OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae719

https://doi.org/10.1093/bib/bbae719

Review

Deep learning in integrating spatial transcriptomics with other modalities

Jiajian Luo, Jiye Fu, Zuhong Lu*, Jing Tu*

State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, 2 Sipailou, Xuanwu District, Nanjing 210096, China

*Corresponding authors. Jing Tu, E-mail: jtu@seu.edu.cn; Zuhong Lu, E-mail: zhlu@seu.edu.cn

Abstract

Spatial transcriptomics technologies have been extensively applied in biological research, enabling the study of transcriptome while preserving the spatial context of tissues. Paired with spatial transcriptomics data, platforms often provide histology and (or) chromatin images, which capture cellular morphology and chromatin organization. Additionally, single-cell RNA sequencing (scRNA-seq) data from matching tissues often accompany spatial data, offering a transcriptome-wide gene expression profile of individual cells. Integrating such additional data from other modalities can effectively enhance spatial transcriptomics data, and, conversely, spatial transcriptomics data can supplement scRNA-seq with spatial information. Moreover, the rapid development of spatial multi-omics technology has spurred the demand for the integration of spatial multi-omics data to present a more detailed molecular landscape within tissues. Numerous deep learning (DL) methods have been developed for integrating spatial transcriptomics with other modalities. However, a comprehensive review of DL approaches for integrating spatial transcriptomics data with other modalities remains absent. In this study, we systematically review the applications of DL in integrating spatial transcriptomics data with other modalities. We first delineate the DL techniques applied in this integration and the key tasks involved. Next, we detail these methods and categorize them based on integrated modality and key task. Furthermore, we summarize the integration strategies of these integration methods. Finally, we discuss the challenges and future directions in integrating spatial transcriptomics with other modalities, aiming to facilitate the development of robust computational methods that more comprehensively exploit multimodal information.

Keywords: deep learning; spatial transcriptomics; integration; image; scRNA-seq; multi-omics

Introduction

Cells within tissues are organized in specific patterns that are crucial for their function. This organization varies significantly depending on the tissue type and its role in the body. Therefore, preserving spatial context is essential for studying tissue architecture and cellular interactions. A key advancement in understanding these complex biological systems has been the advent of spatial transcriptomics, which reveals spatial patterns of gene expression. By providing insights into the spatial organization of gene activity, spatial transcriptomics has significantly enhanced our knowledge of tissue function and disease processes [1–4]. Spatial transcriptomics technologies are primarily classified into two categories: image-based and sequencing-based [5]. Early-stage image-based technologies, such as MERFISH [6], STARmap [7], and seqFISH+ [8], offer single-cell resolution but are generally limited to a few hundred genes, which may not fully capture the transcriptome's complexity. Nevertheless, recent commercial image-based platforms, including CosMx (NanoString), MERSCOPE (Vizgen), and Xenium In Situ (10x Genomics), enable measurement of the expression levels of hundreds to thousands of genes at a subcellular level. Conversely, 10x Visium [9], a popular sequencing-based technology, quantifies transcriptome-wide gene expression levels within ~55 μm spots that often contain multiple cells. More recent sequencing-based technologies, like Stereo-seq [10] and Seq-Scope [11], offer a subcellular spatial resolution of <1 μm but suffer from high dropout events.

Moreover, platforms that provide spatial transcriptomics data often include paired high-resolution image data, such as histology and chromatin images, which complement cellular morphology and chromatin organization information [7, 9, 12, 13]. Single-cell RNA sequencing (scRNA-seq) profiles the whole transcriptome at the single-cell level despite lacking spatial information [14]. These additional data can serve as valuable resources to enhance the utility of spatial transcriptomics data. For example, integrating histology images can improve the resolution of spatial transcriptomics data and accurately identify spatial domains [12]. Furthermore, integrating scRNA-seq data enables missing gene imputation and cell deconvolution for spatial transcriptomics data [15, 16]. In turn, spatial transcriptomics data aid in spatial location reconstruction for scRNA-seq data [15, 16]. As spatial multi-omics technologies advance, integrating spatial multi-omics data is essential to reveal refined tissue architecture and facilitate downstream analysis.

However, integrating multimodal data presents significant computational challenges due to their inherent complexity and variability [17]. Spatial transcriptomics data and scRNA-seq data, typically represented as gene-by-cell matrices, still have a domain gap (Fig. 1A and C). Besides, spatial transcriptomics data

Received: September 9, 2024. **Revised:** November 5, 2024. **Accepted:** December 30, 2024

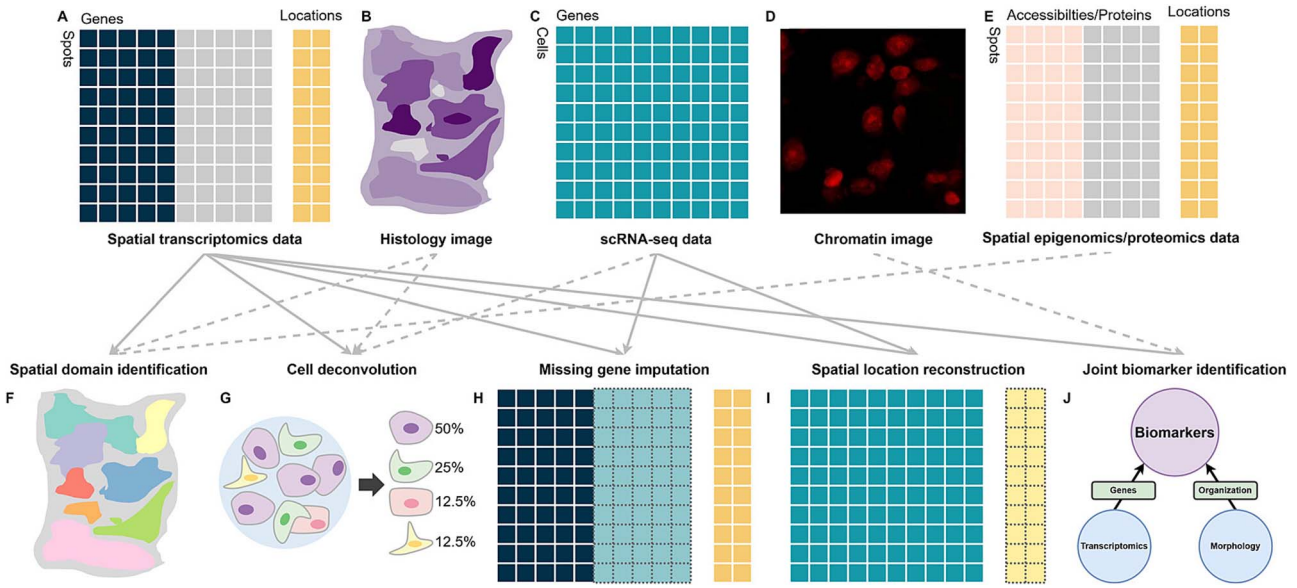

Figure 1. Spatial transcriptomics data, histology images, scRNA-seq data, and prevalent tasks in integrating spatial transcriptomics data with other modalities. (A) Spatial transcriptomics provides gene expression patterns along with corresponding spatial locations, with gray squares indicating unmeasured genes. (B) Histology images, paired with spatial transcriptomics data, capture cellular morphology and tissue architecture. (C) scRNA-seq technologies measure gene expression levels of individual cells, though spatial information is lost during tissue dissociation. (D) Chromatin images, paired with spatial transcriptomics data, reveal nuclear morphology and chromatin organization of individual cells. Reproduced with permission from *Nature Communications* under the Creative Commons Attribution 4.0 International License [67]. (E) Spatial epigenomics/proteomics data, paired with spatial transcriptomics data, offer chromatin accessibility and protein abundance information, with gray squares indicating unmeasured features. (F) Spatial domain identification involves identifying regions of a tissue sample with coherent gene expression and histology. (G) Cell deconvolution infers cell-type composition of spots. (H) Missing gene imputation predicts unmeasured gene expression levels. The blue-green squares with gray dashed lines refer to the imputed gene expression levels in spatial transcriptomics data. (I) Spatial location reconstruction predicts spatial locations of cells in scRNA-seq data. (J) Joint biomarker identification aims to identify multimodal biological markers. The light-yellow squares with gray dashed lines refer to the predicted spatial location of cells in scRNA-seq data. Solid arrows denote indispensable modalities for the task, while dashed arrows represent optional modalities.

include unique spatial information. Histology and chromatin images, which are essentially 2D arrays of pixel values, exhibit high heterogeneity compared to sequencing data (Fig. 1B and D). Furthermore, in spatial multi-omics data, the distribution disparities among different modalities are considerable, and the number of features across these modalities is imbalanced (Fig. 1E). Therefore, sophisticated computational methods are critical to fully exploit the potential of these multimodal data.

