the spatially guided cross attention or multi-scale feature fusion were removed from the framework (Supplementary Table 1), indicating that these two modules contributed greatly the representations learned from the noisy spatial

multi-omics data. Finally, we investigated the computational resources occupied by these methods, and the results showed that SpaMM-Net was run efficiently in all selected datasets (Supplementary Tables 2 and 3).

Discussion and conclusion

In this work, we proposed the SpaMM-Net to generate the latent representations by combining the spatially-guided multimodal feature fusion and multiscale feature integration. As spatial sequencing technologies continue to evolve, future platforms are expected to support simultaneous profiling of an increasing number of omics modalities. In general, spatial multi-omics data are high-dimensional and noisy, even with inconsistencies among different modalities in some local regions. Based on the encoder-decoder framework, SpaMM-Net utilizes spatially-guided cross attention and multi-scale feature fusion modules to leverage spatial context to guide multimodal integration. The ablation study indicated that these two components could capture the relevant information from neighbors to reduce the local noises, thus providing a new avenue to effectively integrate spatial multi-omics data. In addition, the complex models often suffer from training instability due to the presence of numerous hyperparameters. To address this issue, SpaMM-Net utilizes an uncertainty-based loss weighting strategy that enabled the model to autonomously learn and adjust the contributions of different loss terms. Our comparisons show that these strategies are effective in different kinds of datasets. Generally, SpaMM-Net achieves similar performance compared to SpatialGlue. SpaMM-Net also shows better performance on some datasets, such as the multi-modal data with higher noise or discrepancy among modalities, partly because SpaMM-Net utilizes the spatially-guided cross attention and multiscale network to integrate modality-specific features and capture spatial structures at different scales. Thus, SpaMM-Net provides an alternative option to perform spatial domain clustering from spatial multi-omics data.

There exist several limitations in our work. First, we used Allen Brain Atlas and histological images to manually annotate the coronal and sagittal mouse brains for method comparisons. Though the major annotated brain regions resemble the reference brain map, these annotations cannot be completely accurate due to the difficulty in exactly aligning histological images to reference brain map, especially for some sub-regions. To mitigate the negative effect of potentially inaccurately annotated regions, we used both supervised metric ARI and unsupervised metric Moran's I to evaluate the methods. Second, we mainly evaluated our methods in three real datasets, which were manually annotated according to reference map or prior knowledge. To overcome the absence of well-defined annotations, we also used the simulated data to evaluate our method. However, the results derived from the simulated dataset and limited real datasets cannot completely represent all real cases. For example, our method achieved better performance than SpatialGlue in some cases, but it did not always outperform SpatialGlue. Third, in the comparisons, we performed spatial clustering by using the same unsupervised clustering algorithm on the latent representations derived from all selected methods. This means that we excluded the effect derived from clustering algorithms implemented in the existing methods when evaluating the learned representations from spatial multi-omics data. Other clustering algorithms or computational settings, such as the high-quality histological images, may be used to achieve better performance in some methods. Finally, we used different random seeds to preliminarily estimate the network uncertainties to some extent. However, the uncertainty estimation and stability selection of these models should be further investigated in the future, since they play crucial roles in extracting information from high-dimensional multi-omics data [22, 23]. Taken all together, further benchmark is still needed to fairly evaluate the methods in representation learning from spatial multi-omics data.

In summary, we developed a computational tool SpaMM-Net to integrate spatially-guided graph attention networks and multiscale feature integration for representation learning from spatial multi-omics data in this work. The model effectively captures and integrates latent representations from multiple omics modalities, while expanding the receptive fields via stacked k-hop graph convolutions to obtain spatially contextualized embeddings at different semantic levels. We evaluated SpaMM-Net on different spatial multi-omics datasets. The results demonstrated that the latent representations generated by SpaMM-Net were effective for downstream spatial clustering tasks with generally better performance in terms of clustering quality and spatial coherence.

Abbreviations

SpaMM-Net	Spatially multimodal and multiscale network
KNN	k nearest neighbor
PCA	Principal component analysis
SC-Attn	Spatially-guided cross-attention
SGU	Cross-scale gating unit
MSF	Multi-scale feature fusion
GAT	Graph attention network

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12945-y.

Supplementary Material 1.

Availability and requirements

Project name: SpaMM-Net
Project home page: https://github.com/CPenglab/SpaMM-Net

Operation System(s): Linux
Programming language: Python 3.10.16
Other requirements: numpy 2.3.2, pandas 2.3.1, scanpy 1.11.4, scikit_learn 1.7.1, scipy 1.16.1, torch 2.5.1+cu124, torch_geometric 2.6.1
License: GPL-3.0

Authors' contributions
F.Z. designed and developed the software, and C.P. administrated the project. F.Z. and C.P. wrote the manuscript.

