Genome Biology

METHOD

Open Access

SMOPCA: spatially aware dimension reduction integrating multi-omics improves the efficiency of spatial domain detection

Mo Chen[1,2†], Ruihua Cheng[3†], Jianuo He[1,2], Jun Chen[4*] and Jie Zhang[1,2*]

†Mo Chen and Ruihua Cheng contributed equally to this work.

*Correspondence: Chen.Jun2@mayo.edu; zhangj_ai@nju.edu.cn

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu, China
[2] School of Artificial Intelligence, Nanjing University, Nanjing, Jiangsu, China
[3] Big Data Statistics Research Center, Tianjin University of Finance and Economics, Tianjin, China
[4] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

Abstract

Technological advances have enabled us to profile multiple omics layers with spatial information, significantly enhancing spatial domain detection and advancing a variety of biomedical research fields. Despite these advancements, there is a notable lack of effective methods for modeling spatial multi-omics data. We introduce SMOPCA, a Spatial Multi-Omics Principal Component Analysis method designed to perform joint dimension reduction on multimodal data while preserving spatial dependencies. Extensive experiments reveal that SMOPCA outperforms existing single-modal and multimodal dimension reduction and clustering methods, across both single-cell and spatial multi-omics datasets derived from diverse technologies and tissue structures.

Background

Humans and many other eukaryota are made from billions of cells, with a variety of cell types, functional states [1], and cellular activities [2–4]. Cellular phenotype and functional states are intrinsically regulated at multiple "omics" layers, involving the genome, epigenome, transcriptome, proteome, and metabolome [5]. In addition, the microenvironment [1] and neighboring cells could also modulate functional states of a cell through cell interactions [3, 6], cell signaling, and other microenvironmental factors.

The rapid development in multimodal sequencing technologies [7] enables us to simultaneously profile different omics layers [8], which provides a complete representation of cellular identity [8], allows the detailed classification of cell types, subtypes and functional states, and facilitates deep biological mechanism investigation in health and disease [9–19]. For example, CITE-seq [20] (cellular indexing of transcriptomes and epitopes by sequencing) and REAP-seq [21] (RNA expression and protein sequencing) have been developed to simultaneously quantify cell surface protein and transcriptomic data within a single-cell readout. Furthermore, emerging single-cell multi-omics technologies, including SNARE-seq [22] (single-nucleus chromatin accessibility and mRNA

expression sequencing) and SMAGE-seq [23] (10X Single-Cell Multiome ATAC + Gene Expression), aim to jointly capture chromatin accessibility and gene expression within individual cells. The biological information provided by each modality generally has its own strength and weakness, and different modalities from the multi-omics dataset are complementary [21, 24–26] in cell type identification.

Recent advances in spatial multi-omics technologies [5, 27, 28] further allow us to profile different aspects of a cell, while retaining spatial information, and are expected to improve spatial domain detection and foster significant progress in a variety of biomedical research fields. In 2020, Liu et al. presented DBiT-seq [29] (deterministic barcoding in tissue for spatial omics sequencing) for co-mapping of mRNAs and proteins in a formaldehyde-fixed tissue slide via next-generation sequencing (NGS). Subsequently, Liu et al. extended co-indexing of transcriptomes and epitopes (CITE) to the spatial dimension (spatial-CITE-seq [27]) and profiled hundreds of proteins and whole transcriptome in human tissues, revealing spatially distinct germinal center reactions in tonsil and early immune activation in skin at the Coronavirus Disease 2019 mRNA vaccine injection site. Jiang et al. in 2023 introduced the microfluidic indexing-based spatial assay for transposase-accessible chromatin and RNA-sequencing (MISAR-seq [28]), offering a method for the spatially resolved joint profiling of chromatin accessibility and gene expression. Through the application of MISAR-seq to the developing mouse brain, they conducted a comprehensive study on tissue organization and spatiotemporal regulatory dynamics during mouse brain development. Despite technical variations, capturing the inherent spatial dependencies in spatial domain detection significantly extends our ability to annotate cell types, improves downstream analysis (e.g., visualization, differential expression analysis, gene set enrichment analysis, trajectory analysis, and so on), and facilitates the investigation of biological mechanisms [30–33].

To fully exploit the complementary and mutual enhancing nature of single-cell multi-omics data, joint dimension reduction and clustering analysis algorithms have been proposed [25, 34, 35]. For example, BREM-SC [24], CiteFuse [36], Seurat V4 [37], Specter [38], TotalVI [34], and scMDC [25] (Additional file 5: Supplementary Note 1) that combine multiple cellular views in an unbiased manner have been proposed for the clustering analysis of the single-cell multi-omics data in the past years. Nevertheless, to our best knowledge, none of the aforementioned methods have the capability to model spatial dependencies in the data. Therefore, current computational methods for single-cell multi-omics data may not be optimal for analyzing spatial multi-omics datasets, as they assume independence among cells or spots and fail to integrate crucial spatial information.

Recently developed methods for analyzing spatially resolved transcriptomics (SRT) data may not be directly applicable to spatial multi-omics data. For example, SpatialPCA [33], which is based on the generalized probabilistic principal component analysis [39] (GPPCA), is a state-of-the-art method to model the spatial correlation structure across tissue locations and learn better low-dimensional representations for the spatial transcriptomics data. However, SpatialPCA cannot be directly applied to spatial multi-omics data. Naïve concatenation of different omics data before applying SpatialPCA ignores the potential varying informativeness of each omics datatype. In contrast, SpaVAE [40] is a spatial variational autoencoder model developed by Tian

et al. in 2024 mainly for modeling the SRT data. Similar to scVI [41] and TotalVI [34], SpaVAE can be extended to characterize spatial multi-omics data by adding additional network branches and modifying loss functions. In 2024, Long et al. proposed SpatialGlue [42], a graph neural network model with a dual-attention mechanism designed to detect spatial domains from spatial multi-omics data. Although SpaVAE and SpatialGlue differ in the implementation, they share the same limitation that they could only process at most two modalities and have to re-design the network architecture completely if three or more modalities are available. To alleviate the restriction, Long et al. provided "SpatialGlue_3M" (https://github.com/JinmiaoChenLab/SpatialGlue_3M), an extended version of SpatialGlue, tailored for integrating data with exactly 3 modalities. However, even with this extension, SpatialGlue still could not handle dataset with ≥ 4 modalities. Unlike SpaVAE and SpatialGlue, MEFISTO [43] is a factor analysis-based method developed for modeling spatial multimodal data. MEFISTO can natively handle data with three or more modalities. However, existing studies [42] have shown that MEFISTO's performance is less competitive compared to SpatialGlue. In addition, training a MEFISTO model takes significantly longer (Additional file 3: Table S2). Methods that can effectively and simultaneously model various data modalities and spatial information are scarce.

To fill the methodological gap, we introduce a novel spatial multi-omics principal component analysis method, named SMOPCA, to optimize the entire computational pipeline for dimension reduction, clustering analysis and spatial domain detection, and deliver more accurate results for downstream analytical tasks in spatial multi-omics studies. To our best knowledge, this is the first principal component analysis model that specifically designed for modeling spatial dependencies in the spatial multi-omics data. SMOPCA simultaneously models different data modalities and spatial information and infers a joint low-dimensional representation over multiple omics data types. The latent factors of SMOPCA encapsulate variations across data modalities, facilitating the discernment of biological signals. We theoretically demonstrate that the latent factors learned by SMOPCA, which integrate information from multi-modal data, are valid and more stable than performing dimension reduction on each modality separately. The learned latent representations could maintain the spatial correlation structure across tissue locations and, therefore, preserve the neighboring similarity of the original spot in the low-dimensional manifold. The latent vectors obtained from SMOPCA can be directly paired with K-means clustering to improve spatial domain detection and enhance downstream analysis for spatial multi-omics data, as verified through experiments conducted on recently published spatial multi-omics datasets. In addition, SMOPCA is readily applicable to model single-cell multi-omics data with the pseudo-spatial coordinates generated from the nonlinear method UMAP [44]. It could effectively borrow the strength from another dimension reduction method to preserve local clustering structures by naturally employing a prior multivariate normal distribution defined on the pseudo-spatial coordinates. When applied to single-cell multi-omics datasets obtained from diverse technologies and tissue structures, SMOPCA consistently delivers superior or, at least, comparable results compared to existing methods, including the best deep learning-based approach. Like PCA, SMOPCA demonstrates enhanced robustness and stability,

when applied to different datasets, and we hope it could contribute to the in-depth understanding of the mechanisms in biological and biomedical studies.

Results

An overview of SMOPCA

Figure 1 displays the schematic diagram of the proposed dependency-aware dimension reduction method SMOPCA. As shown in Fig. 1, SMOPCA takes multiple modalities $Y = \{Y_1, \ldots, Y_K\}$ and location information $S = \{s_1, \ldots s_n\}$ as inputs, and learns joint latent factors Z through factor analysis models. To integrate spatial location information, we assume each latent factor Z_l adheres to a multivariate normal (MVN) prior distribution with expected values $\mathbf{0}$ and covariance matrix $\Sigma_l \in \mathbb{R}^{n \times n}$ calculated based on the location information S to explicitly capture the spatial dependencies in the latent space across cells or spots. Technical details and rationale of SMOPCA are given in " Methods" and Additional file 5: Supplementary Note 3. The posterior distribution of Z encapsulates valuable correlation information and could be utilized to explain the variance of the data. The outputs of SMOPCA can be seamlessly integrated with existing analytical tools to facilitate and improve downstream analyses, including dimension reduction, visualization, clustering analysis, differential expression analysis, and many other analytical tasks, in spatial multi-omics studies.

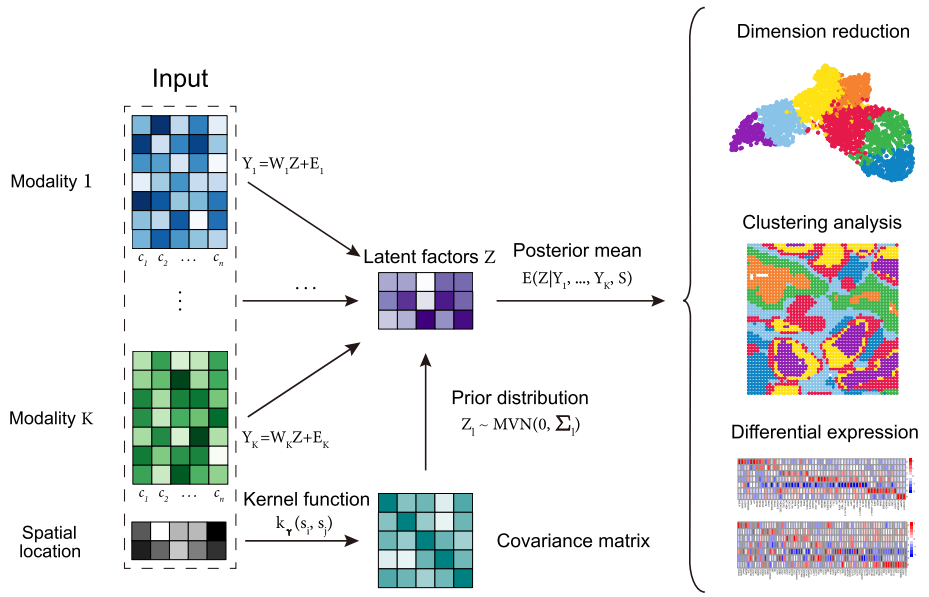

Fig. 1 The architecture of SMOPCA model. SMOPCA is a dependency-aware dimension reduction method that accepts multiple modalities $Y = \{Y_1, \ldots, Y_K\}$ and location information $S = \{s_1, \ldots s_n\}$ as inputs. It models each modality Y_i through a factor analysis model with joint latent factors Z. We assume that each latent factor Z_l follows a multivariate normal (MVN) prior distribution with covariance matrix calculated based on the location information S to explicitly capture the spatial dependencies in the latent factors Z across cells or spots $\{c_1, c_2, \ldots, c_n\}$. As a result, the obtained low-dimensional components Z from SMOPCA encapsulate valuable spatial correlation information and can be seamlessly integrated with existing analytical tools to facilitate and enhance downstream analyses (e.g., dimension reduction, visualization, clustering analysis, differential expression analysis, and many other analytical tasks) for spatial multi-omics studies

Simulation studies

Simulation study I

To numerically evaluate the effectiveness and robustness of SMOPCA, we conducted simulation experiments using real single-cell multi-omics datasets with simulated cell locations. Here, we propose to evaluate the performance of SMOPCA on five CITE-seq datasets, including the Peripheral Blood Mononuclear Cells (PBMC) dataset [38] and a series of Mouse Spleen Lymph Node datasets (SLN111D1, SLN111D2, SLN208D1, and SLN208D2) [34]. Additionally, we incorporate two Single-cell Multiome ATAC and Gene Expression (SMAGE-seq) datasets [22, 23] derived from human peripheral blood mononuclear cells, denoted as PBMC3K and PBMC10K, comprising approximately 3000 and 10,000 cells, respectively. Detailed information on the experimental datasets is provided in section " Public real datasets" and Additional file 2: Table S1. To make a comprehensive comparison, we evaluate our model against widely used cell clustering methods developed for CITE-seq data (i.e., BREM-SC, CiteFuse, SC3, scMDC, Seurat, TotalVI, Tscan, SpatialPCA, and PCA + K-means) and additional methods designed specifically for SMAGE-seq data (i.e., chromVAR [45], cisTopic [46], LSA [47], PeakVI [48], SCALE [49], scMDC, Seurat, SpatialPCA, and PCA + K-means), as detailed in section " Competing methods." Notably, the competing methods include approaches specifically designed for multimodal data clustering analysis (e.g., scMDC, BREM-SC, CiteFuse, and Seurat) and models developed for learning low-dimensional embeddings for either single-modal or multimodal data (e.g., TotalVI). Furthermore, the evaluation incorporates two prominent clustering tools, SC3 and Tscan, specifically designed for single-cell data analysis, along with a widely used general clustering framework PCA + K-means, serving as baseline methods. It is worth noting that SpatialPCA, a spatially aware dimension reduction method proven to be effective in spatial domain detection for spatial transcriptomics, is also utilized as a robust baseline method for thorough comparisons across all analyses.