In this context, deep learning (DL) has emerged as a transformative approach in computational biology, characterized by its ability to process and analyze complex, high-dimensional data through multilayered neural networks [18, 19]. Its robust capabilities make DL adapted for integrating spatial transcriptomics data with other modalities, tackling the inherent complexity of these datasets. Compared to conventional statistical methods, DL excels in extracting meaningful features from sequencing data and images and is more versatile in integrating spatial transcriptomics with other modalities, especially heterogeneous data like histology and chromatin images. The recent emergence of various DL methods for spatial transcriptomics integration demonstrates its effectiveness. However, a comprehensive review of DL approaches for integrating spatial transcriptomics data with other modalities is lacking, and existing reviews do not cover the latest developments in DL approaches.

In this study, we identify 22 DL methods for spatial transcriptomics integration with other modalities, including histology images, scRNA-seq, chromatin images, and other spatial omics data. We classify these methods based on integrated modality and key tasks while summarizing their respective integration strategies. Our review begins with an overview of DL techniques in the context of spatial transcriptomics data integration, followed by an introduction to the prevalent key tasks in this domain. Subsequently, we elaborate on the DL methods tailored for spatial transcriptomics data integration. The manuscript concludes with a discussion of the prevailing challenges and future directions.

Overview of DL techniques in spatial transcriptomics data integration with other modalities

A range of DL techniques are employed for integrating spatial transcriptomics data with other modalities.

Convolutional neural networks (CNNs) and graph neural networks (GNNs) are specialized architectures for image and graph data, respectively. CNNs use convolutional and pooling layers alongside nonlinear activations to efficiently capture local patterns and construct complex representations. This design reduces the parameter count compared to fully connected neural networks, mitigating the risk of overfitting. CNNs are frequently used for feature extraction from histology and chromatin images. Moreover, spatial transcriptomics data, which can be considered as images with hundreds to thousands of channels, can be processed by CNNs potentially. GNNs iteratively update node representations by aggregating information from adjacent nodes, encapsulating both local and global contexts. Recent advancements in GNNs, such as graph convolutional networks (GCNs) and graph attention networks (GATs), have been widely applied to integrate gene expression data with spatial and histology information.

Unsupervised DL techniques for spatial transcriptomics data integration with other modalities encompass contrastive,

generative, and adversarial approaches. Contrastive learning utilizes the concept of similarity, training models to bring closer the representations of similar samples while pushing apart those of dissimilar ones. This promotes the model to learn common features among similar samples. Generative models, including autoencoders (AEs) and variational autoencoders (VAEs), are trained using a reconstruction loss. An AE comprises an encoder that maps input to a latent space and a decoder that reconstructs the input, facilitating the learning of latent representations from gene expression data and images. VAEs extend AEs by introducing probabilistic encodings, ensuring that the latent representations closely resemble the prior distribution. Compared to AEs, VAEs excel in both feature extraction and data generation, such as imputing missing gene expression levels. Adversarial learning employs a discriminator to differentiate latent representations across modalities and trains the encoders to fool the discriminator. Through this adversarial process, the encoders map modalities (e.g. scRNA-seq data and spatial transcriptomics data) to a shared latent space, extracting underlying features common to multiple modalities.

Additionally, attention mechanisms have been developed to enable models to weigh the importance of each part of the input data, regardless of their distance. This allows models to capture long-range dependencies more effectively and improve model interpretability. For example, when integrating representations from multiple modalities, an attention layer can dynamically assign weights based on the specific input data, avoiding the need to set a fixed hyperparameter rigidly. Multihead attention further refines this approach by employing multiple sets of attention weight. The Transformer [20] architecture harnesses these mechanisms within its structure, achieving unprecedented success in natural language processing and computer vision [21, 22] and serving as a powerful choice for processing gene expression and image data.

Prevalent tasks in integrating spatial transcriptomics data with other modalities

In this review, we examine 9 DL methods for integrating spatial transcriptomics data with histology images, 10 methods with scRNA-seq data, 1 method with chromatin images, and 2 methods for spatial multi-omics integration. These methods are further categorized based on key task and integration strategy. Table 1 summarizes current DL methods for spatial transcriptomics integration with other modalities. Supplementary Table S1 provides descriptions of datasets applied to these DL methods. Supplementary Table S2 presents prevalent metrics utilized for benchmarking. Before delving into these methods, we introduce prevalent tasks in integrating spatial transcriptomics data with other modalities, including spatial domain identification, cell deconvolution, missing gene imputation, and spatial location reconstruction. Figure 1 illustrates multiple data modalities and these tasks.

'Spatial domain identification' is a critical step in spatial transcriptomics data analysis (Fig. 1F). It aims to identify regions of a tissue sample with coherent gene expression and histology. Traditional clustering methods rely exclusively on gene expression data, overlooking the inherent spatial relationships between cells. Recent methods leverage additional information, such as spatial coordinates and histology images, to more accurately identify spatial domains.

Many spatial transcriptomics technologies capture gene expression from multiple cells within a single spot, rather than at single-cell resolution [5]. 'Cell deconvolution' refers to inferring the proportions of different cell types present in each spatial location (Fig. 1G). Approaches for cell deconvolution can be divided into two classes: reference-based and reference-free. Reference-based methods use scRNA-seq data, which undergo clustering and differential expression analysis for cell type composition prediction. Reference-free methods utilize techniques like latent Dirichlet allocation [23] and archetypal analysis [24] to infer transcriptional patterns for each cell state.

'Missing gene imputation' in spatial transcriptomics predicts the expression levels of unmeasured genes in spatial transcriptomics data (Fig. 1H), primarily applied to image-based spatial transcriptomics technologies [5]. This approach addresses the limitations of these technologies, which offer superior spatial resolution but typically measure only hundreds of preselected genes due to indexing scheme constraints. By leveraging complementary scRNA-seq data, missing gene imputation can enhance spatial transcriptomics data with transcriptome-wide information.

'Spatial location reconstruction' integrates scRNA-seq data with spatial transcriptomics data, assigning single-cell gene expression profiles to precise spatial coordinates within a tissue sample (Fig. 1I). This allows for the identification of spatially variable cell subpopulations and the investigation of cell–cell interactions in their spatial context.

Biomarkers, encompassing genes, proteins, and morphological features, function as measurable indicators of biological states or conditions. 'Joint biomarker identification' integrates multiple data modalities to identify multimodal biological markers (Fig. 1J). Such biomarkers provide a comprehensive understanding of disease mechanisms and progression, offering a more robust and reliable perspective.

Spatial transcriptomics data integration with histology images

Histology images, such as those obtained through hematoxylin and eosin (H&E) staining, are often provided by sequencing-based platforms [12, 13]. These images offer high-resolution insights into cellular morphology, capturing critical details about cell size, shape, and arrangement, despite inherent noise. Integrating histology images is motivated by the premise that spots with similar histology characteristics are more likely to belong to the same spatial domain and are more likely to share similar cell compositions. The primary objective of most integration methods is to harness the detailed cellular morphology captured in histology images to facilitate spatial transcriptomics data analysis. Additionally, some methods leverage the superior resolution of histology images to improve the resolution of spatial transcriptomics data.

Spatial domain identification

Most studies integrating spatial transcriptomics data with histology images aim to enhance spatial domain identification. Tools such as SpaCell [25], stMVC [26], DeepST [27], ConGI [28], and TransformerST [29] are designed to learn a joint latent representation, thereby improving domain segmentation, while SpaGCN [30] incorporates clustering within its training process to directly identify spatial domains. To evaluate the efficacy of these methods, researchers often rely on established datasets. Although different studies use varying datasets to assess their methodologies, the human dorsolateral prefrontal cortex dataset [31], which is manually annotated, has emerged as a popular benchmark for evaluating spatial domain identification methods. This dataset provides a reliable reference point for assessing the accuracy of algorithms. In terms of performance metrics, the adjusted Rand index (ARI), normalized mutual information, and homogeneity (HOM) are frequently utilized to provide insights into how

Table 1. Deep learning methods for integrating spatial transcriptomics data with other modalities.

