nature communications

Article

https://doi.org/10.1038/s41467-026-70821-5

SMART: spatial multi-omic aggregation using graph neural networks and metric learning

Received: 26 January 2025

Accepted: 6 March 2026

Published online: 27 March 2026

Check for updates

Zhihua Du[1], Qiyi Chen[1,2], Weiliang Huang[1,2,3], Jinmiao Chen[4,5,6] ✉ & Xubin Zheng[2,3] ✉

Spatial multi-omics enables the exploration of tissue microenvironments and heterogeneity from the perspective of different omics modalities across distinct spatial domains within tissues. To jointly analyze the spatial multi-omics data, computational methods are desired to integrate multiple omics with spatial information into a unified space. Here, we present SMART (Spatial Multi-omic Aggregation using gRaph neural networks and meTric learning), a computational framework for spatial multi-omic integration. SMART leverages a modality-independent modular and stacking framework with spatial coordinates and adjusts the aggregation using triplet relationships. SMART excels at accurately identifying spatial regions of anatomical structures, compatible with spatial datasets of any type and number of omics layers, while demonstrating exceptional computational efficiency and scalability on large datasets. Moreover, a variant of SMART, SMART-MS, expands its capabilities to integrate spatial multi-omics data across multiple tissue sections. In summary, SMART provides a versatile, efficient, and scalable solution for integrating spatial multi-omics data.

In recent years, spatial single-omics technologies, such as spatial transcriptomics, epigenomics, and metabolomics have established a vital role and are regarded as next generation techniques in biological and medical research. The emergence of spatial multi-omics enables simultaneous detection of different modal omics within the same tissue section, which offers valuable perspectives for in-depth analysis and understanding of gene regulation and microenvironment. Currently, the predominant spatial multi-omics technologies encompass simultaneous detection of spatial transcriptomics and epigenomics, such as CUT&Tag-RNA-seq[1] and MISAR-seq[2], as well as the concurrent detection of spatial transcriptomics and proteomics including SM-Omics[3], SPOTS[4], GeoMX DSP[5], Stereo-CITE-seq[6] and 10x Genomics Visium CytAssist. Different omics can provide complementary information of the tissues, but also pose an urgent demand for computational methods capable of integrating multi-omics data and accurately identifying distinct spatial domains, such as anatomical structures or cell types, within tissues.

The main challenge in integrating spatial multi-omics data lies in the inherent biological heterogeneity across different omics, which often exhibit distinct data dimensionalities and distributions. Moreover, the incorporation of spatial coordinate information complicates the integration of these diverse omic data into a unified representation. Graph neural network-based approaches have demonstrated promising performance in modeling spatial information within spatial

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. [2]Dongguan Key Laboratory for AI and Dynamical Systems, School of Computing and Information Technology, Great Bay University, Guangdong, China. [3]Institute for Artificial Intelligence, Great Bay University, Guangdong, China. [4]Center for Biomedical Data Science and Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore, Singapore. [5]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [6]Immunology Translational Research Program, Department of Microbiology and Immunology, Yong Loo Lin School of Medicine, National University of Singapore (NUS), Singapore, Singapore.
✉e-mail: micchenj@nus.edu.sg; xbzheng@gbu.edu.cn

transcriptomics, such as SpaGCN[7], CCST[8], STAGATE[9], GraphST[10], conST[11], DeepST[12], and SpaceFlow[13], but these approaches were originally developed for spatial transcriptomics data and have not yet been applied to spatial multi-omics data integration. Recently, multimodal machine learning methods have been proposed such as contrastive learning-based methods[14–17], dynamically learning modality gap[18], and hierarchical multimodal metric learning[19]. Although numerous computational methods have been proposed for multi-omic integration, such as MOFA+[20], StabMap[21], scMM[22], Seurat WNN[23], CiteFuse[24], totalVI[25], SNF[26], MEFISTO[27], and multiVI[28], they do not incorporate the spatial coordinates and are therefore not demonstrated the capability to identify spatial domains within tissue samples. Recently, SpatialGlue[29], CellCharter[30], MISO[31], COSMOS[32], SpaMultiVAE[33], and PRESENT[34] were proposed to integrate multi-omics data with spatial information. However, MISO, CellCharter, and SpaMultiVAE integrate multi-omics features merely through simple concatenation or basic operations, failing to effectively capture and utilize the complementary characteristics among different omics layers, which results in suboptimal integration performance. Although COSMOS constructs graph models based on spatial adjacency, such approaches primarily capture local spatial information and may overlook functionally similar regions that are spatially distant. SpatialGlue constructs separate graphs for spatial coordinates and omics features rather than generating a unified graph representation that jointly models both types of information. Moreover, its dual-attention mechanism leads to increased computational complexity with the number of omics modalities and has not demonstrated to support multi-section analysis. PRESENT mainly models specific omics modalities, and its applicability and scalability to emerging omics technologies remain to be further evaluated. In addition, its computationally intensive statistical parameter estimation process further limits its performance. Therefore, there is an urgent need for a computation method that can construct a unified spatial omics graph representation and efficiently process the spatial multi-omics data from both single and multiple tissue sections.