Funding
This work was supported by National Natural Science Foundation of China [32170662] and Yunnan Fundamental Research Project [202401AS070131].

Data availability
The software and source code are freely available at GitHub (https://github.com/CPenglab/SpaMM-Net). The data that support the findings of this study are available from NCBI and NODE. Mouse brain coronal section and mouse embryo: GEO accession number GSE205055, (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205055); mouse brain sagittal section: NODE, National Omics Data Encyclopedia, (https://www.biosino.org/node/project/detail/OEP003285). The reference maps of coronal mouse brain section and sagittal mouse brain section were downloaded from Allen Brain Atlas with following URLs (https://atlas.brain-map.org/atlas?atlas=1#atlas=1&plate=100960340&structure=961&x=5279.99951171875&y=3744&zoom=-4&resolution=16.75&z=5, https://atlas.brain-map.org/atlas?atlas=181276160#atlas=181276160&plate=100837249&structure=15818&x=3406.999755859375&y=2468.000012397766&zoom=-3&resolution=7.92&z=6).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 2 November 2025 / Accepted: 11 May 2026
Published online: 19 May 2026

References

1. Cheng M, Jiang Y, Xu J, Mentis AA, Wang S, Zheng H, Sahu SK, Liu L, Xu X. Spatially resolved transcriptomics: a comprehensive review of their technological advances, applications, and challenges. J Genet Genomics. 2023;50(9):625–40.
2. Dou SQ, Zhang FY, He YJ, Du P, Deng Y, Shao MK, Yang GX, Wu KJ, Zeng YQ, Peng C, et al. PRRX1 is a key regulator in the phenotypic transition between human normal dermal and keloid fibroblasts. J Dermatol Sci. 2025;119(1):28–39.
3. Hong Y, Song K, Zhang Z, Deng Y, Zhang X, Zhao J, Jiang J, Zhang Q, Guo C, Peng C. The spatiotemporal dynamics of spatially variable genes in developing mouse brain revealed by a novel computational scheme. Cell Death Discov. 2023;9(1):264.
4. Yang J, Li Y, Tang Y, Yang L, Guo C, Peng C. Spatial transcriptome reveals the region-specific genes and pathways regulated by Satb2 in neocortical development. BMC Genomics. 2024;25(1):757.
5. Liu L, Chen A, Li Y, Mulder J, Heyn H, Xu X. Spatiotemporal omics for biology and medicine. Cell. 2024;187(17):4488–519.
6. Bressan D, Battistoni G, Hannon GJ. The dawn of spatial omics. Science. 2023;381(6657):eabq4964.
7. Jiang F, Zhou X, Qian Y, Zhu M, Wang L, Li Z, Shen Q, Wang M, Qu F, Cui G, et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. Nat Methods. 2023;20(7):1048–57.
8. Zhang D, Deng Y, Kukanja P, Agirre E, Bartosovic M, Dong M, Ma C, Ma S, Su G, Bao S, et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. Nature. 2023;616(7955):113–22.
9. Velten B, Stegle O. Principles and challenges of modeling temporal and spatial omics data. Nat Methods. 2023;20(10):1462–74.
10. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87. e3529.
11. Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, Yosef N. MultiVI: deep generative model for the integration of multimodal data. Nat Methods. 2023;20(8):1222–31.
12. Coleman K, Schroeder A, Loth M, Zhang D, Park JH, Sung JY, Blank N, Cowan AJ, Qian X, Chen J, et al. Resolving tissue complexity by multimodal spatial omics modeling with MISO. Nat Methods. 2025;22(3):530–8.
13. Long Y, Ang KS, Sethi R, Liao S, Heng Y, van Olst L, Ye S, Zhong C, Xu H, Zhang D, et al. Deciphering spatial domains from spatial multi-omics with SpatialGlue. Nat Methods. 2024;21(9):1658–67.
14. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382–3.
15. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
16. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2018;2018:7482–7491.
17. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep-Uk. 2019;9.
18. Anselin L. Local Indicators of Spatial Association - Lisa. Geogr Anal. 1995;27(2):93–115.
19. Li HC, Zhang ZQ, Squires M, Chen X, Zhang XW. scMultiSim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell-cell interactions. Nat Methods. 2025;22(5):982–93.
20. Petar Veličković GC, Casanova A, Romero A. Pietro Liò, Yoshua Bengio: Graph Attention Networks. in Proc 6th ICLR. 2018. p. 1–12.
21. Daigle TL, Madisen L, Hage TA, Valley MT, Knoblich U, Larsen RS, Takeno MM, Huang L, Gu H, Larsen R, et al. A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced Brain-Cell-Type Targeting and Functionality. Cell. 2018;174(2):465–80.
22. Meinshausen N, Bühlmann P. Stability selection. J Roy Stat Soc B. 2010;72:417–73.
23. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High Throughput. 2019;8(1).

Publisher's Note