Integrated modality	Year	Tool	Model	Integration strategy	Key task
Integrating spatial transcriptomics with histology images	2019	SpaCell [25]	CNN + AE	Link-based	Spatial domain identification
	2021	SpaGCN [30]	GCN	Graph-based	Spatial domain identification
	2022	stMVC [26]	CNN + GAE	Sum-based & Graph-based	Spatial domain identification
		DeepST [27]	GVAE	Graph-based	Spatial domain identification
	2023	ConGI [28]	Contrastive learning	Sum-based & Fusion-based	Spatial domain identification
		TESLA [40]	CNN	Link-based	Spatial domain identification
		stLearn [34]	CNN	Graph-based	Spatial domain identification
	2024	TransformerST [29]	Transformer	Link-based	Spatial domain identification
		Starfysh [24]	VAE	Sum-based	Cell deconvolution
Integrating spatial transcriptomics with scRNA-seq	2019	gimVI [54]	VAE	Fusion-based	Missing gene imputation
	2021	DSTG [48]	GCN	Graph-based	Cell deconvolution
		stPlus [57]	AE	Fusion-based	Missing gene imputation
	2022	CellDART [51]	ADDA	Fusion-based	Cell deconvolution
		DestVI [44]	VAE	Fusion-based	Cell deconvolution
		SD^2 [49]	GCN	Graph-based	Cell deconvolution
	2023	GraphST [53]	Contrastive learning + GAE + AE	Fusion-based	Cell deconvolution
		GTAD [50]	GAT	Graph-based	Cell deconvolution
	2024	STEM [61]	AE	Fusion-based	Spatial location reconstruction
		ENVI [60]	VAE	Fusion-based	Missing gene imputation
Integrating spatial transcriptomics with chromatin images	2022	STACI [67]	GVAE+VAE	Fusion-based	Joint biomarker identification
Spatial multi-omics integration	2024	SpatialGlue [68]	GAE	Sum-based	Spatial domain identification
		PRAGA [69]	GAE	Link-based	Spatial domain identification

well the clustering results align with the true spatial domains (Supplementary Table S2).

SpaCell [25] is the first study to combine histology images and gene expression data for cell clustering. It independently normalizes histology images and gene expression data, segments the images into 299 × 299 pixel tiles, each containing a single spot, and employs a pretrained ResNet50 [32] model to extract tile embeddings. These embeddings and the corresponding gene expression data are fed into two AEs, and their latent representations are concatenated for *K*-means or Louvain clustering [33]. Additionally, SpaCell uses a two-layer Deep Neural Network (DNN) to predict the disease stage of spots based on their image embeddings and gene expression data. Despite its promising performance, SpaCell learns latent representations without incorporating spatial information.

Inspired by SpaCell, Pham *et al.* [34] developed stLearn to incorporate gene expression data, spatial and histology information for normalization. A pretrained ResNet50 network extracts features from histology images to compute histological similarity. For each spot S_i, the model identifies the three neighboring spots with highest weights for normalization. The weighting matrix is calculated as follows:

$$W_{i,j} = \frac{GD_{i,j} \cdot MD_{i,j}}{\sum_{j=1}^{n} (GD_{i,j} \cdot MD_{i,j})}, \quad (1)$$

where $GD_{i,j}$ and $MD_{i,j}$ represent gene expression correlation and histology similarity between spot S_i and S_j, respectively, and n denotes the number of selected spots. stLearn normalizes gene expression data as follows:

$$GE_i' = \frac{1}{2} GE_i + \frac{1}{2} \left\{ \sum_{j=1}^{n} (W_{i,j} \cdot GE_j) \,|S_j \right\}, \quad (2)$$

where S_j denotes the selected spots for normalization. GE'_i represents normalized gene expression of spot S_i, while GE_i and GE_j represent raw gene expression of spots S_i and S_j. Next, stLearn

identifies broad clusters at a global level using normalized data, followed by a refinement phase that enables subclustering based on spatial segregation within tissue sections. In addition, stLearn includes two other modules for spatio-temporal cell trajectories reconstruction and cell–cell interaction analysis. However, both SpaCell and stLearn employ a pretrained ResNet50 network trained on a dataset of nonhistology images, which may not fully capture significant patterns from histology images. An independent benchmarking study [35] shows that the old version of stLearn has generally surpassed SpaCell.

In contrast, SpaGCN [30] converts spatial transcriptomics data into a graph-structured format and utilizes the GCN to integrate gene expression, histology images, and spatial information. SpaGCN employs PCA for dimension reduction of gene expression data after normalization with the top 50 principal components serving as node embeddings. The edge weight between any two spots is calculated using the Euclidean distance, which incorporates the spatial coordinates and an additional third dimension, z, derived from the the mean color value for the red, green, and blue (RGB) channels in the histology image. The GCN takes the graph as input and outputs aggregated representations, upon which the Louvain method is applied for clustering. The parameters of the GCN are optimized with the iterative refinement of clusters. Furthermore, SpaGCN employs the Wilcoxon rank-sum test to identify spatially variable genes (SVGs). If no SVG is identified, SpaGCN uses a set of genes to formulate a meta gene significantly highly expressed within a specific spatial domain. Although spaGCN outperforms the older version of stLearn, a recent benchmarking study [36] indicates that the latest version of stLearn exhibits slightly better clustering accuracy than spaGCN.

Zuo *et al.* [26] pointed out SpaGCN's inability to utilize the textural features of spots. They introduced stMVC, semisupervised graph attention autoencoders that learn robust representations of spatial transcriptomics data and histology images. stMVC initializes histology images via a ResNet50 model [32], which is trained by SimCLR framework [37] and outperforms the ResNet50 model pretrained by the ImageNet, which is used in SpaCell and stLearn. The model constructs two graphs: one representing histological similarity and the other representing spatial proximity between spots. The model utilizes an AE to compress the gene count matrix, encapsulating nonlinear gene relationships. The compressed gene count matrix is used for the feature matrix of the graphs. Graph attention autoencoders (GATEs) are used to learn representations of nodes. Considering the quality of the information in different views may be different, an attention layer is followed for representation weights assignment. The weighted sum of the two representations is applied for spot classification with manual region segmentation serving as a reference, enabling representations learned by GATEs under weak supervision simultaneously. Finally, the robust representations are used for spot clustering by the "FindNeighbors" and "FindClusters" function with default parameters from the Seurat package [38]. stMVC's flexible framework supports the addition of various omics data, enabling the integration of spatial multi-omics data with histology images. Using average silhouette width (ASW) as the metric (Supplementary Table S2), stMVC shows improved clustering accuracy over the previous version of stLearn, although its relative performance compared to other spatial domain identification methods integrating histology information remains unclear.

Similar to stLearn, DeepST [27] conducts data augmentation on gene expression data to integrate histology information as follows:

$$\widetilde{GE_i} = GE_i + \frac{\sum_{j=1}^{n} GE_j \cdot MS_{ij} \cdot GC_{ij}}{n}, \tag{3}$$

where GE_i and GE_j are the raw gene expressions for spot S_i and n adjacent spots S_j, GC_{ij} is the gene expression correlation between spot S_i and spot S_j, and MS_{ij} is the morphological similarity between spot S_i and adjacent spot S_j. DeepST utilizes a denoising autoencoder to effectively reduce the dimensionality of gene expression data. It then constructs a spatial graph and employs a graph variational autoencoder (GVAE) to learn the graph embeddings. Besides, DeepST applies domain adversarial neural networks to align embeddings from multiple batches or different spatial transcriptomics technologies. The Leiden method is applied for graph embedding clustering. Despite the advancements of stLearn and DeepST, they both adhere to a fixed radius for determining neighboring points during gene expression data augmentation. A benchmarking study for GNN-based methods reveal that DeepST's clustering accuracy is typically higher than SpaGCN's, while DeepST requires more computational resources since it comprises two deep neural networks.

Zeng *et al.* [28] indicated that SpaGCN only integrates histology images prior to training, and image noise may result in inaccurate spot relationships. They developed ConGI, which utilizes contrastive learning to filter noise within histology images for enhanced integration. ConGI augments image data and gene expression data by adding noise. It employs two independent encoders to extract features from gene expression and image data separately, subsequently projecting the embeddings from two modalities to a shared space. ConGI applies three contrastive learning losses to pull representations of paired data closer and push those of unpaired data apart: one for gene expression, one for images, and one for cross-modality (images to gene expression). At the end, the image and gene expression representations are combined and used for spatial domain identification by mclust [39]. ConGI outperforms SpaGCN in terms of ARI but suffers from poor interpretability and neglects the spatial information, potentially constraining its efficacy.

Hu *et al.* [40] proposed TESLA, a method utilizing the CNN for domain segmentation. Developed by the same research group behind the SpaGCN, TESLA employs the Canny edge detection algorithm to identify tissue regions, which are then divided into equal squares significantly smaller than the spots. Gene expression levels for these squares are imputed using the 10 nearest measured spots based on the Euclidean distance metric, similar to SpaGCN. TESLA constructs a meta gene from multiple marker genes and creates a meta gene image where each pixel represents the meta gene's expression level. The grayscale histology image and the meta gene image are then combined into a two-channel image, which undergoes unsupervised segmentation via a CNN, facilitating annotations from cell type to structure.