In this work, we propose SMART (Spatial Multi-omic Aggregation method using gRaph neural networks and meTric learning), an unsupervised deep learning framework that leverages graph sampling, aggregation and metric learning for spatial multi-omic integration. SMART constructs a unified graph by incorporating spatial coordinates and omic features. It learns the graph representations using graph neural networks and refines the aggregation process based on the relationships among omic spots through metric learning, thereby integrating multi-omics data with spatial context into a unified latent space. Furthermore, SMART is applicable to datasets at different spatial resolutions and across multiple tissue sections. To validate the effectiveness of SMART, we first conducted experiments on simulated tri-omics data, which demonstrated that SMART theoretically outperforms other methods in both qualitative and quantitative assessments. Subsequently, we applied SMART to jointly detected spatial transcriptomics and proteomics data from human lymph nodes, as well as jointly detected spatial transcriptomics and epigenomics data from mouse brains. These datasets were generated using a variety of spatial multi-omics technologies, encompassing methods such as 10x Genomics Visium CytAssist, MISAR-seq[2], SPOTS[4], Stereo-CITE-seq[6]. SMART demonstrated superior quantitative performance and more accurate spatial domain identification on real-world datasets. Finally, we applied the SMART-MS model, a variant of SMART for Multi-Section analysis, to integrate spatial transcriptomics and proteomics data from mouse spleen and thymus across multiple tissue sections. This approach enabled more precise identification of cell types and anatomical structures within the tissues. These experimental results underscore the advantages of SMART in analyzing spatial multi-omics data.

Results

Overall structure of SMART

SMART is a computational method using graph neural networks combined with metric learning to integrate multiple omics modalities and spatial tissue distribution into a unified latent representation (Fig. 1a). It constructs a graph using spatial coordinates and omic-derived principal components to capture spatial correlations across tissue spots. Some spots may belong to the same cell type or anatomical structure but be spatially distant. To account for this, we applied metric learning with triplet loss to adaptively adjust the latent representation. The input of SMART includes omic matrices from transcriptomics, proteomics, or epigenomics, along with corresponding spatial coordinates indicating locations. SMART aims to derive an integrated representation that can provide a more comprehensive understanding of spatial domains within tissue samples and is applicable to datasets collected by diverse platforms, including 10x Genomics Visium CytAssist, MISAR-seq[2], SPOTS[4] and Stereo-CITE-seq[6].

Instead of analyzing raw or highly variable omic features, we applied principal components, which reflect omic variation and synergy for further analysis, aligning with the biological premise that genes function cooperatively. Moreover, spatial omics data are typically sparse due to limitations of current sequencing technologies. Directly similarity calculations based on raw omic features can be misleading, as zero expression from different spots may be incorrectly interpreted as similar. Therefore, we applied principal component analysis (PCA) to each omics, reducing high-dimensional expression data to lower-dimensional representations that retain biologically meaningful patterns. Its variant, SMART-MS, designed for multi-section multi-omics integration, utilizes PCA and Harmony to remove batch effects, and uses the processed data as input to the model (Fig. 1b).

As cells communicate with their neighbors, each cell within a tissue is influenced by its surrounding environment[35,36]. To model this spatial context, SMART constructs spatial neighbor graphs based on the spatial coordinates of tissue sections through k-nearest-neighbor algorithm (KNN). To integrate omic features with spatial information, we applied graph sampling and aggregation (GraphSAGE[37]) -based encoders to embed both the spatial neighbor graph and the low-dimensional features obtained through PCA. A fully connected network then integrates the embeddings from different omics to generate unified representations, which can reconstruct omic-specific features through GraphSAGE-based decoders.

However, the spatial neighbor graph only captures local spatial context and neglects spots that are of the same type but located distally within the tissue. Therefore, our model assesses the similarity of principal component between spots to identify similar groups. A mutual nearest neighbor algorithm (MNN) is employed to detect pairs of spots that in each other's similar group. These spot pairs are then designated as anchor-positive pairs for metric learning. For single-section data, positive samples are selected from within the same section, whereas for multi-section data, positive samples are drawn from different sections. In both cases, negative samples are consistently selected from within the same section as anchors (Fig. 1c). A triplet loss is constructed to guide the spatial graph by adjusting the representation between spots based on their mutual similarity in principal component space, even when they are spatially distant within the tissue.