Due to the absence of cell location information in the single-cell multi-omics datasets (e.g., PBMC, SLN111D1, SLN111D2, SLN208D1, SLN208D2, PBMC3K, and PBMC10K), we try to simulate the spatial location for each cell. Inspired by the six-layered human dorsolateral prefrontal cortex (DLPFC) datasets [50], we assume that cells from the same cluster are proximate, and distinct cell types exhibit layer-wise patterns. For a single-cell multi-omics dataset with n cells, we map the cells onto a simulated 2D grid (as detailed in " Data simulation" section). Specifically, we shuffle the cells by randomly ordering the cell types and, within each cell type, randomly permutating the cells. We allocate the cells to the grid locations, proceeding row by row and column by column. Additional file 1: Fig. S1-S7 ("Simulated Coordinates") display an example of simulated cell locations for different datasets. As our simulation studies are based on real multi-omics data with zero inflation and over-dispersion (e.g., gene and protein expressions from the CITE-seq data, and gene and peak measurements from the SMAGE-seq data), the simulated datasets are also zero-inflated and over-dispersed. Without loss of generality, the simulated coordinates have the flexibility to undergo transformations or rotations, yielding a distinct set of locations while preserving the cell-to-cell distances. To ensure a thorough comparison, for each real

single-cell multi-omics dataset, we randomly generated 10 simulated locations resulting in 10 simulated datasets and subsequently assessed the clustering performance of different methods for each simulated dataset.

We visualized the simulated spatial coordinates and the low-dimensional representations learned by SpatialPCA and SMOPCA for a single simulation in Additional file 1: Fig. S1-S7 ("Simulated Coordinates," "SpatialPCA-Sim," and "SMOPCA-Sim"). As shown in Additional file 1: Fig. S1-S7 ("Simulated Coordinates," "SpatialPCA-Sim," and "SMOPCA-Sim"), the latent vectors learned by SMOPCA exhibit improved capacity to distinguish between different cell types compared to SpatialPCA, where cells from different types are mixed together.

Figure 2a,b illustrate the clustering performance of SMOPCA and the competing methods on five CITE-seq and two SMAGE-seq datasets with simulated spatial coordinates, respectively. The detailed results are shown in Additional file 1: Fig. S8-S9. The evaluation is based on three distinct metrics: Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) (Please refer to the " Evaluation metrics" section for details). These metrics act as quantitative measurements for evaluating the quality of clustering results obtained from various methods. For each dataset, we conducted the experiments 10 times with different simulated coordinates. As shown in Fig. 2a,b and Additional file 1: Fig. S8-S9, both SMOPCA and SpatialPCA exhibit superior performance compared to PCA +K-means, highlighting the efficacy of simulated spatial coordinates in helping to capture the inherent structure of the data. It also demonstrates from another perspective that the capability of SMOPCA and SpatialPCA in modeling location information or cell-to-cell dependency is instrumental in achieving more accurate clustering results. Moreover, SMOPCA outperforms SpatialPCA, which takes concatenated data from different modalities as inputs, across all three metrics (AMI, NMI, and ARI).

In addition, SMOPCA demonstrates much better results compared to all other methods, including scMDC, which is recognized as a leading method for clustering analysis in single-cell multi-omics data. This observation holds true for both simulated CITE-seq and SMAGE-seq datasets across all three metrics (i.e., AMI, NMI, and ARI). We further present the results in a unified manner in Additional file 1: Fig. S10-S12,S14-S16. It is evident that when two metrics are considered simultaneously, SMOPCA consistently maintains its superior position.

We rank all methods according to their performance for each dataset. The ranking is determined based on the median of the metric values across multiple runs for each dataset, thereby incorporating the consideration of performance stability. Additional file 1: Fig. S13 and Additional file 1: Fig. S17 show the averaged rank and standard error of each method across different experimental datasets. It is evident that, when examining the averaged rank, SMOPCA consistently outperforms other methods.

We also compared to recently developed methods, SpaVAE and SpatialGlue. As shown in Additional file 1: Fig. S18-S19, SMOPCA outperforms SpaVAE and SpatialGlue on the real single-cell multi-omics datasets with simulated spatial coordinates. In addition, based on the simulated datasets, we demonstrate that SMOPCA, which integrates multiple modalities, outperforms methods that operate on each modality individually (Additional file 1: Fig. S20-S21). To study the robustness of SMOPCA, we

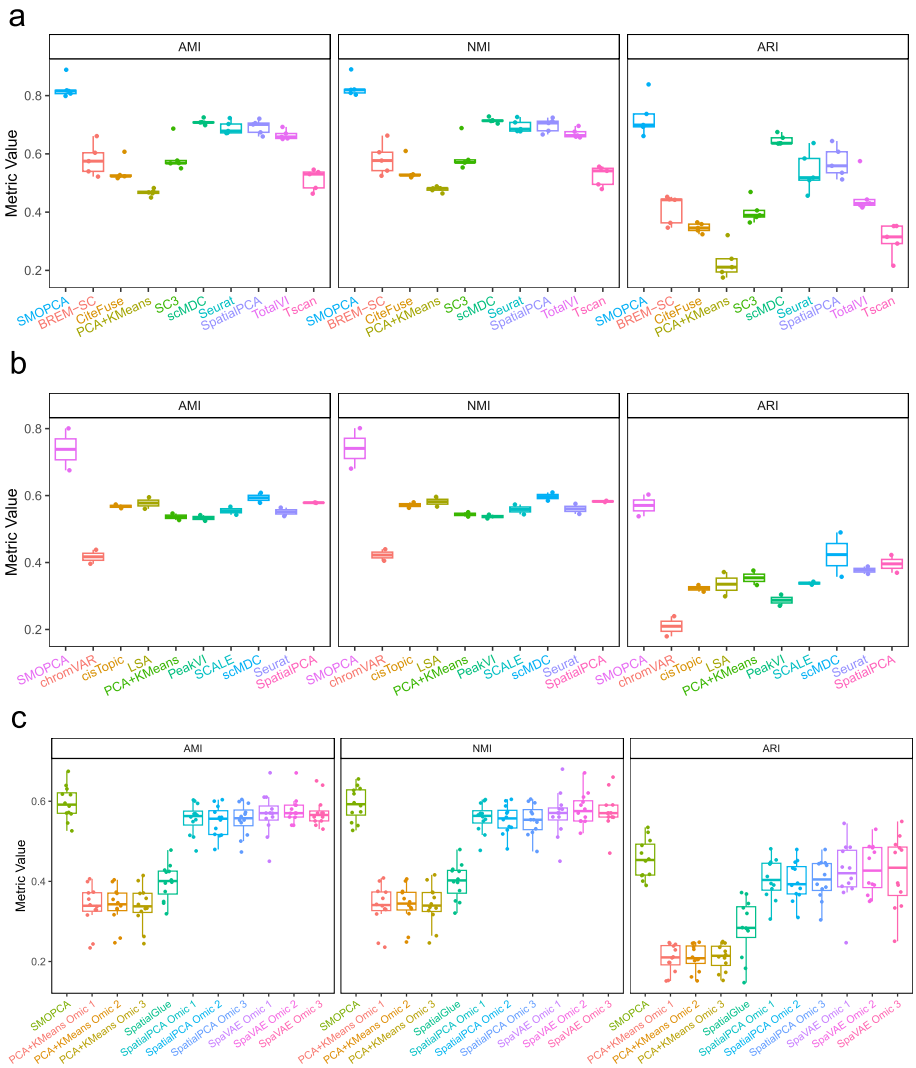

Fig. 2 Clustering performance of SMOPCA and competing methods on simulated data. **a** Clustering results of SMOPCA and competing methods on 5 CITE-seq datasets with simulated spatial coordinates. **b** Clustering results of SMOPCA and competing methods on 2 SMAGE-seq datasets with simulated spatial coordinates. **c** Clustering results of SMOPCA and competing methods on 12 simulated 3-omics samples. SpatialGlue uses the implementation "SpatialGlue_3M" (https://github.com/JinmiaoChenLab/SpatialGlue_3M) for modeling the 3 omics data. Each dot represents an experiment on an individual dataset. All boxplots are standard boxplots, which display the distribution of data by presenting the inner fence (the whisker, taken to 1.5 × the interquartile range, or IQR, from the quartile), first quartile, median, third quartile, and outliers

performed sensitivity analysis by examining its performance with different kernels, kernel parameters ($\nu = 0.5$, $\nu = 1.5$, and $\nu = 2.5$), number of components, data normalization methods (i.e., LogNormalize, SCTransform, and VST), and gene selection methods (i.e., SVG [51] and HVG). Results in Additional file 1: Fig. S22-S31 show that SMOPCA is quite robust to different parameter settings and preprocessing procedures. Taken together, these results highlight the effectiveness and robustness of the proposed method in exploiting spatial information to achieve improved dimension reduction and clustering performance.

Simulation study II

Clustering analyses in existing single-cell multi-omics and spatial multi-omics research (e.g., TotalVI, MOFA [35], scMDC, MEFISTO, SpaVAE, SpatialGlue, and many others) have mainly focused on two omics modalities. To demonstrate the capabilities of SMOPCA in analyzing data with more modalities, we simulated datasets with three modalities. Following SpaVAE [40], we employed SRTsim [52] to simulate three omics modalities based on the LIBD human dorsolateral prefrontal cortex (DLPFC) dataset [50]. The DLPFC data have twelve tissue sections, spanning six neuronal layers and the white matter from three human brains. For each tissue section, we simulated three count matrices for the three modalities based on the SRT data (detailed in " Data simulation" section). We next applied SMOPCA to the twelve simulated sections. The results in Fig. 2c and Additional file 1: Fig. S46-S48 demonstrate that SMOPCA outperforms PCA + K-means, SpatialPCA, SpaVAE, and SpatialGlue (SpatialGlue_3M), and the performance differences between SMOPCA and other methods are statistically significant under the paired *t*-test (Additional file 4: Table S3). The performance of SMOPCA with different kernels and gene selection methods (i.e., SVG [51] and HVG) are provided in Additional file 1: Fig. S49-S50. These findings highlight the distinct advantage of our model in analyzing multi-omics data.

Application to spatial multi-omics data

Application to spatial CITE-seq data

Recent advancements in sequencing technologies have enabled simultaneous profiling of gene expression and surface protein abundance at the level of individual cells or spots within spatial genomics data [27, 29]. To demonstrate the capability in modeling real spatial multi-omics data, we applied SMOPCA to analyze a real spatial-CITE-seq data of human tonsil tissue [27] (Additional file 2: Table S1). Following existing studies [51], we employed SPARK [53] to identify spatially variable genes (SVGs), benefiting from its higher statistical power. We selected up to 3000 significant SVGs, with a false discovery rate (FDR) ≤ 0.05, as input to our model. Note that the filtering process was limited to genes (mRNA data), and all proteins were included.