TransformerST [29] contains two advanced Transformer-based models—a Vision Transformer and an adaptive Graph Transformer—and a cross-scale internal graph network. It demonstrates remarkable efficiency in clustering and enhancing spatial transcriptomics data at single-cell resolution. The Vision Transformer processes spot-centric image patches and employs its decoder for gene expression prediction, compelling the encoder to learn a joint latent representation of gene expression and histology information. TransformerST then constructs a spatial graph for spots, concatenating the latent image embeddings with the gene expression matrix to form spot feature vectors. Consequently, the adaptive Graph Transformer incorporates spatial

information into the feature vectors, generating graph embeddings for spatial domain identification. Finally, the cross-scale internal graph network infers gene expression embeddings for each sliding-window histology image patch by initializing the patches through the Vision Transformer's encoder. Unlike methods that directly use image similarity to estimate gene expression, this model introduces learnable weights for each neighboring node. This allows for a more nuanced estimation of gene expression for sliding-window patches. Furthermore, besides achieving super-resolution clustering, TransformerST demonstrates superior clustering accuracy compared to SpaGCN, DeepST, and the latest version of stLearn.

Cell deconvolution

Similar to methods integrating histology images for spatial domain identification, the sole reference-free DL method for cell deconvolution also seeks to derive a robust joint latent representation of gene expression data and histology images for each spot.

Among the methods for cell deconvolution of spatial transcriptomics data, only a few are reference-free. Though they enable cell deconvolution without scRNA-seq data, He *et al.* [24] noted that previous reference-free methods fall short in refined cell state inference, scalability, multislice analysis capability, and utilization of histological image information. Inspired by the application of deep generative models in single-cell omics, they developed Starfysh, a VAE designed for cell deconvolution in spatial transcriptomics data without scRNA-seq as a reference. If available, the model can be enhanced by integrating paired histology images. It first determines the number of cell states and their marker genes through archetypal analysis [41]. Starfysh integrates multiple samples by identifying and updating gene markers based on sample-specific anchors, aggregating these markers across all samples. The model learns latent representations denoted as u_k for each cell state k via a neural network. The model uses a Dirichlet distribution to model c_k, the proportion of cell state k. View-specific encoders, θ_1 and θ_2, map spatial transcriptomics data and histology images to latent space, conforming to distributions $Normal\left(\mu_1, \sigma_1^2\right)$ and $Normal\left(\mu_2, \sigma_2^2\right)$, respectively. The posterior distribution $q_\theta\left(\mathrm{z}|\mathrm{c}, \mathrm{x}, \mathrm{y}\right)$ is parameterized as follows:

$$q_\theta\left(\mathrm{z}|\mathrm{c}, \mathrm{x}, \mathrm{y}\right) = \frac{\mu_1/{\sigma_1}^2 + \mu_2/{\sigma_2}^2}{1/{\sigma_1}^2 + 1/{\sigma_2}^2}, \tag{4}$$

where z represents the joint latent variables of gene expression data x and histology images y. The posterior distribution is constrained to approximate the prior distribution $p\left(\mathrm{z}|\mathrm{c}, u; \sigma\right)$ as follows:

$$p\left(\mathrm{z}|\mathrm{c}, u; \sigma\right) = Normal\left(\sum_k c_k u_k, \sum_k c_k \sigma_k\right), \tag{5}$$

where σ_k represents cell state-specific heterogeneity. The posterior c is parameterized by a neural network, while the prior c is determined by A, the enrichment score of the marker genes for cell states. View-specific decoders then reconstruct the original gene expression data and histology images from the latent variables z. Consequently, Starfysh disentangles spots across slices within the latent space and reconstructs cell type–specific gene expression levels to reveal cell states. It outperforms previous reference-free methods, including CARD [42], BayesTME [43], and STdeconvolve [23], in deconvolving both major and finer cell types. Additionally, it demonstrates comparable performance to state-of-the-art reference-based methods such as DestVI [44], Cell2location [45], Tangram [46], and BayesPrism [47] across Jensen–Shannon divergence (JSD) and root-mean-squared error (RMSE) metrics in both simulated and real datasets (Supplementary Table S2).

Spatial transcriptomics data integration with scRNA-seq data

scRNA-seq technologies enable the profiling of transcriptome-wide gene expression at the individual cell level, although spatial information is lost during tissue dissociation [14]. Assuming that the biological processes captured by scRNA-seq and spatial transcriptomics are fundamentally the same, the integration of spatial transcriptomics data and scRNA-seq can combine the strengths of both technologies. For low-resolution, sequencing-based spatial transcriptomics data, scRNA-seq data serve as a reference for cell deconvolution. For image-based transcriptomics data with limited gene measurement, scRNA-seq data assist in missing gene imputation. Conversely, for scRNA-seq data lacking spatial information, spatial transcriptomics data facilitate the spatial location reconstruction of individual cells.

Cell deconvolution

Methods for cell deconvolution assume that each spot in spatial transcriptomics data represents a mixture of multiple cell types. Therefore, some DL methods integrating scRNA-seq data for cell deconvolution generate pseudo-spots, which are mixtures of cells from scRNA-seq data. These pseudo-spots simulate complex cellular mixtures in real spatial transcriptomics data and serve as the ground truth for model training.

DSTG [48], SD^2 [49], and GTAD [50] all employ the GNN to integrate spatial transcriptomics data with scRNA-seq data for cell deconvolution. In preprocessing, DSTG and GTAD select highly variable genes, as is common in spatial transcriptomics integration methods. Conversely, SD^2 selects genes with a significantly high dropout rate. For graph construction, DSTG and SD^2 use the k-nearest neighbors (k-NN) method, whereas GTAD adopts a random projection forest, which obviates the need to determine a neighborhood size. Ultimately, GNN is applied to learn spot embeddings and predict the cell composition of pseudo-spots. Specifically, DSTG and SD^2 utilize the GCN, and GTAD employs the GAT to capture the correlations between pseudo-spots and real spots more effectively. However, these methods do not model the distribution discrepancy between pseudo-spots and real spots, which may diminish the deconvolution performance.

CellDART [51] considers the discrepancy and employs adversarial discriminative domain adaptation (ADDA) [52] to transfer the ability of cell composition prediction from pseudo-spots to real spots. It first trains a feature extractor and predictor to predict cell composition within pseudo-spots. Subsequently, real spots are fed to the feature extractor, and a discriminator is applied to map embeddings of pseudo-spots and real spots to a shared latent space. This adaptation enhances the model's performance in inferring cell composition for spatial transcriptomics data. Nevertheless, cell deconvolution methods based on pseudo-spots fail to infer the specific cell state for cells within spots.

Lopez *et al.* [44] observed that conventional methods, which treat cell type as a categorical variable, exhibit declining performance as the cell clustering resolution increases, due to the neglect of differences in similarity between cell types. To address this issue, they developed DestVI, a model that utilizes continuous latent embeddings to characterize cell types. DestVI assumes that cell states within a spot are similar for a specific cell type, thereby simplifying the task. A conditional VAE is employed to infer cell

type composition for each spot and the latent embedding of each cell type, representing the cell state. Furthermore, the decoder can reconstruct cell-type-specific gene expression for each spot, facilitating downstream analysis. An independent benchmark systematically assessed DestVI, DSTG, and SD^2 for cell deconvolution. It reveals that DestVI exhibits the most robust performance with both simulated and real-world datasets, achieving the highest Pearson correlation coefficient (PCC) and the lowest RMSE and JSD among these three methods (Supplementary Table S2). However, DestVI tends to yield average cell-type proportions in the Slide-seqV2 and stereo-seq datasets, resulting in suboptimal performance.

Unlike previous methods, GraphST [53] aims to learn a mapping matrix to project scRNA-seq data onto spatial transcriptomics for cell deconvolution through three modules. Module 1 employs a GAE with augmentation-based self-supervised contrastive learning to learn spot embeddings. Specifically, it constructs a neighborhood graph for spots and then randomly shuffles gene expression vectors across spots to generate a corrupted neighborhood graph. During model training, the spot embeddings are pulled closer to their neighborhood's embeddings in the real neighborhood graph and pushed away from their neighbor's embeddings in the corrupted neighborhood graph. This approach ensures that the spot embeddings effectively capture the local spatial context Module 2 aligns multiple slices and constructs a shared graph, enabling joint analysis of multiple tissue slices. Module 3 utilizes an AE to acquire cell embeddings from scRNA-seq data, reducing noise from sequencing technology. It then learns a mapping matrix denoting the probability of cells being projected onto each spot of the spatial transcriptomics data. In addition to the reconstruction loss, it employs contrastive learning to minimize differences in cell composition between adjacent spots, thereby capturing spatial information. Similar to generative models for cell deconvolution, such as DestVI, GraphST can reconstruct cell-type-specific gene expression to reveal cell states within spots. However, the neural networks for deriving cell and spot embeddings are trained independently, potentially leading to distribution discrepancies between these embeddings, which could impair the performance of cell deconvolution.