SMART is designed to preserve the original omics-specific features in the latent space while maintaining similarity between spots and incorporating spatial context. Despite its simplicity, SMART is effective for integrating datasets containing two or three omics layers and is theoretically scalable to both single-omics and more complex multi-omics scenarios. Moreover, SMART supports the integration of spatial multi-omics data across different technologies, resolutions, and tissue sections (Fig. 1d). To validate the design of SMART, we

https://doi.org/10.1038/s41467-026-70821-5

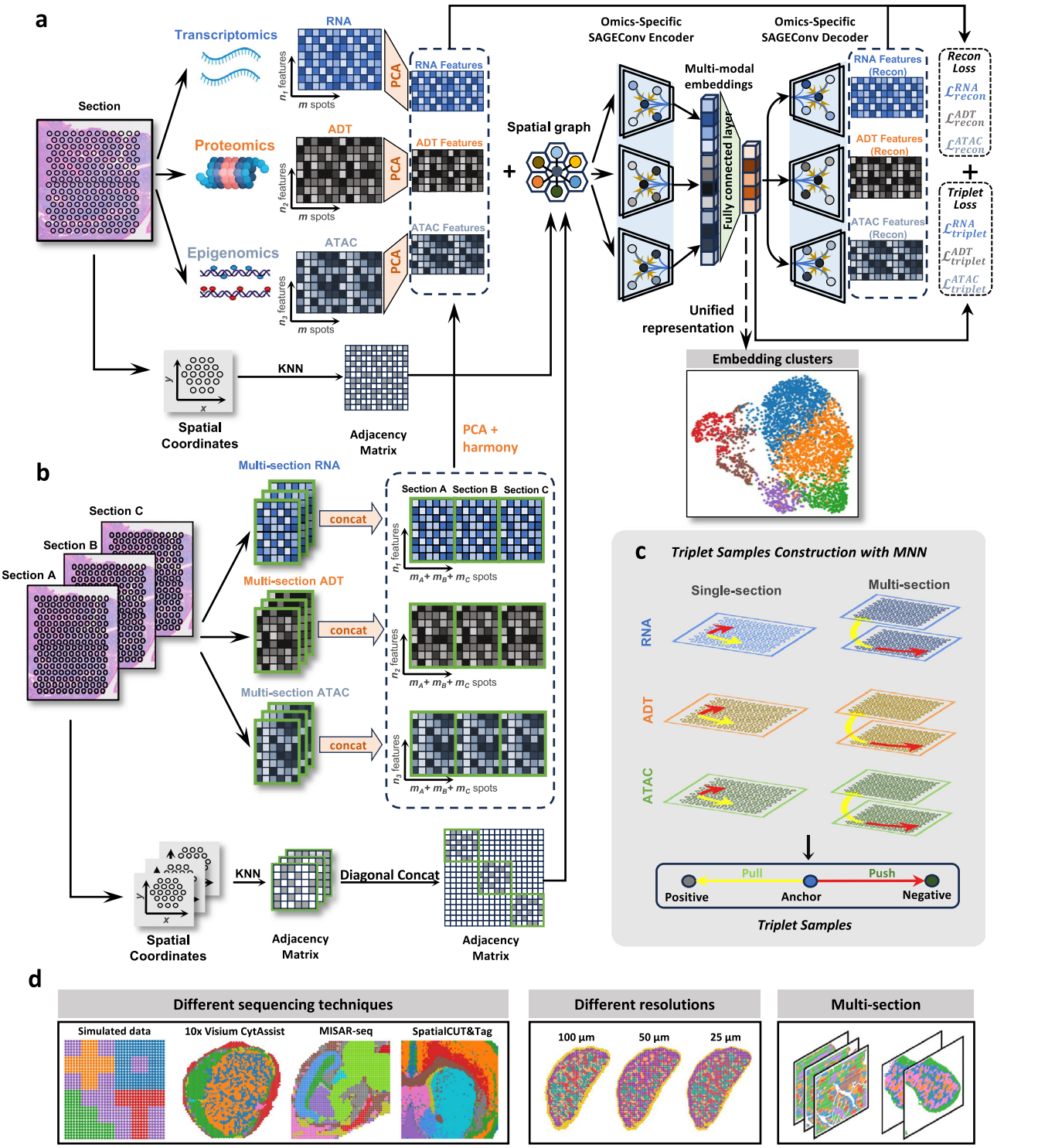

Fig. 1 | Framework of SMART designed for spatial multi-omics integration. **a** SMART overview. The SMART framework constructs a spatial neighbor graph from spatial coordinates using k-nearest neighbors (KNN). For each modality, normalized features are dimensionally reduced using principal component analysis (PCA) and used as model inputs. Anchor, positive, and negative samples are determined via mutual nearest neighbor (MNN) matching for metric learning. SAGEConv encoder processes the spatial graph and input features by sampling and aggregating neighborhood information to generate node embeddings. Multi-modal embeddings are concatenated and integrated through a dense layer to produce unified representations. The model is optimized by jointly minimizing feature reconstruction (Recon) loss and triplet margin loss in the SAGEConv decoder. The transcriptomics, proteomics, and epigenomics icons were adapted from BioRender.com and are licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). **b** SMART-MS overview. The SMART-MS framework integrates multi-section omics data through horizontal concatenation of expression matrices across tissue sections, followed by joint dimensionally reduction and batch correction with Harmony to generate a unified input matrix. For spatial graph construction, the model employs diagonal concatenation of spatial graphs from individual sections. The remaining pipeline maintains consistency with the standard SMART framework. **c** Triplet samples construction. Single-section triplets use intra-section sampling, while multi-section triplets select positives from other sections and negatives from the anchor's section with MNN. **d** Downstream applications of SMARTS and SMART-MS for spatial multi-omics data across technologies, resolutions, and multi-section analyses.

conducted a series of ablation studies, evaluating different graph neural network encoders (e.g., GCN[38], GAT[39], GraphConv[40]), strategies for selecting anchor and positive pairs in metric learning, alternative loss functions and other architectural components (Supplementary Table 1 and Supplementary Figs. 1–4).

Moreover, SMART allows for tuning the number of neighbors K in the spatial graph construction and the weight of the triplet loss to bias the embedding toward either cell type or spatial domain preferences (Supplementary Fig. 5). SMART also supports multi-omics integration in scenarios where spots are partially missing within a single section (Supplementary Fig. 6a–e). These experiments demonstrate the robustness and flexibility of the SMART framework.