We compared our model with state-of-the-art methods, including scMDC, Seurat, TotalVI, and many others, commonly employed for integrating CITE-seq data. Similar to SMOPCA, both SpatialPCA and TotalVI are generative models. In addition, the weighted-nearest neighbor (WNN) functionality in Seurat package, which integrates gene and protein modalities, was used as a competing method. We also compared SMOPCA with newly developed spatial multi-omics methods, including MEFISTO, SpatialGlue, and SpaVAE. Following Tian et al. [40], we applied the *K*-means clustering to the latent embeddings learned by dimension reduction methods, namely SMOPCA, SpaVAE, SpatialPCA, PCA, and TotalVI, to obtain 7 clusters.

In Fig. 3a,b and Additional file 1: Fig. S51, we show the overall clustering results and the learned latent vectors from competing methods. The clustering pattern from SMOPCA is spatially smoother and is more concordant with the tissue image (Fig. 3a,b). Evaluating the spatial smoothness using Moran's I and LISI scores [54] reveals that the clustering results of scMDC, Seurat, TotalVI, and PCA + K-means are spatially noisier (i.e.,

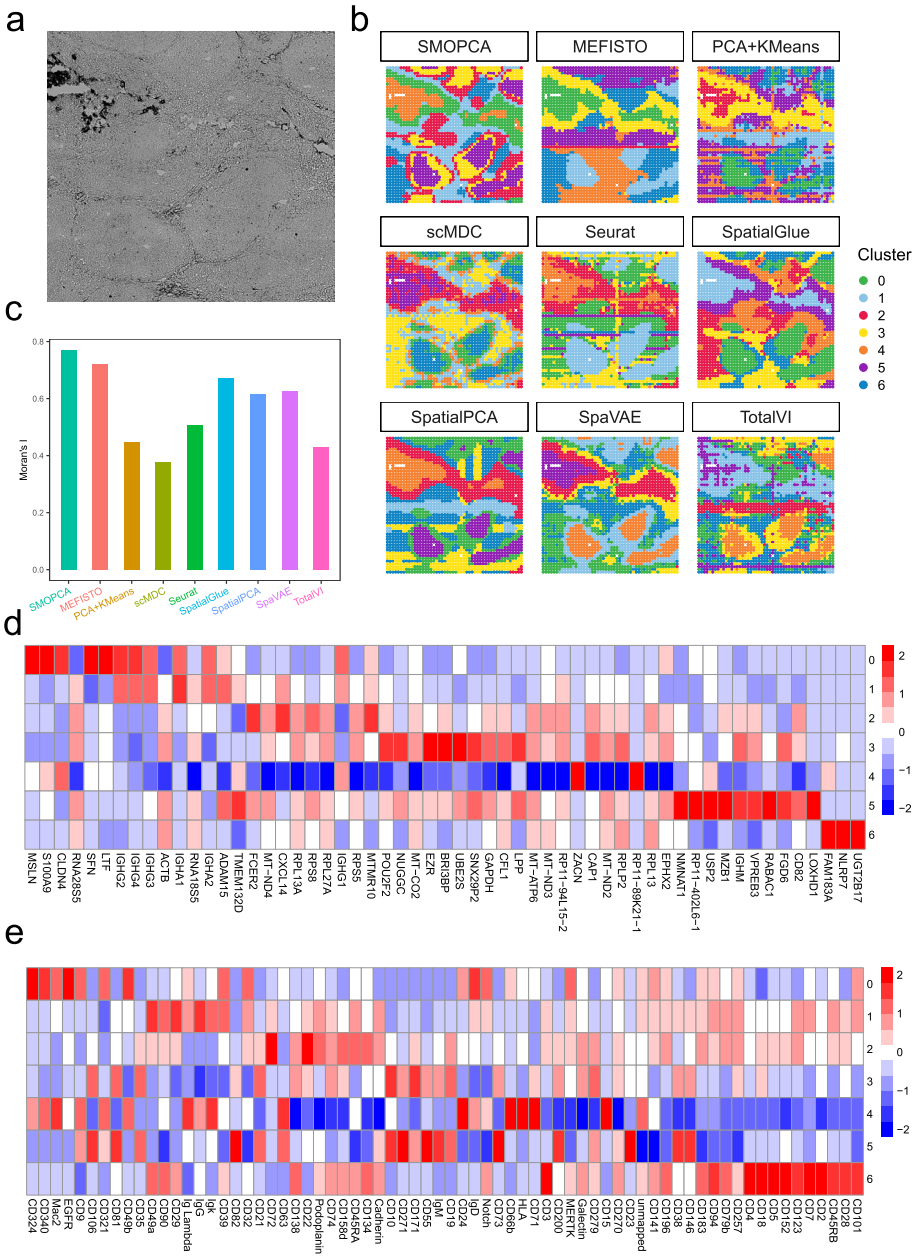

Fig. 3 Application of SMOPCA to a real spatial-CITE-seq data. **a** Tissue image of the analyzed human tonsil sample. **b** Clustering results of SMOPCA and different competing methods, with colors denoting the predicted cluster labels. **c** Moran's *I* score of different methods. **d,e** Top genes (**c**) and top proteins (**d**) identified by the differential expression analysis conducted on the predicted labels of SMOPCA. Heatmaps visually display the relative averaged expression levels across predicted clusters

lower Moran's I score and higher LISI score) than SMOPCA and SpatialPCA (Fig. 3a–c and Additional file 1: Fig. S52), highlighting the importance of using spatial information. Notably, SMOPCA achieves the highest Moran's I score and the lowest LISI score compared to SpatialPCA, MEFISTO, SpatialGlue, and SpaVAE, indicating reduced mixing of different cell types and increased homogeneity within clusters. As expected, SMOPCA also outperforms methods applied to individual modalities (Additional file 1: Fig. S54).

The tissue image and spatial clusters derived from SMOPCA exhibit a robust correlation between anatomical features and tissue/cell types (Fig. 3a,b). Following Liu et al. [27], we could approximately annotate the clusters obtained by SMOPCA as follows: Cluster 0 corresponds to the crypt epithelia; Cluster 6 indicates specific T cell zones; Clusters 5 and 3 represent the germinal center (GC) dark and light zones; Clusters 1 and 2 are localized in extrafollicular regions; and Cluster 4 contains peripheral blood cells in vasculature. We performed differential expression (DE) analyses for both genes and proteins to identify the top genes and proteins within each cluster reported by SMOPCA (Fig. 3d,e). Specifically, significant genes and proteins (adjusted p-value <0.05) were detected by Seurat's "*FindAllMarkers*" function with the cluster labels reported by SMOPCA. The top 10 most significant genes and proteins for each cluster were visualized in Fig. 3d,e. We observed that T cell markers, such as CD3, CD4, and CD49a [55], were upregulated in cluster 6, which indicated specific T cell zones [27]. CD21 [56] and CD23, found on mature B cells [27], along with IgM, whose expression is restricted to GC B cells [27], exhibit upregulation in cluster 3 and 5. CD90 (Thy- 1) is associated with a wide range of cell types (enriched in clusters 1 and 6) but completely absent in GCs (clusters 3 and 5). CD32 is an Fc receptor that regulates B cell activation [57] and was found mainly outside GCs [27] (only upregulated in cluster 0, 1, and 2). Mac2/Galectin3 is highly enriched in the crypt zone [27] (cluster 0). The identified differentially expressed protein markers, including markers associated with B cells and GC B cells (CD21, CD23, IgM, and IgD), markers enriched in the extracellular region (CD90 and Mac2), T cell markers (CD3, CD4, and CD45RA), and other related protein markers (CD32, CD9, and CD171), are visually presented in Additional file 1: Fig. S53. We conducted gene set enrichment analysis (GSEA), following the differential gene expression analysis. The significant pathways (adjusted p-value <0.05) for the clusters identified by SMOPCA on the real spatial-CITE-seq dataset are shown in Additional file 1: Fig. S55.

Application to Stereo-CITE-seq data

To demonstrate the broad applicability of SMOPCA across various technology platforms, we further applied it to analyze Stereo-CITE-seq [58] data. Liao et al. combined CITE-seq and Stereo-seq to develop the Stereo-CITE-seq [58] workflow, which could capture mRNA and protein expression with high spatial resolution, reproducibility, and accuracy. The Stereo-CITE-seq was used by Liao et al. to analyze a mouse thymus section (a small gland surrounded by a capsule of fibers and collagen) and obtain mRNA and protein measurements for each spot. Following Long et al. [42], we preprocessed the data and selected up to 2000 HVGs. The filtering process was applied only to mRNA data, while all proteins were included. Following SpatialGlue, we tested 9 methods, including MEFISTO, SpaVAE, SpatialGlue, scMDC, TotalVI, Seurat, SpatialPCA, PCA + K-means, and SMOPCA, to obtain 8 clusters for this dataset [42]. The learned latent vectors for different methods were visualized in Additional file 1: Fig. S56.

SMOPCA, scMDC, MEFISTO, SpaVAE, and SpatialGlue all detected similar patterns, indicating their abilities to reveal meaningful biological structure and provide useful low dimensional representation for each spot. The detected patterns/clusters are corroborated by protein marker expression, as shown in Additional file 1: Fig. S59. Following Long et al., we could approximately annotate some detected clusters as follows:

0—Medulla, 1—Middle cortex region, 3—Connective tissue capsule, 5—Outer cortex region, 6—Inner cortex region, 7—Connective tissue capsule and Subcapsular zone. Additional file 1: Fig. S57-S58 display the Moran's I and LISI scores for different methods. Overall, SMOPCA achieves the best scores, indicating its superior performance in analyzing this Stereo-CITE-seq [58] data. We further conducted differential expression analysis based on cluster labels obtained by SMOPCA and identified differentially expressed top genes (Fig. 4b) and top protein markers (Fig. 4c, Additional file 1: Fig. S59). Based on gene set enrichment analysis, the significant pathways (adjusted p-value < 0.05) for the clusters identified by SMOPCA are presented in Additional file 1: Fig. S61. We also analyzed the effectiveness of each modality in clustering analysis and demonstrated that methods integrating information from multiple modalities could deliver better results as shown in Additional file 1: Fig. S60.

Application to spatial ATAC and mRNA data

Microfluidic indexing-based spatial assay for transposase-accessible chromatin and RNA-sequencing (MISAR-seq), proposed by Jiang et al., is a new sequencing method

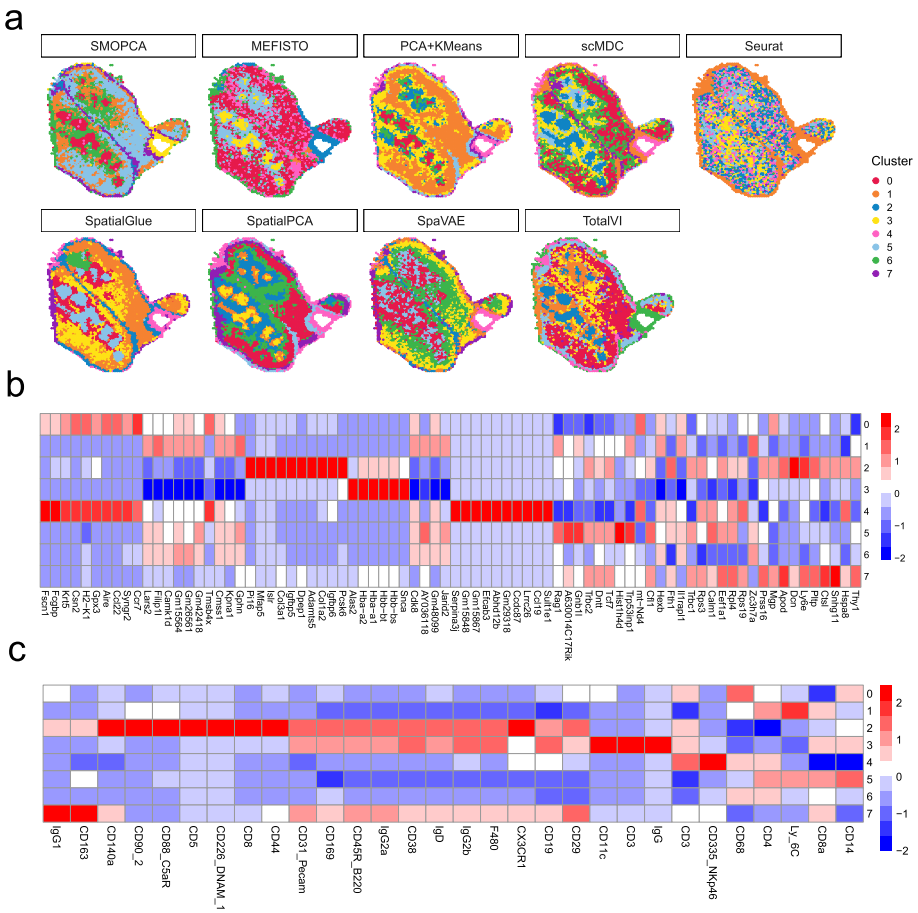

Fig. 4 Application of SMOPCA to a real Stereo-CITE-seq data. **a** Clustering results of SMOPCA and different competing methods, with colors denoting the predicted cluster labels. **b**,**c** Top genes (**b**) and top proteins (**c**) identified by the differential expression analysis conducted on the predicted labels of SMOPCA. Heatmaps visually display the relative averaged expression levels across predicted clusters

for spatially resolved joint profiling of chromatin accessibility and gene expression [28]. By applying MISAR-seq to the developing mouse brain, Jiang et al. [28] measured ATAC and mRNA counts in mouse embryonic brain across various regions (Additional file 1: Fig. S62), and studied tissue organization and spatiotemporal regulatory logics during mouse brain development. To demonstrate the performance of SMOPCA in analyzing real spatial ATAC-seq data, we analyzed a MISAR-seq dataset of mouse embryonic (E15.5) brain [28] (Additional file 2: Table S1).