Missing gene imputation

Lopez *et al.* [54], from the same lab that developed DestVI, observed that previous methods embed sequencing data via linear model, which fail to capture nonlinear gene relationships. Moreover, these methods tend to overlay samples that do not exhibit substantial biological resemblance, as the alignment is conducted in an *ad hoc* manner. To address these issues, they extended scVI [55] to develop gimVI, a VAE augmented with an auxiliary binary neuron. Spatial transcriptomics data and scRNA-seq are fed to distinct nonlinear encoders with a shared final layer, followed by a shared nonlinear decoder for reconstruction. Both the encoder and decoder are conditioned on an auxiliary binary neuron within, which acts as a modality indicator. To align representations from two modalities, the model incorporates H-divergence between the two latent spaces to the loss function. gimVI accounts for biases inherent to diverse sequencing technologies by adopting conditional distributions specific to each method: Poisson for smFISH [56], negative binomial (NB) for STARmap [7], and either zero-inflated negative binomial (ZINB) or NB for scRNA-seq data. Gene imputation is performed by designating the modality indicator to scRNA-seq for representations derived from spatial transcriptomics data.

stPlus [57] employs a tailored loss function to efficiently generate embeddings for spatial transcriptomics and scRNA-seq data through an AE. Genes only measured by scRNA-seq are masked. The AE aims to learn meaningful embeddings by minimizing the reconstruction loss for spatial transcriptomics data and the masked value prediction error for masked scRNA-seq data. stPlus highlights the function designed to assess prediction error, which considers the dropout events in scRNA-seq and penalizes the prediction error of scRNA-seq data with high sparsity. To predict the expression of unmeasured genes in spots, stPlus employs the 50 nearest scRNA-seq cells for spot gene imputation, using a weighted k-NN method based on the distance between their embeddings. Spatial transcriptomics data imputed by stPlus demonstrate better clustering accuracy than those imputed by gimVI, indicating that stPlus identifies cell populations more effectively. However, compared to generative models, directly using scRNA-seq data to impute missing genes in spatial transcriptomics data may introduce technical bias. Indeed, independent benchmarks [58, 59] highlight gimVI's superior predictive capabilities to stPlus. For example, when assessing the average accuracy score (AS) across multiple datasets, gimVI's AS is 0.84, while stPlus's AS is 0.31 (Supplementary Table S2).

ENVI [60] postulates that a spot's microenvironment correlates with its gene expression. The model innovatively incorporates spatial context for gene imputation via a conditional VAE, where an auxiliary binary neuron specifies the input modality for both the encoder and decoder. This study introduces the Covariance Environment (COVET) matrix, which effectively integrates cellular gene expression with their microenvironment. ENVI employs a shared encoder to map spatial transcriptomics data and scRNA-seq data into a common latent space. During training, an expression decoder reconstructs expression profiles, while an environment decoder predicts the COVET matrices from spatial transcriptomics embeddings, thereby incorporating spatial context into the embeddings. In this way, the expression decoder can impute unmeasured genes for spatial data, and the environment decoder can reconstruct the spatial context of scRNA-seq data. The effectiveness of ENVI is benchmarked using the PCC and the multiscale spectral similarity index, a spatially aware metrics developed by the authors (Supplementary Table S2). Both metrics demonstrate ENVI's superior performance over gimVI in gene imputation.

Spatial location reconstruction

Hao *et al.* [61] developed STEM, an AE designed to predict a mapping matrix from scRNA-seq to spatial transcriptomics. This model leverages the biological assumption that gene expression profiles contain rich spatial information, which can be used to infer cellular localization and spatial relationships. STEM utilizes a shared encoder to learn embeddings from both spatial transcriptomics data and scRNA-seq data. The spatial adjacency matrix of spatial transcriptomics data is reconstructed via two methods: first, through the inner product of spatial transcriptomics embeddings, and second, by generating scRNA-seq to spatial transcriptomics and spatial transcriptomics to scRNA-seq mapping matrices through the inner product of single-cell and spatial transcriptomics embeddings. The model is trained using two reconstruction losses, combined with a maximum mean discrepancy loss to minimize the mean distance between spatial transcriptomics and single-cell embeddings. After prediction, the integrated gradient [62] is applied for model interpretation to identify spatially dominant genes. STEM significantly outperforms CellTrek [63], scSpace [64], Seurat [38], SpaOTsc [65], and

Tangram [46] in correctness of predicted cell-to-cell adjacency and is the only method that effectively reconstructs the spatial distribution of cells in simulated datasets.

Spatial transcriptomics data integration with chromatin images

Several spatial transcriptomics technologies, such as STARmap [7] and 10x Visium [9], provide spatial transcriptomics data and paired chromatin images. These images capture the nuclear morphology and chromatin organization of individual cells, which are valuable for reflecting tissue development and disease progression [66]. However, few methods utilize these images as complementary information in spatial transcriptomics data analysis. To our knowledge, there is only one study designed for the integration of spatial transcriptomics data and chromatin images.

Joint biomarker identification

STACI [67] comprises an overparameterized GVAE and a variational CNN autoencoder. Initially, it obtains latent representations of spatial transcriptomics data via the GVAE. The overparameterized architecture, where the size of hidden layers exceeds the dimensionality of the input feature space, is designed for batch effect correction in spatial transcriptomics data. Next, the variational CNN autoencoder is employed to acquire representations of chromatin images. In addition to the standard loss function of the VAE, STACI minimizes the distance between the spatial transcriptomic and image representation of each cell, thereby providing joint representations of both spatial transcriptomics and chromatin images. These joint representations are then utilized for downstream analysis, identifying combined morphometric and molecular biomarkers of disease progression.

Spatial multi-omics integration

Spatial multi-omics is an advanced field that combines multiple omics layers, such as genomics, transcriptomics, and proteomics, along with spatial context to provide a comprehensive view of the molecular landscape of tissues. However, integrating spatial multi-omics is challenging due to the significant discrepancies in distribution between different modalities and the imbalance in the number of features they possess. For example, proteins usually have only dozens to hundreds of features, while genes number in the tens of thousands.

Spatial domain identification

Long *et al.* [68] developed SpatialGlue, the first integration tool designed for spatial multi-omics data. It employs a GAE with a dual-attention mechanism at two levels. Considering that the spatial distribution of each cell type can be either discrete or continuous, SpatialGlue constructs two distinct graphs for features and spatial information within each modality. It utilizes a shared GCN in conjunction with an attention layer to integrate feature modality and spatial information representations. Another attention layer is then used to combine representations from multiple modalities. In addition to the conventional reconstruction loss, the model also incorporates a correspondence loss to align representations from various modalities.

Huang *et al.* [69] noted that SpatialGlue captures limited relations through the feature adjacency graphs constructed by K-Nearest Neighbors (KNN). To address this, they develop PRAGA, setting the adjacency matrix of the feature graphs as learnable parameters. PRAGA calculates the weighted sum of the adjacency matrices of both the feature graph and spatial graph, which are used as the edge weights between spots. A GAE is applied to learn the joint representation of modalities, while individual GCNs initially process distinct modalities. To avoid unstable training caused by significant changes in the adjacency matrices, the model adds an HOM loss to penalize changes between the adjacency matrix before and after updates. PRAGA exhibits higher clustering accuracy than SpatialGlue on identical benchmarking datasets, demonstrating its effectiveness.

Integration strategies

Integration strategies are methods for integrating information from various modalities. We identify four primary integration strategies: link-based, sum-based, graph-based, and fusion-based. Link-based and sum-based strategies are typically used to integrate spatial transcriptomics data with paired histology images or spatial omics data, while graph-based and fusion-based methods can combine spatial transcriptomics data with both images, scRNA-seq data and spatial omics data. Besides the integrated modality, the selection of integration strategy is also influenced by the key task. For example, three DL methods integrating scRNA-seq data for missing gene imputation all employ the fusion-based strategy.

Link-based

The link-based strategy involves concatenating representations of multiple modalities, which can occur in either data space or latent space (Fig. 2A). For example, SpaCell outputs concatenated histology and gene expression representations, while PRAGA follows this with an multilayer perceptron (MLP) for the final representation. TESLA merges histology image with a meta-gene image generated from spatial transcriptomics data to serve as the input for a CNN. Similarly, TransformerST combines histology representations with the gene expression matrix to characterize spots.

Sum-based

The sum-based strategy integrates data representations from various modalities by computing their weighted sum (Fig. 2B). Weights can be assigned as hyperparameters, as seen in ConGI, or learned within the neural network, as exemplified by stMVC, Starfysh, and SpatialGlue.

Graph-based

The graph-based strategy constructs graphs using multimodal information, often followed by a GNN-based model to learn node representations. It can be divided into two classes: "nodes and edges" and "dual nodes" (Fig. 2C). In the "nodes and edges" strategy, one modality is used as nodes, while another modality is used to calculate edge weight. stLearn, SpaGCN, stMVC, and DeepST use this method to integrate spatial transcriptomics data with histology images, where spots are nodes and histology images contribute to edge weight calculations alongside spatial information. In the "dual nodes" strategy, both modalities serve as nodes, with edges representing the similarity between nodes. DSTG, SD^2, and GTAD adopt this strategy to merge spatial transcriptomics data with scRNA-seq data, where each node represents a spatial transcriptomics spot or a cell in scRNA-seq data, and edges reflect gene expression similarity.

Fusion-based

The fusion-based strategy aims to map data from multiple modalities to a joint latent space to capture common features (Fig. 2D). This strategy employs two primary techniques: shared encoder projection and representation alignment. gimVI, ConGI, DestVI,

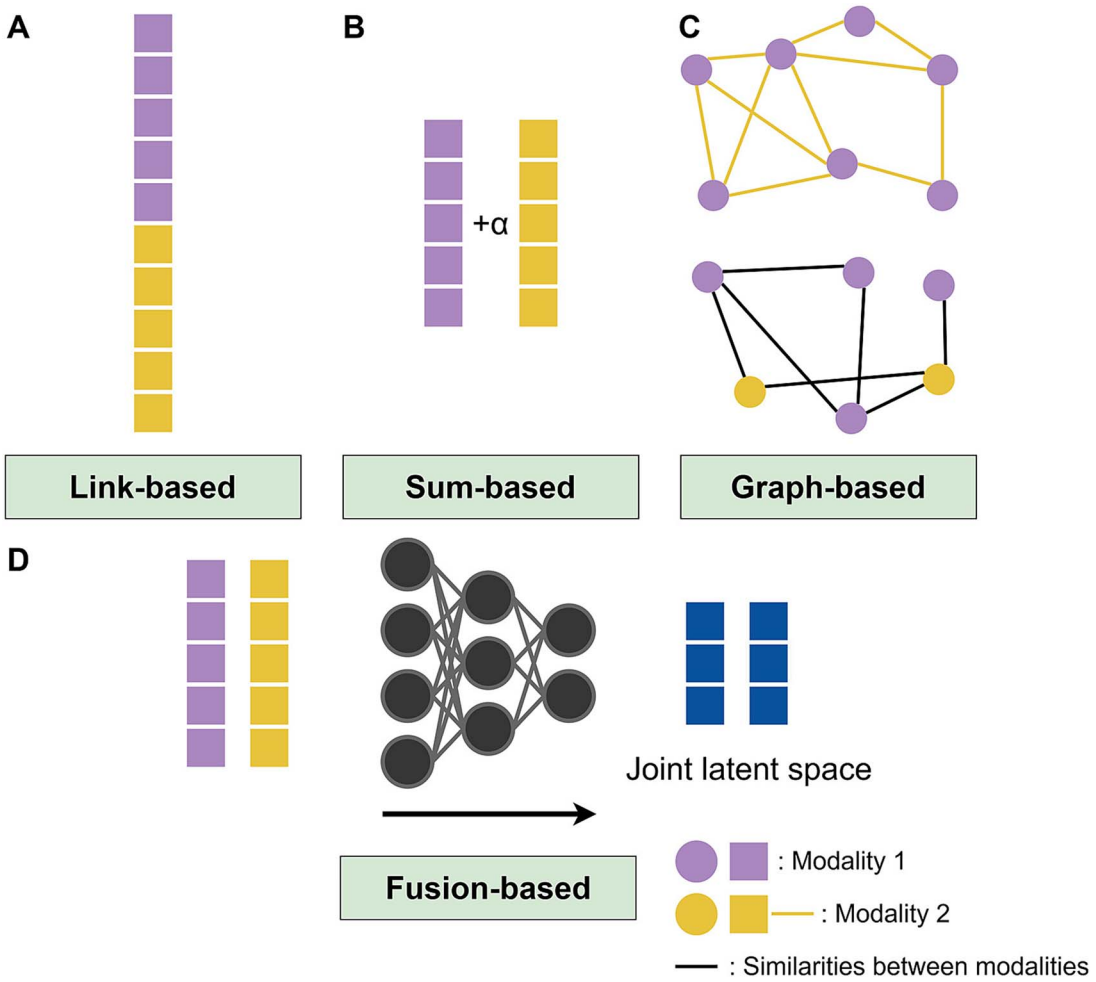

Figure 2. Integration strategies. (A) Link-based strategy: The representations of Modality 1 and Modality 2 are concatenated to form the final representations. (B) Sum-based strategy: The weighted sum of the representation of Modality 1 and the representation of Modality 2 serves as the final representation. (C) Graph-based strategy: In the "nodes and edges" subcategory (top), Modality 1 features nodes, while Modality 2 contributes to edges. In the "dual nodes" subcategory (bottom), both Modality 1 and Modality 2 feature nodes with edges representing the similarity between nodes. (D) Fusion-based strategy: Modality 1 and Modality 2 are mapped to a joint latent space.

and STACI use distinct encoders for each modality but align representations in the latent space. In contrast, methods like CellDART, STEM, stPlus, and ENVI adopt a shared encoder for multiple modalities, with CellDART and STEM ensuring representation alignment in the latent space. Additionally, both gimVI and ENVI are augmented with an auxiliary binary neuron. GraphST does not use specific methods to align the two modalities' representations. However, it learns a matrix to map one modality to another, assuming they share similar underlying features. Thus, it is classified within this category. These diverse implementations highlight the flexibility and effectiveness of the fusion-based strategy in integrating multimodal data.

Challenges and future directions

Biological systems operate through intricate networks, where various molecular components interact dynamically. While single data modalities provide valuable insights, they often fail to capture the comprehensive view necessary to fully understand these interactions. Leveraging the strengths of multiple data types allows researchers to achieve a more robust and nuanced understanding of biological processes and disease mechanisms.

Despite the emergence of DL methods for integrating spatial transcriptomics data with other modalities, several challenges remain. Firstly, the majority of DL integration methods are unavailable for multislice analysis of spatial transcriptomics data. An increasing number of spatial studies generate data from multiple slices to construct extensive spatial atlases, but batch effects within slices hinder their joint analysis. As such, it is imperative to develop tools equipped with batch effects removal capabilities to achieve multislice analysis. To address this need, STACI [67] employs an overparameterized autoencoder, and DeepST [27] uses adversarial learning. However, both methods do not account for spatial dependency between slices. GraphST [53] uses the PASTE algorithm [70] to align slices and construct a shared graph across multiple slices for batch correction via GNNs. Nevertheless, the PASTE algorithm itself does not incorporate histology information within slices.

The second challenge is the inherent noise within the data. Histology and chromatin images not only capture morphological features but also include extraneous information irrelevant to specific tasks, such as spatial domain identification and cell deconvolution. Integrating such images can potentially hamper the overall performance of models. For example, independent benchmarks [35, 36, 71] indicate that adding H&E staining images does not consistently enhance SpaGCN's performance. Furthermore, dropouts and technical effects in spatial omics and scRNA-seq data can confound biological variations. Hence, it

Table 2. Independent benchmarking studies of methods for specific tasks.

Benchmarking study	Task
Cheng *et al.* [35]	Spatial domain identification
Liu *et al.* [71]	Spatial domain identification
Yuan *et al.* [36]	Spatial domain identification
Li *et al.* [58]	Cell deconvolution & missing gene imputation
Yan and Sun [75]	Cell deconvolution
Li *et al.* [76]	Cell deconvolution
Sang-aram *et al.* [77]	Cell deconvolution
Avsar and Pir [59]	Missing gene imputation

is essential to filter out noise in multimodal integration via advanced DL techniques.

Thirdly, the scalability of computational methods is increasingly critical as spatial transcriptomics datasets grow in size. Early computational methods were designed for datasets with no more than 10 000 spots, as was the limit of spatial technology at the time. Nevertheless, rapid advancements in spatial technologies enable the simultaneous measurement of hundreds of thousands of locations per section [72]. Consequently, existing methods may struggle to maintain efficiency at such scales. To improve the scalability of models, one potential solution is to employ efficient neural network architectures that offer comparable or even higher performance with fewer computations.