Figure 5 displays the results of SMOPCA along with other competing methods. Figure 5a and Fig. 5c visualize the ground-truth labels and predicted clusters generated by various methods. SMOPCA and SpatialPCA, designed to capture spatial correlations among spots, yield notably smoother results (Additional file 1: Fig. S63-S64) compared to other competing methods. Additionally, SMOPCA consistently delivers the best results across all three metrics, including AMI, NMI, and ARI (Fig. 5b). We visualize the latent embeddings of various methods in Fig. 5d, with ground-truth labels indicated by different colors. Overall, we observe that the latent representation of SMOPCA effectively separates different labels, whereas forebrain, midbrain, and hindbrain tend to be entangled in the embeddings produced by other methods (e.g., chromVAR, cisTopic, LSA, PeakVI, SCALE, scMDC, and Seurat). Note that SMOPCA also outperforms methods applied to each individual modality (Additional file 1: Fig. S66-S67).

While SMOPCA outperforms other methods, there are still some mismatches between the ground-truth groups and the predicted clusters, particularly in the case of the forebrain region. Therefore, we apply the McFadden-adjusted pseudo R^2 value [59] to quantify the latent representation's capability to predict the true cell types. A higher pseudo R^2 value indicates a greater likelihood of accurately deriving true cell types from the latent representations. We presented the McFadden's pseudo R^2 value of all methods in Additional file 1: Fig. S65. As presented in Additional file 1: Fig. S65, SMOPCA stands out with the highest pseudo R^2 value, indicating its effectiveness in capturing and explaining the majority of the variance present in the MISAR-seq data. This finding emphasizes the superior capability of SMOPCA in extracting meaningful latent representations from the spatial bimodal ATAC-seq data.

We also applied our method to analyze a mouse brain dataset obtained through recently developed spatial ATAC–RNA-seq, a spatially resolved, genome-wide technique that co-maps the epigenome and transcriptome by simultaneous profiling of chromatin accessibility and mRNA expression. Following Zhang et al. [60] and Long et al. [42], we processed the data and clustered the spots into 18 clusters. As shown in Additional file 1: Fig. S68-S71, the results from SMOPCA aligned well with the annotated reference of the mouse brain coronal section from the Allen Mouse Brain Atlas. In addition, SMOPCA provided smoother results compared to all other competing methods, as evidenced by Moran's I and LISI scores.

Application to single-cell multi-omics data

Although the motivation of the development of SMOPCA is to utilize the spatial information to improve the efficiency and interpretability of dimension reduction and spatial domain detection based on multi-omics data, it is readily applicable to single-cell multi-omics data with the pseudo-spatial coordinates generated from

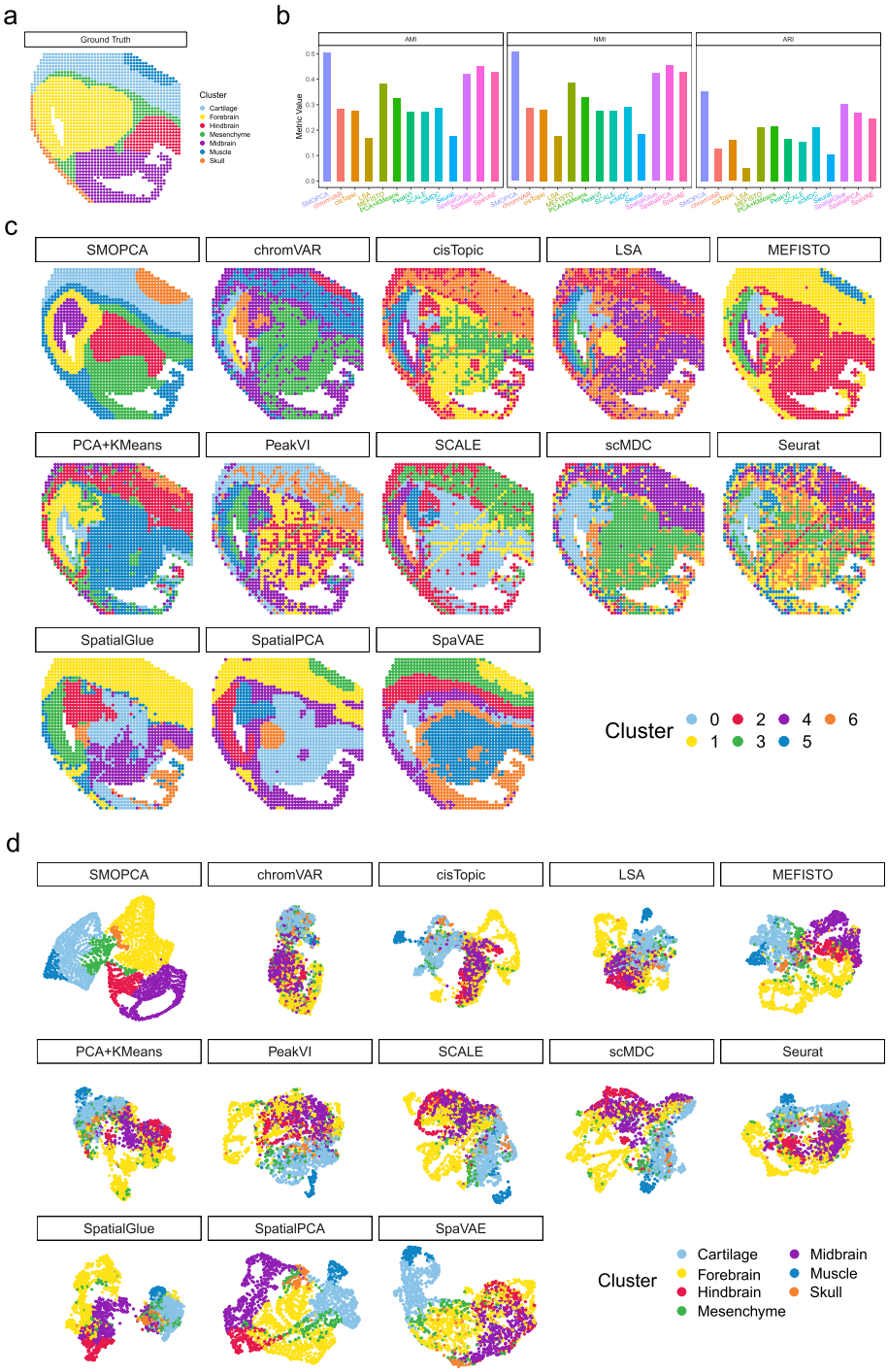

Fig. 5 Application of SMOPCA to a real MISAR-seq data. **a** Manually annotated labels of mouse embryonic (E15.5) brain tissues in the MISAR-seq dataset. **b** Clustering accuracy of SMOPCA and different competing methods, quantified by AMI, NMI, and ARI. **c** Clustering results of SMOPCA and different competing methods, with colors denoting the predicted cluster labels. **d** 2D UMAP visualization of latent representations learned by SMOPCA and different competing methods, with colors denoting manually annotated labels

other dimension reduction techniques. Thus, it effectively borrows the strength from another dimension reduction method. It is worth noting that existing methods, including scvis [61] and scDHMap [62], have been introduced to preserve local clustering structures and capture underlying low-dimensional embeddings in scRNA-seq data through the application of t-SNE [63] regularizations. In the following applications, we show that by using the UMAP [44] coordinates, we improve the efficiency of the cell clustering, compared to the version of SMOPCA without using any coordinate information.

In this section, we propose to generate the pseudo-spatial coordinates in a two-dimensional space from the data and then compute distance and covariance matrices. Specifically, the pseudo-spatial coordinates are achieved through the application of the UMAP [44] transformation to the top 2000 highly variable genes (CITE-seq data) or mapped genes (SMAGE-seq data) through Seurat package (as detailed in section "UMAP location generation for real single-cell multi-omics data"). The computed cell pseudo-spatial coordinates, along with gene and protein/ATAC measurements, serve as inputs for fitting a SMOPCA model. As illustrated in Additional file 1: Fig. S1-S7 ("UMAP Coordinates"), while there are some overlaps between different cell types, the calculated pseudo-coordinates consistently manifest a pattern where cells of the same type tend to exhibit smaller distances between them. In the subsequent paragraphs, we demonstrate that with the calculated spatial dependency information, the proposed SMOPCA can still effectively learn latent representations that preserve correlations and biological information for cells, resulting in improved clustering performance.

We evaluate the clustering performance of SMOPCA on the real CITE-seq and SMAGE-seq datasets. The pseudo-spatial coordinates and low-dimensional representations learned by SpatialPCA and SMOPCA are visualized in Additional file 1: Fig. S1-S7 ("UMAP Coordinates", "SpatialPCA", and "SMOPCA"). Additional file 1: Fig. S32-S33 and Additional file 1: Fig. S34-S36,S39-S41 illustrate the clustering performance of SMOPCA and the competing methods on the real CITE-seq and SMAGE-seq datasets with UMAP coordinates. Note that methods, such as scMDC, Seurat, and PCA + K-means, do not use any cell location information, and their results align with those presented in simulation studies. When compared with the state-of-the-art deep learning-based method scMDC, SMOPCA consistently yields superior or at least comparable results across all metrics and datasets, as shown in Additional file 1: Fig. S32-S33 and Additional file 1: Fig. S34-S36,S39-S41. Additionally, SMOPCA stands out by providing significantly better results than many other competing methods. The results, quantified by the averaged rank, are presented in Additional file 1: Fig. S37 and Additional file 1: Fig. S42.

We ran a special version of SMOPCA by explicitly setting the covariance matrix to a diagonal matrix. The results are shown in Additional file 1: Fig. S38 and Additional file 1: Fig. S43. It is interesting that, with diagonal covariance matrix, our method still delivers good results and actually outperforms many existing single-cell multi-omics methods in clustering analysis, which suggests that our method could be directly applied to single-cell multi-omics data even without assuming any dependency among cells. Finally, we conducted additional studies to demonstrate that SMOPCA, which models multiple modalities simultaneously, outperforms methods that operate on each modality individually (Additional file 1: Fig. S44 and Additional file 1: Fig. S45).

In summary, the proposed method can be applied to single-cell multi-omics datasets by directly computing a cell-to-cell distance matrix, constructed using domain knowledge, marker genes, or other relevant information, or by directly assigning a prior distribution with diagonal covariance matrix for the latent embeddings. Therefore, our model is flexible and capable of incorporating diverse information to model the single-cell multi-omics data.

Discussion

Our study presents SMOPCA, a novel dependency aware dimension reduction method that is tailored for spatial multi-omics data. SMOPCA explicitly integrates spatial location information into the learned latent factors, preserving neighborhood similarity from the original data onto the low-dimensional manifold. Therefore, the low-dimensional components derived from SMOPCA encapsulate valuable spatial correlation information, potentially enhancing the capabilities of existing tools and facilitating a range of downstream analyses in spatial multi-omics data analysis. In this paper, our focus lies in the joint analysis of mRNA alongside protein (spatial-CITE-seq and Stereo-CITE-seq) or ATAC data (MISAR-seq). It is worth noting that our method can be directly applied to analyze single-cell multi-omics data and is readily available to concurrently analyze more sources of data ($K > 2$). Our experimental results on real single-cell (CITE-seq and SMAGE-seq) and spatial multi-omics (spatial-CITE-seq, Stereo-CITE-seq and MISAR-seq) data demonstrate that the proposed dependency-aware dimension reduction approach can characterize different sources of data and model correlations and dependencies among neighboring cells/spots effectively and efficiently, and could deliver better or at least comparable results compared with existing methods in dimension reduction and clustering analysis.