Another challenge is the lack of interpretability in DL methods. For example, it is difficult to assess the contribution of spatial transcriptomics data and histology images to their joint representation obtained through graph-based or fusion-based strategies. Moreover, most studies in this field overlook interpretability analysis, which is essential for dissecting model decision-making processes, refining models, and uncovering biological insights [18]. For example, STEM [61] identifies spatially dominant genes by interpreting how the model predicts cell locations from gene expression data.

Moving forward, there remain new possibilities for the integration of spatial transcriptomics and other modalities. By combining spatial transcriptomics data with histology images and spatial omics data, researchers can improve cancer grading, subtyping, drug response predictions, and patient prognosis assessments. These areas are active fields of research, whereas previous studies were often limited to using unpaired multi-omics data and (or) histology images due to technical limitations. Nowadays, an increasing number of datasets provide spatial omics data paired with histology images. Integrating these data types is expected to significantly advance the field. Moreover, most research identifies spatial domains through transcriptomic heterogeneity. Alternatively, cell-type composition derived from cell deconvolution promises refined domain segmentation, simplifying the task. Recently, Ma *et al.* [73] have introduced the first method to leverage cell type composition for spatial domains detection via machine learning, assuming that spots within similar spatial domains share similar cell type composition. Therefore, we anticipate the development of DL methods based on this assumption for enhanced performance.

The rapid evolution of spatial technologies has created new opportunities for the integration of multimodal data. For example, emerging computational methods are specifically designed for spatial multi-omics integration. Given the success and versatility of DL in single-cell multi-omics data integration [74], researchers might extend these approaches to spatial data, thus driving forward the field of spatial multi-omics. Additionally, akin to transcriptomic data, spatial Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) can be integrated with scATAC-seq to complement their respective limitations, thereby offering a more comprehensive landscape of spatial chromatin accessibility.

As numerous DL integration methods have emerged, independent benchmarks are crucial for unbiased and comprehensive assessment. Table 2 presents recent benchmarking studies, which suggest that DL approaches have not demonstrated a notable advantage over non-DL methods in cell deconvolution and missing gene imputation. However, these studies have not evaluated the latest DL integration methods developed in the past 2 years. In addition, these benchmarks suffer from imbalanced data, as most spatial transcriptomics data selected for benchmarks are from normal tissues, potentially leading to bias in method evaluation. Future benchmarks are expected to incorporate more spatial transcriptomics datasets from diseased tissues and evaluate more recent DL integration methods.

In conclusion, our review systematically explores advancements in DL methods for integrating spatial transcriptomics data with other modalities. This emerging field holds significant promise for unlocking new biological insights and enhancing our understanding of complex biological systems. Future algorithmic development needs to not only focus on specific tasks but also consider critical factors such as multislice analysis capabilities, noise reduction, scalability, and interpretability. Moreover, developing computational methods for novel tasks and unexplored data types integration remains an important area for further research. We hope that the insights elucidated in this review offer a critical reference point for forthcoming studies and advance the field.

Key Points

- Additional data, including paired histology and chromatin images, as well as scRNA-seq data from matching tissues, serve as valuable resources to enhance the utility of spatial transcriptomics data.
- The integration strategy of deep learning methods in this review can be divided into four categories: concatenation, weighted-sum, graph-based, and joint embedding.
- Challenges in integrating spatial transcriptomics data with other modalities include scalability, inconsistent enhancement of spatial domain identification, and interpretability.
- Future directions for integrating spatial transcriptomics data with other modalities include possibilities such as cancer grading, subtyping, drug response predictions, patient prognosis assessments, and integration of unexplored data like scATAC-seq and spatial ATAC-seq.
- The datasets and prevalent evaluation metrics used in these integration methods are provided in the supplementary data.

Supplementary data

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

Funding

This work was supported by the National Natural Science Foundation of China (62371128), the Key Research and Development Project of Jiangsu Province (BE2022804), and the Fundamental Research Funds for the Central Universities (2242023 K5005).

Author contributions

J.T. and Z.L. conceived the review; J.L. and J.F. wrote the manuscript; J.T. and Z.L. revised the manuscript.

References

1. Hildebrandt F, Andersson A, Saarenpää S. *et al.* Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat Commun* 2021;**12**:7046. https://doi.org/10.1038/s41467-021-27354-w.
2. Crosse EI, Gordon-Keylock S, Rybtsov S. *et al.* Multi-layered spatial transcriptomics identify secretory factors promoting human hematopoietic Stem cell development. *Cell Stem Cell* 2020;**27**:822–839.e8. https://doi.org/10.1016/j.stem.2020.08.004.
3. Denisenko E, de Kock L, Tan A. *et al.* Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones. *Nat Commun* 2024;**15**:2860. https://doi.org/10.1038/s41467-024-47271-y.
4. Boyd DF, Allen EK, Randolph AG. *et al.* Exuberant fibroblast activity compromises lung function via ADAMTS4. *Nature* 2020;**587**:466–71. https://doi.org/10.1038/s41586-020-2877-5.
5. Rao A, Barkley D, França GS. *et al.* Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;**596**:211–20. https://doi.org/10.1038/s41586-021-03634-9.
6. Chen KH, Boettiger AN, Moffitt JR. *et al.* RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**:aaa6090. https://doi.org/10.1126/science.aaa6090.
7. Wang X, Allen WE, Wright MA. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**:eaat5691. https://doi.org/10.1126/science.aat5691.
8. Eng CL, Lawson M, Zhu Q. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 2019;**568**:235–9. https://doi.org/10.1038/s41586-019-1049-y.
9. Larsson L, Frisén J, Lundeberg J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat Methods* 2021;**18**:15–8. https://doi.org/10.1038/s41592-020-01038-7.
10. Chen A, Liao S, Cheng M. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;**185**:1777–1792.e21. https://doi.org/10.1016/j.cell.2022.04.003.
11. Cho CS, Xi J, Si Y. *et al.* Microscopic examination of spatial transcriptome using Seq-scope. *Cell* 2021;**184**:3559–3572.e22. https://doi.org/10.1016/j.cell.2021.05.010.
12. Hu J, Schroeder A, Coleman K. *et al.* Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Comput Struct Biotechnol J* 2021;**19**:3829–41. https://doi.org/10.1016/j.csbj.2021.06.052.
13. Burgess DJ. Spatial transcriptomics coming of age. *Nat Rev Genet* 2019;**20**:317. https://doi.org/10.1038/s41576-019-0129-z.
14. Aldridge S, Teichmann SA. Single cell transcriptomics comes of age. *Nat Commun* 2020;**11**:4307. https://doi.org/10.1038/s41467-020-18158-5.
15. Longo SK, Guo MG, Ji AL. *et al.* Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* 2021;**22**:627–44. https://doi.org/10.1038/s41576-021-00370-8.
16. Yan C, Zhu Y, Chen M. *et al.* Integration tools for scRNA-seq data and spatial transcriptomics sequencing data. *Brief Funct Genomics* 2024;**23**:295–302. https://doi.org/10.1093/bfgp/elae002.
17. Zahedi R, Ghamsari R, Argha A. *et al.* Deep learning in spatially resolved transcriptfomics: a comprehensive technical view. *Brief Bioinform* 2024;**25**:bbae082. https://doi.org/10.1093/bib/bbae082.
18. Novakovsky G, Dexter N, Libbrecht MW. *et al.* Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* 2023;**24**:125–37. https://doi.org/10.1038/s41576-022-00532-2.
19. Sapoval N, Aghazadeh A, Nute MG. *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* 2022;**13**:1728. https://doi.org/10.1038/s41467-022-29268-7.
20. Vaswani A. Attention is all you need. arXiv preprint arXiv:1706.03762 2017.
21. Wolf T, Debut L, Sanh V. *et al.* Transformers: State-of-the-Art Natural Language Processing. In: Liu Q, Schlangen D (eds). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45. Association for Computational Linguistics.
22. Khan S, Naseer M, Hayat M. *et al.* Transformers in vision: a survey. *ACM Comput Surv* 2022;**54**:1–41. https://doi.org/10.1145/3505244.
23. Miller BF, Huang F, Atta L. *et al.* Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat Commun* 2022;**13**:2339. https://doi.org/10.1038/s41467-022-30033-z.
24. He S, Jin Y, Nazaret A. *et al.* Starfysh integrates spatial transcriptomic and histologic data to reveal heterogeneous tumor-immune hubs. *Nat Biotechnol* 2024;1–13. https://doi.org/10.1038/s41587-024-02173-8.
25. Tan X, Su A, Tran M. *et al.* SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* 2020;**36**:2293–4. https://doi.org/10.1093/bioinformatics/btz914.
26. Zuo C, Zhang Y, Cao C. *et al.* Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nat Commun* 2022;**13**:5962. https://doi.org/10.1038/s41467-022-33619-9.
27. Xu C, Jin X, Wei S. *et al.* DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res* 2022;**50**:e131–1. https://doi.org/10.1093/nar/gkac901.
28. Zeng Y, Yin R, Luo M. *et al.* Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Brief Bioinform* 2023;**24**:bbad048. https://doi.org/10.1093/bib/bbad048.
29. Zhao C, Xu Z, Wang X. *et al.* Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Brief Bioinform* 2024;**25**:bbae052. https://doi.org/10.1093/bib/bbae052.
30. Hu J, Li X, Coleman K. *et al.* SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**:1342–51. https://doi.org/10.1038/s41592-021-01255-8.
31. Maynard KR, Collado-Torres L, Weber LM. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**:425–36. https://doi.org/10.1038/s41593-020-00787-0.