The lengthscale/bandwidth parameter in the kernel function could control the smoothness of the detected spatial patterns. SMOPCA and SpatialPCA differ in the ways to choose the lengthscale parameter. SpatialPCA applies heuristic methods (i.e., Silverman's "rule-of-thumb" bandwidth [64] for large datasets and non-parametric Sheather & Jones's bandwidth [65] for small datasets) to compute the lengthscale hyperparameter for the kernel function and keeps it fixed during model training. Note that the heuristic methods do not use spatial location information to determine the bandwidth parameter. Consequently, given the same expression matrix, SpatialPCA will be trained with the same bandwidth hyperparameter, regardless of the cell locations. In contrast, SMOPCA, by default, automatically learns the lengthscale through the optimization of marginal likelihood [66]. However, when the data are very sparse and noisy, or the spatial structure is very complicated, it may be challenging to learn the appropriate lengthscale parameter. Therefore, SMOPCA also provides an option for users to specify the lengthscale. The users can tune the lengthscale to achieve the expected spatial clustering pattern based on prior knowledge.

In this study, we assume that $\Sigma_1 = \cdots = \Sigma_d = \Sigma$, meaning Σ_l (for $l = 1, \ldots, d$) is calculated with the same kernel and length scale parameter. The same assumption is used in PPCA (probabilistic principal component analysis [67]), SpatialPCA [33], and SpaVAE [40]. This assumption is imposed to facilitate computation, as there exists closed-form solution for $\widehat{W}_k$. When this assumption is relaxed, allowing each latent

dimension to have its own length scale parameter, the theoretical properties remain valid. However, unlike the case where the covariance is shared, no closed-form solution for $\widehat{W_k}$ can be found. When the covariance functions are different, the maximum marginal likelihood estimation of the matrices W_k is equivalent to an optimization problem with orthogonal constraints $W_k^T W_k = I_d$, often referred as the Stiefel manifold [68, 69]. A numerical optimization algorithm that preserves the orthogonal constraints may be introduced to solve the problem [68]. We will explore this option in our future work.

Different kernel functions can be employed by our model for capturing spatial dependencies. However, SMOPCA does not appear to be very sensitive to the choice of kernel functions and parameters (Additional file 1: Fig. S22, Additional file 1: Fig. S23 and Additional file 1: Fig. S49). For numerical stability, SMOPCA defaults to the Matern kernel (with $\nu = 3/2$), while users also have the flexibility to opt for other kernels (e.g., Cauchy kernel or Gaussian kernel) or other parameters (e.g., $\nu = 5/2$ also achieves good results, as shown in Additional file 1: Fig. S24 and Additional file 1: Fig. S25). The Matern kernel is given by:

$$k(x_i, x_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{\gamma} d(x_i, x_j) \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\gamma} d(x_i, x_j) \right),$$

with positive parameters ν and γ, where $d(\bullet, \bullet)$ is the Euclidean distance, $K_\nu(\bullet)$ is a modified Bessel function and $\Gamma(\bullet)$ is the gamma function. The Matern kernel encompasses a wide range of different kernel functions, named the Matern class after the work of Matern [70]. Particularly, with $\nu = 1/2$, the Matern kernel becomes equivalent to the exponential kernel, and with $\nu \to +\infty$, it is equivalent to Gaussian kernel. The flexibility of the Matern kernel makes it widely applicable for modeling spatially correlated data [71]. Note that the Matern covariance function becomes especially simple and has a closed-form expression as a product of an exponential and a polynomial of order p, when ν is half-integer: $\nu = p + 1/2$, where p is a non-negative integer. As mentioned by Rasmussen et al. [66], $\nu = 3/2$ and $\nu = 5/2$ may be the most interesting cases for machine learning. When $\nu = 1/2$, the process becomes very rough [66], and for $\nu \geq 7/2$, in the absence of explicit prior knowledge about the existence of higher order derivatives, it is probably very hard from finite noisy training examples to distinguish between values of $\nu \geq 7/2$.

Due to the eigen decomposition, the time complexity and space complexity of our current implementation are $O(n^3)$ and $O(n^2)$, respectively, where n is the number of cells/spots in the process. Note that this is similar to SpatialPCA. Additional file 3: Table S2 lists the running time of different methods on different datasets. As shown in Additional file 3: Table S2, SMOPCA is much faster than many existing methods. Matern kernels, with $\nu = (2p + 1)/2$ for all $p \in N$, are of the desired rational function form as analyzed by Jouni Hartikainen and Simo Särkkä [72], and the marginal likelihood of this model can be reformulated as a product of the multivariate normal distributions, which may potentially be solved exactly with classical Kalman filtering theory [72, 73]. In our future work, we aim to further optimize the multi-omics model to reduce both time and space complexity.

Similar to Seurat [37] (e.g., https://satijalab.org/seurat/articles/pbmc3k_tutorial), PCA, and SpatialPCA [33], our method utilizes normalized data, rather than raw expression measurements, as inputs for dimension reduction analysis. Due to various technical limitations, current spatial multi-omics data often contain a significant number of dropout events, leading to excessive zeros in the expression matrix [74]. Therefore, our method may be suboptimal, as it does not consider the dropout events and the inherent mean–variance relationship presents in raw counts, potentially resulting in a loss of inference accuracy and subsequent reduction in analytical power. Similar power losses have been extensively documented in prior studies [53, 75–78] for methods exclusively analyzing normalized data. One possible extension of our model is to assume zero-inflated models such as zero-inflated Gaussian distribution [74] for the normalized data or zero-inflated negative binomial distribution [40] for raw count data. Direct modeling of count data through deep-learning-based models, such as auto-encoder and variational auto-encoder models, is also possible, similar to our previously developed spatial deep generative model, SpaVAE [40], which directly analyzes raw count data and models spatial dependencies of neighboring spots through a Gaussian process (GP) prior [79]. As additional layers of complexity are introduced, the computation required for these models may increase significantly. Therefore, developing computationally efficient methods that account for the count sampling process represents an important direction for future research.

Conclusions

SMOPCA serves as a powerful and flexible framework for dependency-aware dimension reduction in spatial multi-omics analysis. By explicitly incorporating spatial location information, SMOPCA preserves neighborhood similarity in the latent space and effectively captures spatial correlations across multiple omics layers. Extensive experiments on both single-cell and spatial multi-omics datasets demonstrate that SMOPCA delivers competitive or superior performance compared to existing approaches, with enhanced efficiency and robustness, laying a solid foundation for advancing spatial multi-omics data analysis.

Methods

SMOPCA overview

We consider a spatial multi-omics dataset consisting of different types of measurements for n spatial locations (n spots or cells) of a tissue. These spots have known spatial coordinates that are recorded during the experiment. We denote s_i as the spatial coordinates for spot i, with $i \in \{1, \cdots, n\}$. Depending on the technology, the spatial coordinates may vary continuously over a two-dimensional space ($s_i = (s_{i1}, s_{i2}) \in \mathbb{R}^2$) or a three-dimensional space ($s_i = (s_{i1}, s_{i2}, s_{i3}) \in \mathbb{R}^3$). Let's denote $k \in \{1, \ldots, K\}$ as the modality of the data. We denote Y_k as the $m_k \times n$ measurement matrix of the k th modality in the study. The ji th value of Y_k, $Y_{k_{ji}}$ represents the k th modality measurement for j th feature (e.g., gene or protein) on i th location. Following previous studies [33, 80–82], we assume that the measurements have already been normalized (see " Data preprocessing"). Our objective is to conduct dimension reduction on the measurement matrices and learn a joint $d \times n$ factor matrix, denoted as Z, representing the low-dimensional embeddings

of spots. The factor matrix Z contains d factors, and its lth row, Z_l, is a n-vector that represents the lth factor values across n spots. Mathematically, we consider the following model

$$Y_k = W_k Z + E_k, \tag{1}$$

where $k \in \{1, \ldots, K\}$ represents the modality index; W_k is a $m_k \times d$ factor loading matrix; and E_k is a $m_k \times n$ matrix of residual errors. Following GPPCA [39] and Shang et al. [33], we assume that the ji-th element of E_k, $E_{k_{ji}}$, follows an independent normal distribution with mean zero and variance σ_k^2, namely $E_{k_{ji}} \sim N(0, \sigma_k^2)$. As shown in Eq. (1), different modalities of data share the same latent factor matrix Z in SMOPCA, indicating the inferred posterior of latent factors will integrate information from multiple modalities.

The factor model (1) is not identifiable, as for any invertible matrix Λ_k, if W_k, Z is the solution, $W_k\Lambda, \Lambda^{-1}Z$ is also a solution. Following the probabilistic principal component analysis model (PPCA [67]), we further place constraints on W_k and Z to ensure model identifiability. We assume that the loading matrix W_k satisfies the orthonormality constraint or $W_k^\top W_k = I_d$. Under this assumption, we derived a closed-form solution for the maximum marginal likelihood estimation of the factor loading matrix W_k when the covariance function of the factor processes is shared. Existing factor analysis models, such as MOFA [35] and its spatial version MEFISTO [43], were formulated in a probabilistic Bayesian framework and prior distributions were placed on all unobserved variables of the model (i.e., the factors Z and the weight matrices W_k). The key determinant of the MOFA and MEFISTO is the two-level sparsity regularizations applied on the weights W_k. Specifically, an Automatic Relevance Determination (ARD) prior [83] and a spike-and-slab prior [84] were combined together to ensure model identifiability and achieve factor-wise sparsity and feature-wise sparsity [35, 43, 69] for Z and W_k, respectively. Compared to MOFA and MEFISTO, the closed-form marginal likelihood obtained in this work is more computationally efficient and feasible [67].

To explicitly model spatial information and encourage neighborhood similarity in factor values [39], we assume that each Z_l follows a multivariate normal distribution

$$Z_l \sim MVN(0, \Sigma_l), l = 1, \ldots, d, \tag{2}$$

where Σ_l is a $n \times n$ covariance matrix that models the correlation among spatial locations. We assume that spatially proximate spots are more likely to exhibit biological similarity, resulting in a relatively larger value at the corresponding entry of the covariance matrix. Here, we construct the covariance matrix by employing a kernel function defined over spatial coordinates. Specifically, the ij th element of Σ_l is in the form of $k_\gamma(s_i, s_j)$, with $k_\gamma(s_i, s_j)$ being the kernel function.

Different kernel functions can be employed by our model for capturing the spatial dependencies. Three kernel functions have been considered (Additional file 5: Supplementary Note 2), and their effects are displayed in Additional file 1: Fig. S22-S25 and Additional file 1: Fig. S49. As shown in Additional file 1: Fig. S22-S25 and Additional file 1: Fig. S49, SMOPCA does not appear to be very sensitive to the choice of these kernels. For numerical stability, by default, we choose the Matern kernel with $\nu = 3/2$. The Matern kernel with $\nu = 3/2$ between spots i and j is defined as

$$k_{\nu=3/2,\gamma}(\boldsymbol{s}_i,\boldsymbol{s}_j) = \left(1+\frac{\sqrt{3}\|\boldsymbol{s}_i-\boldsymbol{s}_j\|}{\gamma}\right)\exp\left(-\frac{\sqrt{3}\|\boldsymbol{s}_i-\boldsymbol{s}_j\|}{\gamma}\right), \tag{3}$$

where s_i and s_j are spatial locations of spot i and j; γ is the length scale parameter. The strength of spatial correlation is jointly determined by both the distance between spots and the length scale parameter. On the one hand, for the same distance measure, a small γ leads to a weak spatial correlation, while a large γ

leads to a strong spatial correlation [85, 86]. On the other hand, given the parameter γ, if two locations are close to each other, then the corresponding element in covariance matrix will be large, leading to similar factor values on the two locations; and vice versa.

Inference

SMOPCA uses a Maximum Likelihood Estimation (MLE) framework to estimate the parameters and latent variables including $\gamma, \sigma_1, \ldots, \sigma_K, W_1, \ldots, W_K$ and Z. The modality-specific parameters σ_k, W_k are updated iteratively within each modality, while the modality-shared parameter γ is updated across modalities. Specifically, we first marginalize out Z to obtain the marginal likelihood of modality k. We then iteratively perform the MLE on the marginal likelihoods to obtain estimates of W_k and σ_k^2. In some scenarios, users may have the desire to balance the contribution of each modality and reduce the influence of some modalities on the parameter estimates. One possible solution is to impose modality-dependent weights to the log likelihood function of $Y_1, \ldots, Y_K$. Without loss of generality, we assume there is a positive weight $\alpha_{\mathrm{k}}, k = 1, \ldots, K$, for the kth modality. Following popular methods, including SpatialPCA, SpaVAE, SpatialGlue, TotalVI, scMDC and many others, by default, we set $\alpha_{\mathrm{k}} = 1$ for each modality. However, our implementation is quite flexible and allows users to explore and specify predefined modality-dependent weights based on their needs and prior knowledge. Given $\{\alpha_1, \alpha_2.., \alpha_K\}$, we use the weighted joint likelihood of $Y_1, \ldots, Y_K$ to update γ. Lastly, with the weighted joint likelihood of $Y_1, \ldots, Y_K, Z$ and the parameter estimates, the posterior mean of Z is computed as the integrated low-dimensional representation of the input data. Note, in the algorithm, we perform eigen decomposition on the kernel matrix Σ to enable scalable computation (Additional file 5: Supplementary Note 3).

Specifically, denote $\theta = \{\gamma, W_1, \ldots, W_k, \sigma_1^2, \ldots \sigma_K^2\}$ and $\theta_k = \{W_k, \sigma_k^2\}, k = 1, \ldots, K$. According to the prior of residual error $E_{kji} \sim N(0, \sigma_k^2), i = 1, \ldots, n, j = 1, \ldots, m_k, k = 1, \ldots, K$, and the dependency relationship between latent variables, we can derive the conditional distribution Y_k in **Lemma 1**.

Lemma 1. *For model (1), the marginal distribution of Y_k conditional on Z is as follows, for $k = 1, \ldots, K$,*

$$p(Y_k|Z;\theta_K) \propto \left(\sigma_k^2\right)^{-\frac{m_k n}{2}} \exp\left(tr\left(-\frac{(Y_k-W_kZ)(Y_k-W_kZ)^T}{2\sigma_k^2}\right)\right). \tag{4}$$

Assume $\Sigma_1 = \cdots = \Sigma_{\mathrm{d}} = \Sigma$. Based on Lemma 1, we can obtain the maximum marginal likelihood estimation of W_k and σ_k^2 iteratively, denoted as $\widehat{W_k}$ and $\widehat{\sigma}_k^2$, and the weighted maximum joint likelihood estimation of γ, denoted as $\widehat{\gamma}$.

Based on conditional marginal distribution of Y_k in formula (4) and the multivariate normal distribution of Z in formula (2), we can derive the weighted posterior

distribution for each Z_l, $l = 1, \ldots, d$. The posterior expectation for each Z_l can be the final low-dimensional representation as shown in Theorem 1 below.

Theorem 1 *For model (1), we can obtain the multivariate normal distribution of each* $Z_l, l = 1, \ldots, d$

$$Z_l^T \sim MVN\left(\tfrac{1}{2}A^{-1}b_l,\ \tfrac{1}{2}A^{-1}\right), \tag{5}$$

where $A = \frac{1}{2}\left(\sum_{k=1}^{K}\frac{\alpha_k}{\sigma_k^2}I_n + \Sigma^{-1}\right), b_l = \sum_{k=1}^{K}\frac{\alpha_k}{\sigma_k^2}Y_k^\top w_{k_l}$. *Further, plugging the estimators* $\{\widehat{\gamma}, \widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_K^2; \widehat{W_1}, \ldots, \widehat{W_K}\}$ *into equation (5), we can obtain the maximum likelihood estimation for each* Z_l, $l = 1, \ldots, d$,

$$\widehat{Z_l^T} = \widehat{E\left(Z_l^T\right)} = \sum_{k=1}^{K}\left(\sum_{k=1}^{K}\frac{\alpha_k}{\widehat{\sigma}_k^2}I_n + \Sigma^{-1}\right)^{-1}\left(\frac{\alpha_k}{\widehat{\sigma}_k^2}Y_k^\top \widehat{w}_{k_l}\right). \tag{6}$$

Corollary 1 *Suppose that dimension reduction is performed on each modality k separately. For* $k = 1, \ldots, K, l = 1, \ldots, d$, *we can obtain the multivariate normal distribution of each* Z_{kl} *on single modality k of data as.*

$$Z_{k_l}^T \sim MVN\left(\tfrac{1}{2}A_k^{-1}b_{kl},\ \tfrac{1}{2}A_k^{-1}\right), \tag{7}$$

where $A_k = \frac{1}{2}\left(\frac{1}{\sigma_k^2}I_n + \Sigma^{-1}\right), b_{kl} = \frac{1}{\sigma_k^2}Y_k^\top w_{k_l}$.

When modality weight $\alpha_k \geq 1$, with some derivations, we show that the latent factors learned by SMOPCA, which integrates information from multimodal data, yields lower uncertainty compared to conducting dimension reduction on each modality of data, separately. The detailed derivation of equation (4-7), the parameter inference algorithm [66, 87, 88] and the stability analysis of SMOPCA can be found in Additional file 5: Supplementary Note 3.

Model implementation

SMOPCA is implemented in Python 3 (version 3.10.0) using scanpy [89] (version 1.9.1), numpy [90] (version 1.23.5), scipy [91] (version 1.10.0), and scikit-learn [92] (version 1.2.1). SMOPCA involves two hyperparameters: the dimensionality d of the latent vector and the γ parameter, which acts as the length scale of the kernel. For the number of PCs d, existing popular works, such as SpaVAE [40] (default $d = 20$), SpatialPCA (default 10 PCs), BayesSpace [32] (default 15 PCs), SpaGCN [31] (default 50 PCs), and PCA + K-means (from Seurat package; default 50 PCs), all use default values. In SMOPCA, the dimension of the latent space d is set to 20 by default, if the user does not specify a value for this parameter. We conducted a sensitivity analysis for the choice of d. Additional file 1: Fig. S26 and Additional file 1: Fig. S27 demonstrate that SMOPCA is robust to the number of spatial PCs. By default, the length scale parameter γ is learned automatically in the training process. K-means clustering ("sklearn.cluster.KMeans") is performed on the low-dimensional vectors to predict the final clustering labels for SMOPCA.

Data preprocessing

In accordance with Lin et al. [25], we preprocessed and normalized the data from each modality separately. Initially, genes and proteins with zero counts were filtered out. Subsequently, the count matrix was normalized by library size (the total read count), followed by log transformation and scaling to achieve zero mean and unit variance. This library size-based normalization method has been used in other algorithms, such as SPARK [53] and BOOST-MI [93], and proved to be effective in practice (Additional file 5: Supplementary Note 4 and Additional file 1: Fig. S28-S29). Following scMDC [25], we mapped ATAC reads to gene regions and collapsed the peak matrix into a gene activity matrix, adhering to the established protocol from the Satija lab. The gene activity matrix was preprocessed and normalized using the same method as applied to mRNA data. Following Tian et al. [40], we utilized SPARK [53] to filter genes, retaining a set of spatially variable genes (SVG) for the analysis of spatial-CITE-seq, MISAR-seq, and spatial ATAC–RNA-seq data. Gene significance is assessed using a False Discovery Rate (FDR) threshold of 0.05, applying the Benjamini-Yekutieli (BY) procedure, which is effective under arbitrary dependence across genes [94].

Data simulation

Simulation setting I: simulate spatial information for real single-cell multi-omics data

In simulation studies, we assume that cells from the same cluster are close to each other and different cell types exhibit layer-wise patterns, as inspired by the six-layered human dorsolateral prefrontal cortex (DLPFC) datasets [50]. Given a single-cell multi-omics dataset with n cells, we assign the cells to a simulated 2D grid represented by $\{1,2,\ldots,m\} \times \{1,2,\ldots,m\}$, where $m = \lceil \sqrt{n} \rceil$ is the smallest integer greater than or equal to $\sqrt{n}$. Initially, we rearrange the cells by randomly ordering the cell types. Subsequently, within each cell type, we randomly permutate the cells. Afterwards, we allocate the cells to the grid locations row by row and column by column, following the order:

$$\begin{cases} (1,1),(1,2),\cdots,\left(\lceil \frac{n}{m} \rceil, m\right), \text{if} n - \lfloor \frac{n}{m} \rfloor \times m = 0 \\ (1,1),(1,2),\cdots,\left(\lceil \frac{n}{m} \rceil, n - \lfloor \frac{n}{m} \rfloor \times m\right), \text{if} n - \lfloor \frac{n}{m} \rfloor \times m \neq 0 \end{cases},$$

where $\lfloor \frac{n}{m} \rfloor$ is the largest integer less than or equal to $\frac{n}{m}$. It is worth noting that the simulated coordinates have the flexibility to undergo transformation or rotation, resulting in a distinct set of locations. To ensure a thorough comparison, we randomly generate 10 simulated locations for each real dataset and then evaluate the clustering performance for each simulated dataset.

Simulation setting II: simulate multi-omics data with three modalities based on real SRT data

Due to a lack of available data, simulations and real data analyses in current single-cell multi-omics and spatial multi-omics methodology research (e.g., TotalVI, MOFA, scMDC, MEFISTO, SpaVAE, SpatialGlue, and others) has primarily focused on two omics datatypes. Here, we simulated datasets with 3 omics modalities to demonstrate the capability of SMOPCA in analyzing multi-omics data with more than two

modalities. SRTsim [52], developed by Zhu et al. in 2023, is an SRT-specific simulator for scalable, reproducible, and realistic SRT simulations. We employed SRTsim to generate simulated data, which not only maintains various expression characteristics of the SRT data but also preserves spatial patterns. We used the LIBD human dorsolateral prefrontal cortex (DLPFC) dataset [50], which includes 12 tissue sections spanning six neuronal layers and the white matter from three human brains, as a reference for tissue-wise model fitting and data simulation. For each section, we simulated three count matrices from three omics modalities with different random seeds.

UMAP location generation for real single-cell multi-omics data

In order to analyze real single-cell multi-omics datasets, we generated a 2D UMAP [44] coordinates by the Seurat [37] package. We normalized the gene expression matrix and selected the top 2000 highly variable genes (CITE-seq data) or mapped genes (SMAGE-seq data) as input features for the principal component analysis (PCA) analysis. In accordance with the online Seurat tutorial, we utilized the top 50 principal components (PCs) for conducting UMAP analysis.

Evaluation metrics

The clustering performance is assessed using three metrics: Adjusted Rand Index [95] (ARI), Normalized Mutual Information [96] (NMI), and Adjusted Mutual Information [97] (AMI). These metrics gauge the concordance between predicted labels and the ground-truth labels. When the ground-truth labels were unavailable, we introduced Moran's I score and local inverse Simpson's index (LISI) score [54] to quantitatively measure the spatial smoothness of the identified spatial domains. Additionally, following Shang et al. [33], we employed the McFadden-adjusted pseudo R^2 [59] to evaluate the predictive ability of the learned latent representations from each method in predicting the true spatial domains.

Given two cluster assignments U and V, we define a, the number of pairs of two objects in the same group in both U and V; b, the number of pairs of two objects in different groups in both U and V; c, the number of pairs of two objects in the same group in U but in different groups in V; and d, the number of pairs of two objects in different groups in U but in the same group in V. The ARI is formally defined as follows:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}.$$

Let $U = \{U_1, U_2, \ldots, U_{C_U}\}$ and $V = \{V_1, V_2, \ldots, V_{C_V}\}$, which are two cluster assignments on a set of n data points and have C_U and C_V clusters, respectively. NMI is defined as the mutual information between U and V divided by the entropy of U and V. Specifically,

$$NMI = \frac{I(U,V)}{\max\{H(U), H(V)\}},$$

where $I(U,V) = \sum_{p=1}^{C_U}\sum_{q=1}^{C_V}|U_p \cap V_q| \log\frac{n|U_p \cap V_q|}{|U_p| \times |V_q|}$ represents the mutual information between U and V; $H(U) = -\sum_{p=1}^{C_U}|U_p| \log\frac{|U_p|}{n}$ and $H(V) = -\sum_{q=1}^{C_V}|V_q| \log\frac{|V_q|}{n}$ are the entropies.

Similarly, AMI is defined as

$$AMI = \frac{I(U,V) - E\{I(U,V)\}}{\max\{H(U),H(V)\} - E\{I(U,V)\}}.$$

The extra component $E\{I(U,V)\}$ is the expected mutual information between two random clusters [97]. Note that ARI, NMI, and AMI will equal 1 when the two cluster assignments are identical, and smaller than one when the two cluster assignments differ in their pairing. The AMI, NMI, and ARI metrics are calculated with scikit-learn [92] functions (i.e., "*sklearn.metrics.adjusted_mutual_info_score*", "*sklearn.metrics.normalized_mutual_info_score*", and "*sklearn.metrics.adjusted_rand_score*").