32. He K, Zhang X, Ren S. *et al.* Deep residual learning for image recognition. In: Anzarouth R (ed). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–8. IEEE.
33. Vincent DB, Jean-Loup G, Renaud L. *et al.* Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008;**2008**:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
34. Pham D, Tan X, Balderson B. *et al.* Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. *Nat Commun* 2023;**14**:7739. https://doi.org/10.1038/s41467-023-43120-6.
35. Cheng A, Hu G, Li WV. Benchmarking cell-type clustering methods for spatially resolved transcriptomics data. *Brief Bioinform* 2023;**24**:bbac475. https://doi.org/10.1093/bib/bbac475.
36. Yuan Z, Zhao F, Lin S. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nat Methods* 2024;**21**:712–22. https://doi.org/10.1038/s41592-024-02215-8.
37. Chen T, Kornblith S, Norouzi M. *et al.* A simple framework for contrastive learning of visual representations. In: Hal D, III, Aarti S (ed). *International conference on machine learning*, Online, 2020, pp. 1597–607. PMLR.
38. Stuart T, Butler A, Hoffman P. *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
39. Scrucca L, Fop M, Murphy TB. *et al.* mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* 2016;**8**:289–317. https://doi.org/10.32614/RJ-2016-021.
40. Hu J, Coleman K, Zhang D. *et al.* Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst* 2023;**14**:404–417.e4. https://doi.org/10.1016/j.cels.2023.03.008.
41. Cutler A, Breiman L. Archetypal analysis. *Dent Tech* 1994;**36**:338–47. https://doi.org/10.1080/00401706.1994.10485840.
42. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol* 2022;**40**:1349–59. https://doi.org/10.1038/s41587-022-01273-7.
43. Zhang H, Hunter MV, Chou J. *et al.* BayesTME: an end-to-end method for multiscale spatial transcriptional profiling of the tissue microenvironment. *Cell Syst* 2023;**14**:605–619.e7. https://doi.org/10.1016/j.cels.2023.06.003.
44. Lopez R, Li B, Keren-Shaul H. *et al.* DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol* 2022;**40**:1360–9. https://doi.org/10.1038/s41587-022-01272-8.
45. Kleshchevnikov V, Shmatko A, Dann E. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:661–71. https://doi.org/10.1038/s41587-021-01139-4.
46. Biancalani T, Scalia G, Buffoni L. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat Methods* 2021;**18**:1352–62. https://doi.org/10.1038/s41592-021-01264-7.
47. Chu T, Wang Z, Pe'er D. *et al.* Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Can* 2022;**3**:505–17. https://doi.org/10.1038/s43018-022-00356-3.
48. Song Q, Su J. DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform* 2021;**22**:bbaa414. https://doi.org/10.1093/bib/bbaa414.
49. Li H, Li H, Zhou J. *et al.* SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics* 2022;**38**:4878–84. https://doi.org/10.1093/bioinformatics/btac605.
50. Zhang T, Zhang Z, Li L. *et al.* GTAD: a graph-based approach for cell spatial composition inference from integrated scRNA-seq and ST-seq data. *Brief Bioinform* 2024;**25**:bbad469. https://doi.org/10.1093/bib/bbad469.
51. Bae S, Na KJ, Koh J. *et al.* CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data. *Nucleic Acids Res* 2022;**50**:e57–7. https://doi.org/10.1093/nar/gkac084.
52. Tzeng E, Hoffman J, Saenko K. *et al.* Adversarial discriminative domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 7167–76. Computer Vision Foundation.
53. Long Y, Ang KS, Li M. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun* 2023;**14**:1155. https://doi.org/10.1038/s41467-023-36796-3.
54. Lopez R, Nazaret A, Langevin M. *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv preprint arXiv:1905.02269 2019.
55. Lopez R, Regier J, Cole MB. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2.
56. Femino AM, Fay FS, Fogarty K. *et al.* Visualization of single RNA transcripts in situ. *Science* 1998;**280**:585–90. https://doi.org/10.1126/science.280.5363.585.
57. Shengquan C, Boheng Z, Xiaoyang C. *et al.* stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* 2021;**37**:i299–307. https://doi.org/10.1093/bioinformatics/btab298.
58. Li B, Zhang W, Guo C. *et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* 2022;**19**:662–70. https://doi.org/10.1038/s41592-022-01480-9.
59. Avşar G, Pir P. A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data. *Mol Omics* 2023;**19**:162–73. https://doi.org/10.1039/D2MO00266C.
60. Haviv D, Remšík J, Gatie M. *et al.* The covariance environment defines cellular niches for spatial inference. *Nat Biotechnol* 2024;1–12. https://doi.org/10.1038/s41587-024-02193-4.
61. Hao M, Luo E, Chen Y. *et al.* STEM enables mapping of single-cell and spatial transcriptomics data with transfer learning. *Commun Biol* 2024;**7**:56. https://doi.org/10.1038/s42003-023-05640-1.
62. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*. *PMLR*, 2017, **70**:3319–28.
63. Wei R, He S, Bai S. *et al.* Spatial charting of single-cell transcriptomes in tissues. *Nat Biotechnol* 2022;**40**:1190–9. https://doi.org/10.1038/s41587-022-01233-1.
64. Qian J, Liao J, Liu Z. *et al.* Reconstruction of the cell pseudo-space from single-cell RNA sequencing data with scSpace. *Nat Commun* 2023;**14**:2484. https://doi.org/10.1038/s41467-023-38121-4.
65. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat Commun* 2020;**11**:2084. https://doi.org/10.1038/s41467-020-15968-5.

66. Uhler C, Shivashankar G. Regulation of genome organization and gene expression by nuclear mechanotransduction. *Nat Rev Mol Cell Biol* 2017;**18**:717–27. https://doi.org/10.1038/nrm.2017.101.
67. Zhang X, Wang X, Shivashankar G. *et al.* Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. *Nat Commun* 2022;**13**:7480. https://doi.org/10.1038/s41467-022-35233-1.
68. Long Y, Ang KS, Sethi R. *et al.* Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nat Methods* 2024;**21**:1658–67. https://doi.org/10.1038/s41592-024-02316-4.
69. Huang X, Ma Z, Meng D. *et al.* PRAGA: prototype-aware graph adaptive aggregation for spatial multi-modal omics analysis. arXiv preprint arXiv:2409.12728 2024.
70. Zeira R, Land M, Strzalkowski A. *et al.* Alignment and integration of spatial transcriptomics data. *Nat Methods* 2022;**19**:567–75. https://doi.org/10.1038/s41592-022-01459-6.
71. Liu T, Fang ZY, Zhang Z. *et al.* A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics. *Comput Struct Biotechnol J* 2024;**23**:106–28. https://doi.org/10.1016/j.csbj.2023.11.055.
72. Bressan D, Battistoni G, Hannon GJ. The dawn of spatial omics. *Science* 2023;**381**:eabq4964. https://doi.org/10.1126/science.abq4964.
73. Ma Y, Zhou X. Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. *Nat Methods* 2024;**21**:1231–44. https://doi.org/10.1038/s41592-024-02284-9.
74. Athaya T, Ripan RC, Li X. *et al.* Multimodal deep learning approaches for single-cell multi-omics data integration. *Brief Bioinform* 2023;**24**:bbad313. https://doi.org/10.1093/bib/bbad313.
75. Yan L, Sun X. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. *Bioinformatics* 2023;**39**:btac805. https://doi.org/10.1093/bioinformatics/btac805.
76. Li H, Zhou J, Li Z. *et al.* A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat Commun* 2023;**14**:1548. https://doi.org/10.1038/s41467-023-37168-7.
77. Sang-Aram C, Browaeys R, Seurinck R. *et al.* Spotless, a reproducible pipeline for benchmarking cell type deconvolution in spatial transcriptomics. *elife* 2024;**12**:RP88431. https://doi.org/10.7554/eLife.88431.

Briefings in Bioinformatics, 2025, **26(1)**, bbae719
https://doi.org/10.1093/bib/bbae719
Review