Moran's I score is a metric used to assess the global spatial autocorrelation of gene expression levels. Let x represent a gene's expression level. Moran's I score is then defined as

$$I = \frac{N}{W}\frac{\sum_i\sum_j\left[w_{ij}(x_i-\overline{x})(x_j-\overline{x})\right]}{\sum_i(x_i-\overline{x})^2},$$

where x_i and x_j are the gene expressions at spot i and j, $\overline{x}$ is the average expression of the gene, N is the number of spots, w_{ij} is spatial weight between spot i and j, and $W = \sum w_{ij}$. We calculate w_{ij} by k-nearest neighbors using spatial coordinates and k is set to 5. The Moran's I score is calculated by "squidpy.gr.spatial_autocorr" from the Squidpy (Spatial Single Cell Analysis in Python) package [98].

The LISI score reflects the effective number of different categories (e.g., clusters) represented in the local neighborhood of each cell/spot and is computed as

$$S = \frac{1}{\sum_{c=1}^{C\prime} p(c)},$$

where $p(c)$ is the probability that the cluster label c is in the local neighborhood, and $C\prime$ is the total number of spatial domains. A smaller LISI score suggests that the cells are less mixed. The LISI score is calculated using the "*compute_lisi*" function from the LISI [54] R package with default parameters.

We also evaluated the information embedded in the outputs of diverse methods used for predicting true spatial domains. To conduct this assessment, we treated the true spatial domains as the outcome and employed a multinomial regression model, with the extracted low-dimensional components serving as predictors. Subsequently, we calculated the McFadden-adjusted pseudo R^2 to evaluate the predictive performance of these variables in predicting the ground truth [59]. A higher pseudo R^2 suggests that the method is capable of extracting informative latent representations for predicting the true spatial domains. The pseudo R^2 Statistics is calculated using the "*PseudoR2*" function from the DescTools [99] R package.

Public real datasets

Multiple CITE-seq datasets are employed in our experiments, encompassing the Peripheral Blood Mononuclear Cells (PBMC) dataset and a series of Mouse Spleen Lymph Node datasets (SLN111D1, SLN111D2, SLN208D1, and SLN208D2). CITE-seq was developed based on the scRNA-seq technology to profile mRNA expression and quantify surface protein simultaneously at the cellular level [20, 21, 25]. Specifically, CITE-seq

enables the counting of Antibody-Derived Tags (ADT) to measure the protein abundance. Each cell, labeled with ADTs and DNA-barcoded microbeads, is encapsulated in a droplet for single-cell sequencing [26]. The PBMC dataset was obtained from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets), and cellular type annotations were provided by Specter [38] (https://github.com/canzarlab/Specter).

As shown in Additional file 2: Table S1, there are 3762 cells, 33,538 genes, and 49 proteins in the downloaded PBMC dataset. The mouse spleen lymph node datasets, along with cell type annotations, were downloaded from the TotalVI [34] Github repository (https://github.com/YosefLab/totalVI_reproducibility). As shown in Additional file 2: Table S1, the SLN111D1, SLN111D2, SLN208D1, and SLN208D2 datasets contain 8853, 6949, 8371, and 6678 cells, respectively. The SLN111D1 and SLN111D2 datasets each measured 13,553 genes and 110 proteins, while the SLN208D1 and SLN208D2 datasets measured 13,553 genes and 207 proteins.

The 10X Single-Cell Multiome ATAC + Gene Expression (SMAGE-seq [22, 23]) datasets (PBMC3K and PBMC10K) are accessible from the 10X Genomics resource center (https://www.10xgenomics.com/resources/datasets). In SMAGE-seq, single-cell ATAC-seq data provides chromatin accessibility information that complements single-cell mRNA-seq data. By learning a joint embedding of both modalities, we can achieve higher-resolution cell types. Both datasets are derived from human peripheral blood mononuclear cells (PBMCs), comprising approximately 3000 and 10,000 cells, referred to as PBMC3K and PBMC10K, respectively. The detailed information is given in Additional file 2: Table S1. For SMAGE-seq datasets, mRNA counts are directly downloaded online, while ATAC gene counts are derived from the raw data using the procedure employed in scMDC [25]. This involves filtering reads based on ATAC peak region fragments, nucleosome signals, and Transcription Start Site (TSS) enrichment. Subsequently, each read is mapped to a gene region using the"*GeneActivity*"function in Signac [100] (v1.4.0), following the established protocol proposed by the Satija lab. PBMC cells are annotated using the label transfer method in Seurat [37], utilizing the reference dataset "pbmc_10k_v3.rds" (https://www.dropbox.com/s/zn6khirjafoyyxl/pbmc_10k_v3.rds?dl=0), as provided by the Satija lab. After preprocessing, the ATAC data, which is mapped to the gene regions, is processed in the same way as for the mRNA data.

The LIBD human dorsolateral prefrontal cortex (DLPFC) datasets [50] were downloaded from http://spatial.libd.org/spatialLIBD/. The 10 × Genomics Visium platform was used by Maynard et al. in 2021 to obtain the DLPFC data and define the spatial topography of gene expression in the six-layered human dorsolateral prefrontal cortex. There are 12 sections/datasets (with section number 151507, 151,508, 151,509, 151,510, 151,669, 151,670, 151,671, 151,672, 151,673, 151,674, 151,675, and 151,676) in the DLPFC data.

The 12 sections contain 3000–4000 spots that span six neural layers and the white matter. The layers were manually annotated by the original authors. We selected 5000 HVGs for each section and used the filtered gene expression matrix as input for SRTsim to generate simulated 3-omics datasets.

The spatial-CITE-seq data [27] can be downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213264). In 2023, Liu et al. [27]

extended co-indexing of transcriptomes and epitopes (CITE) into the spatial dimension and developed the spatial co-indexing of transcriptomes and epitopes for multi-omics mapping by highly parallel sequencing (spatial-CITE-seq). This technique uses a cocktail of approximately 200–300 ADTs to stain a tissue slide, followed by deterministic in-tissue barcoding of both antibody-derived DNA tags and mRNAs, enabling spatially resolved, high-plex protein, and transcriptome co-profiling. They profiled >200 proteins and the whole transcriptome in human tissues, revealing spatially distinct germinal center reactions in tonsil. Here, we used human tonsil sample for the analysis, which contains 2491 spots with 28,417 genes and 283 proteins. Following SpaVAE, the mRNA counts were filtered by SPARK [53] with the default settings, resulting in 959 retained spatially variable genes.

The Stereo-CITE-seq mouse thymus dataset [58] was downloaded from https://zenodo.org/records/10362607. Liao et al. combined CITE-seq and Stereo-seq to develop the Stereo-CITE-seq technology, and studied the murine thymus tissue samples by co-detection of mRNAs and protein markers. There are 4 sections in the dataset, and for our study, we used the first section, which had 4697 spots, 23,622 genes and 51 protein markers. Following SpatialGlue, the mRNA counts were filtered by Seurat to select 2000 highly variable genes.

The microfluidic indexing-based spatial ATAC and RNA sequencing (MISAR-seq [28]) data, which profiled the mouse embryonic E15.5 brain, could be downloaded from https://doi.org/10.5281/zenodo.7480069. Manually annotated labels were obtained based on Allen Mouse Brain Atlas [101] and H&E image (Additional file 1: Fig. S62). MISAR-seq, introduced by Jiang et al. in 2023, is a spatially resolved method for joint profiling of chromatin accessibility and gene expression [28]. It was first applied to study consecutive developmental stages of the mouse brain, and revealed the dynamic spatiotemporal regulatory mechanisms involved in establishing the brain's complex architecture, owing to the high-quality transcriptome-open chromatin status in the different anatomical brain regions. We exactly follow the pipeline provided by the author of MISAR-seq to preprocess and filter the data. After preprocessing, 1949 out of the initial 2500 spots were retained. ArchR calls macs2 [102] (2.2.7.1) to detect the peak regions, subsequently computing counts for each peak per cell. The raw peak matrix contains 105,350 peaks across 1949 spots. We then use the "*FindTopFeatures*" function in Signac [100] (v1.4.0) to select the peaks that are at least detected in 200 cells for the analysis. 47,287 out of 105,350 peaks passed the filtering criteria. The peaks are further mapped to gene regions using the "*GeneActivity*" function in Signac [100] (v1.4.0), following the established protocol proposed by the Satija lab. Following SpaVAE, 2144 spatially variable genes and 3000 spatially variable mapped genes were selected by SPARK [53] with the default parameters.

The spatial ATAC–RNA-seq mouse brain data [60] was downloaded from https://zenodo.org/records/10362607. Zhang et al. [60] presented two technologies (i.e., spatial ATAC–RNA-seq and spatial CUT&Tag-RNA-seq) for spatially resolved, genome-wide, joint profiling of the epigenome, and transcriptome on the same tissue section at near-single-cell resolution. The developed spatial ATAC–RNA-seq was applied to analyze juvenile (P22) mouse brain coronal sections for joint profiling of chromatin accessibility with transcriptome. The downloaded dataset contains 9215 cells, 22,914 genes, and 121,068 peaks. The peaks were mapped to gene regions using the"*GeneActivity*"function in Signac [100] (v1.4.0),

following the established protocol proposed by the Satija lab. Following existing works [40, 42], 2420 spatially variable genes and 3000 spatially variable mapped genes were selected by SPARK [53] with the default parameters.

Competing methods

The following state-of-the-art tools are employed as competing methods for the analysis of protein data: BREM-SC (https://github.com/tarot0410/BREMSC), CiteFuse (https://github.com/SydneyBioX/CiteFuse), SC3 (https://github.com/hemberg-lab/SC3), scMDC (https://github.com/xianglin226/scMDC), Seurat (https://github.com/satijalab/seurat), TotalVI (https://scvi-tools.org), Tscan (https://github.com/zji90/TSCAN), MEFISTO (https://biofam.github.io/MOFA2/MEFISTO), SpaVAE (https://github.com/ttgump/spaVAE/) and SpatialGlue (https://spatialglue-tutorials.readthedocs.io/en/latest/index.html). In the context of ATAC data analysis, the utilized methodologies include chromVAR (https://github.com/GreenleafLab/chromVAR), cisTopic (https://github.com/aertslab/cisTopic), LSA (implemented by scikit-learn), PeakVI (https://scvi-tools.org), SCALE (https://github.com/jsxlei/SCALE), scMDC (https://github.com/xianglin226/scMDC), Seurat (https://github.com/satijalab/seurat), MEFISTO (https://biofam.github.io/MOFA2/MEFISTO), SpaVAE (https://github.com/ttgump/spaVAE/), and SpatialGlue (https://spatialglue-tutorials.readthedocs.io/en/latest/index.html). MEFISTO, SpaVAE, and SpatialGlue were recently developed to model spatial correlations among neighboring spots in the spatial datasets. Furthermore, SpatialPCA (https://github.com/shangll123/SpatialPCA) and PCA + K-means have been introduced as additional competing methods for comprehensive comparisons across all analyses. Methods that model multimodal data, such as BREM-SC, CiteFuse, Seurat, scMDC, MEFISTO, SpaVAE, and SpatialGlue utilize mRNA data in conjunction with either protein or ATAC data as inputs. Conversely, methods built upon ATAC data, including PeakVI, SCALE, and cisTopic, use ATAC data as their primary input. For ATAC data, we utilize a cell-to-gene matrix as input for scMDC, Seurat, MEFISTO, SpatialGlue, SMOPCA, SpatialPCA, and PCA + K-means. This matrix is constructed by mapping ATAC reads onto the gene regions. Seurat, originally designed for CITE-seq data, is adapted in this context by applying the weighted-nearest neighbor (WNN) algorithm to the SMAGE-seq data. For all other methods, such as SpatialPCA and PCA + K-means, different types of data are pre-processed and normalized separately and then concatenated into a unified input. We follow the tutorials provided by the original authors to preprocess data, configure model parameters, and conduct experiments for the competing methods, including BREM-SC, CiteFuse, SC3, scMDC, Seurat, Tscan, PeakVI, SCALE, cisTopic, MEFISTO, SpaVAE, SpatialGlue, and SpatialPCA. To maintain consistency, all methods utilize identical sets of genes/features in both RNA and ATAC data, and they incorporate all proteins in the CITE-seq or spatial-CITE-seq dataset. When normalization is necessary, we adhere to the normalization method outlined in scMDC [25]. Note that *K*-means clustering (implemented using "sklearn.cluster.KMeans") is performed on the low-dimensional representations learned by dimension reduction methods, such as SMOPCA, TotalVI, SpaVAE, SpatialPCA, and PCA, to derive the final cell type results.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03576-9.

Additional file 1: Fig. S1-S71.

Additional file 2: Table S1.

Additional file 3: Table S2.

Additional file 4: Table S3.

Additional file 5: Supplementary Note 1-4.

Additional file 6: Review history.

Acknowledgements
We thank Dr. Xiang Lin from New Jersey Institute of Technology for providing detailed information about data preprocessing procedures used in scMDC.

Review history
The review history is available as Additional File 6.

Peer review information
Veronique van den Berghe and Kevin Pang were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
M.C., R.C. and J.Z. designed the method. M.C., R.C., J.H., and J.Z. conducted experiments. R.C., M.C., and J.Z. derived the theoretical properties. J.C. and J.Z. supervised the study. M.C., R.C., J.H., J.C., and J.Z. wrote the manuscript. All authors read and approved the final manuscript.

Funding
J.Z. is supported by the Natural Science Foundation of Jiangsu Province (grant no. BK20230781) and the National Natural Science Foundation of China (grant no. 62306134). R.C. is supported by the National Natural Science Foundation of China (grant no. 72371186). The work was also supported by Mayo Clinic Center for Individualized Medicine (J.C.).

Data availability
The data that support the findings of this study are publicly available online. Preprocessed genomics datasets can be found at Zenodo: https://doi.org/10.5281/zenodo.15187362 [103]. PBMC dataset is available on 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets) and the cell type labels can be downloaded from the Github of Specter (https://github.com/ canzarlab/Specter). The mouse spleen lymph node datasets [104] (SLN111D1, SLN111D2, SLN208D1, and SLN208D2) and the cell type labels are provided by TotalVI (https://github.com/YosefLab/totalVI_reproducibility). These datasets were sequenced in two batches. The SMAGE-seq datasets (PBMC3K and PBMC10K) can be downloaded from the 10X Genomics website (https://www.10xgenomics.com/resources/datasets). Labels are transferred by Signac [100] (v1.4.0) from the annotated datasets. The DLPFC datasets [105] can be downloaded from http://spatial.libd.org/spatialLIBD. The spatial-CITE-seq data [106] of human tonsil sample can be downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213264). The Stereo-CITE-seq mouse thymus dataset and the spatial ATAC–RNA-seq mouse brain data [107] can be downloaded from https://zenodo.org/records/10362607. The mouse embryonic E15.5 brain [108] (MISAR-seq) data can be obtained from https://doi.org/10.5281/zenodo.7480069.

Code availability
An open-source software implementation of SMOPCA is publicly available under the MIT license on GitHub: https://github.com/cmhimself/SMOPCA [109] and Zenodo: https://doi.org/10.5281/zenodo.15187362 [103].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 February 2024 Accepted: 12 April 2025
Published online: 21 May 2025

References

1. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24:494–515.

2. Zhou X, Franklin RA, Adler M, Jacox JB, Bailis W, Shyer JA, Flavell RA, Mayo A, Alon U, Medzhitov R. Circuit design features of a stable two-cell system. Cell. 2018;172:744-757.e717.
3. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet. 2021;22:71–88.
4. Bich L, Pradeu T, Moreau JF. Understanding multicellularity: the functional organization of the intercellular space. Front Physiol. 2019;10:1170.
5. Deng Y, Bai Z, Fan R. Microtechnologies for single-cell and spatial multi-omics. Nature Reviews Bioengineering. 2023;1:769–84.
6. Dimitrov D, Türei D, Garrido-Rodriguez M, Burmedi PL, Nagai JS, Boys C, Ramirez Flores RO, Kim H, Szalai B, Costa IG, et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun. 2022;13:3224.
7. Method of the year 2019: single-cell multimodal omics. Nat Methods. 2020;17:1-1.
8. Zhang Y, Tang Y, Sun S, Wang Z, Wu W, Zhao X, Czajkowsky DM, Li Y, Tian J, Xu L, et al. Single-cell codetection of metabolic activity, intracellular functional proteins, and genetic mutations from rare circulating tumor cells. Anal Chem. 2015;87:9761–8.
9. Li YE, Preissl S, Hou X, Zhang Z, Zhang K, Qiu Y, Poirion OB, Li B, Chiou J, Liu H, et al. An atlas of gene regulatory elements in adult mouse cerebrum. Nature. 2021;598:129–36.
10. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581:303–9.
11. Klein Allon M, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz David A, Kirschner Marc W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.
12. Macosko Evan Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas Allison R, Kamitaki N, Martersteck Emily M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
13. Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al. A human cell atlas of fetal chromatin accessibility. Science. 2020;370: eaba7612.
14. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, et al. A human cell atlas of fetal gene expression. Science. 2020;370: eaba7721.
15. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, et al. Molecular architecture of the developing mouse brain. Nature. 2021;596:92–6.
16. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184:5985-6001.e5919.
17. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature. 2019;566:490–5.
18. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566:496–502.
19. The Tabula Sapiens C, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaup B, Brown P, et al: The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science, 376:eabl4896.
20. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, Roush T, Herrera A, Papalexi E, Ouyang Z, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat Methods. 2019;16:409–12.
21. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35:936–9.
22. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37:1452–7.
23. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Tay T, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell. 2020;183:1103-1116.e1120.
24. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, Duerr RH, Chen K, Ding Y, Chen W. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 2020;48:5814–24.
25. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 2022;13:7705.
26. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14:865–8.
27. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, Deng Y, Nam J, Gao F, Bordignon P, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Nat Biotechnol. 2023;41:1405–9.
28. 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;20:1048–57.
29. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. Cell. 2020;183:1665-1681.e1618.
30. Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. Nat Commun. 1873;2021:12.
31. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M. 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.
32. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol. 2021;39:1375–84.
33. Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun. 2022;13:7203.
34. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N. Joint probabilistic modeling of single-cell multiomic data with totalVI. Nat Methods. 2021;18:272–82.

35. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14: e8124.
36. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics. 2020;36:4137–43.
37. 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;184:3573-3587.e3529.
38. Do VH, Rojas Ringeling F, Canzar S. Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data. Genome Res. 2021;31:677–88.
39. Gu M, Shen W. Generalized probabilistic principal component analysis of correlated data. J Mach Learn Res. 2020;21:Article 13.
40. Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Dependency-aware deep generative models for multitasking analysis of spatial omics data. Nat Methods. 2024;21:1501–13.
41. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
42. 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. Nature Methods. 2024;21(9):1658–67.
43. Velten B, Braunger JM, Argelaguet R, Arnol D, Wirbel J, Bredikhin D, Zeller G, Stegle O. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. Nat Methods. 2022;19:179–86.
44. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018. https://arxiv.org/pdf/1802.03426.
45. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14:975–8.
46. Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, Aerts S. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat Methods. 2019;16:397–400.
47. Dumais ST. Latent semantic analysis. Annual Review of Information Science and Technology (ARIST). 2004;38:189–230.
48. Ashuach T, Reidenbach DA, Gayoso A, Yosef N. PeakVI: a deep generative model for single-cell chromatin accessibility analysis. Cell Reports Methods. 2022;2: 100182.
49. Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, Zhang M, Jiang T, Zhang QC. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat Commun. 2019;10:4576.
50. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.
51. Li Z, Zhou X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. Genome Biol. 2022;23:168.
52. Zhu J, Shang L, Zhou X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. Genome Biol. 2023;24:39.
53. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.
54. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods. 2019;16:1289–96.
55. Reilly EC, Sportiello M, Emo KL, Amitrano AM, Jha R, Kumar ABR, Laniewski NG, Yang H, Kim M, Topham DJ. CD49a identifies polyfunctional memory CD8 T cell subsets that persist in the lungs after influenza infection. Front Immunol. 2021;12: 728669.
56. Fischer MB, Goerg S, Shen L, Prodeus AP, Goodnow CC, Kelsoe G, Carroll MC. Dependence of germinal center B cells on expression of CD21/CD35 for survival. Science. 1998;280:582–5.
57. Takai T. Roles of Fc receptors in autoimmunity. Nat Rev Immunol. 2002;2:580–92.
58. Liao S, Heng Y, Liu W, Xiang J, Ma Y, Chen L, Feng X, Jia D, Liang D, Huang C, et al. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. bioRxiv. 2023:2023.2004.2028.538364. https://www.biorxiv.org/content/10.1101/2023.04.28.538364v1.
59. McFadden D. Conditional logit analysis of qualitative choice behavior. 1973.
60. 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;616:113–22.
61. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2002;2018:9.
62. Tian T, Zhong C, Lin X, Wei Z, Hakonarson H. Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning. Genome Res. 2023;33:232–46.
63. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:9.
64. Silverman BW. Density estimation for statistics and data analysis. Routledge; 2018.
65. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J Roy Stat Soc: Ser B (Methodol). 1991;53:683–90.
66. Williams CK, Rasmussen CE. Gaussian processes for machine learning. MA: MIT press Cambridge; 2006.
67. Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc Ser B Stat Methodol. 1999;61:611–22.
68. Wen Z, Yin W. A feasible method for optimization with orthogonality constraints. Math Program. 2013;142:397–434.
69. Gu M, Shen W. Generalized probabilistic principal component analysis of correlated data. J Mach Learn Res. 2020;21:1–41.
70. Stein ML: Interpolation of spatial data: some theory for kriging. Springer Science & Business Media; 2012.
71. Gelfand AE, Diggle P, Guttorp P, Fuentes M. Handbook of spatial statistics. CRC Press; 2010.

72. Hartikainen J, Särkkä S: Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In,. IEEE international workshop on machine learning for signal processing. IEEE. 2010;2010:379–84.
73. Gu M, Xu Y. Fast nonseparable Gaussian stochastic process with application to methylation level interpolation. J Comput Graph Stat. 2020;29:250–60.
74. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241.
75. Kim TH, Zhou X, Chen M. Demystifying "drop-outs" in single-cell UMI data. Genome Biol. 2020;21:196.
76. Sun S, Hood M, Scott L, Peng Q, Mukherjee S, Tung J, Zhou X. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 2017;45: e106.
77. Lea AJ, Tung J, Zhou X. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. PLoS Genet. 2015;11: e1005650.
78. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics. 2019;35:487–96.
79. Casale FP, Dalca AV, Saglietti L, Listgarten J, Fusi N: Gaussian process prior variational autoencoders. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 10390–10401. Montréal, Canada: Curran Associates Inc.; 2018:10390–10401.
80. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
81. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.
82. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
83. MacKay DJ: Bayesian methods for backpropagation networks. In: Models of neural networks III: association, generalization, and representation. New York City: Springer; 1996. pp. 211–254.
84. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. J Am Stat Assoc. 1988;83:1023–32.
85. Vaart A, Zanten J. Adaptive bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. The annals of statistics 37. Ann Stat. 2009;37:2655.
86. Bhattacharya A, Pati D. Posterior contraction in Gaussian process regression using Wasserstein approximations. Information and Inference: A Journal of the IMA. 2017;6:416–40.
87. Saad Y: Numerical methods for large eigenvalue problems: revised edition. SIAM; 2011.
88. Kokiopoulou E, Chen J, Saad Y. Trace optimization and eigenproblems in dimension reduction methods. Numerical Linear Algebra with Applications. 2011;18:565–602.
89. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
90. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ. Array programming with NumPy. Nature. 2020;585:357–62.
91. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J. SciPy 1.0: fundamental algorithms for scientific computing in python. Nature methods. 2020;17:261–72.
92. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
93. Jiang X, Xiao G, Li Q. A Bayesian modified Ising model for identifying spatially variable genes from spatial transcriptomics data. Stat Med. 2022;41:4647–65.
94. Yoav B, Daniel Y. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29:1165–88.
95. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50.
96. Strehl A, Ghosh J. Cluster ensembles –- a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2003;3:583–617.
97. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res. 2010;11:2837–54.
98. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19:171–8.
99. Signorell A, Aho K, Alfons A, Anderegg N, Aragon T, Arppe A, Baddeley A, Barton K, Bolker B, Borchers HW. DescTools: tools for descriptive statistics. R package version 099. 2019;28:17. https://andrisignorell.github.io/DescTools/authors.html.
100. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18:1333–41.
101. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.
102. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.
103. Chen M, Cheng R, He J, Chen J, Zhang J. SMOPCA: spatially aware dimension reduction integrating multi-omics improves the efficiency of spatial domain detection. Zenodo. 2025. https://doi.org/10.5281/zenodo.15187362.
104. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N. Joint probabilistic modeling of single-cell multiomic data with totalVI. Datasets. 2021. https://github.com/YosefLab/totalVI_reproducibility.
105. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Datasets. 2021. http://spatial.libd.org/spatialLIBD.

106. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, Deng Y, Nam J, Gao F, Bordignon P, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Datasets. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213264.
107. 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. Datasets. 2024. https://zenodo.org/records/10362607.
108. 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. Datasets. 2023. https://doi.org/10.5281/zenodo.7480069.
109. Chen M, Cheng R, He J, Chen J, Zhang J. SMOPCA: spatially aware dimension reduction integrating multi-omics improves the efficiency of spatial domain detection. Github. 2025. https://github.com/cmhimself/SMOPCA.

Publisher's Note