Benchmarking SMART's performance using simulated data

We initially evaluated the performance of SMART using simulated spatial multi-omics data with known ground truth, generated following the method proposed by Townes et al.[41]. The ground truth of the simulated data defines five distinct spatial factors: factors 1 through 4, representing different cell types, and a background category (Fig. 2a). The simulation includes three distinct modalities, RNA, ADT, and ATAC, which exhibit markedly different expression distributions, reflecting the complementary nature of multi-omics data (Fig. 2b). We compared SMART against several representative methods for multi-omics integration at a clustering resolution of 5 clusters, including MOFA + [20], MEFISTO[27], SpatialGlue[29], SNF[26], CellCharter[30], MISO[31], Seurat WNN[23], PRESENT[34], COSMOS[32] and SpaMultiVAE[33]. Among existing approaches, these are currently the only methods that support the integration of three omics modalities.

Principal components extracted from each individual omic partially revealed 3–5 distinct factors but failed to clearly distinguish all spatial regions (Fig. 2c), highlighting the necessity of integrating complementary multimodal and spatial information. Although MOFA +, MISO, SNF, PRESENT and SpaMultiVAE can integrate multiple omics, they failed to resolve certain regions. SpatialGlue, MEFISTO and COSMOS successfully identified all factors but struggled with precise boundary delineation between spatial regions, whereas WNN introduced substantial noise into the segmentation results. In contrast, SMART and CellCharter not only accurately identified all factors but also closely matched the ground truth, demonstrating superior performance in spatial domain identification (Fig. 2c, d, Supplementary Fig. 6f).

Integration methods aim to generate a unified embedding (representation) of multi-omics data that preserves the original relationships within each omic. To assess how well these relationships are maintained, we evaluated the consistency between the pairwise distances among spots in the original omics and their corresponding distances in the embedded space. Specifically, we calculated the Euclidean distances for all pairs of spots before and after embedding, and based on these, computed the Pearson correlation coefficient to assess the similarity between the original features and the embedded representations. SMART and MOFA+ demonstrated the highest consistency between the original and embedded representations across all modalities, particularly with RNA and ADT (Fig. 2e). These results indicate that both methods maximally preserved the original features of each omics modality during integration, while SMART further incorporated spatial information throughout this process. Within the clusters identified by SMART, we also demonstrated the correlation between each cluster and the individual omics modalities. This analysis reveals that certain clusters tend to be more associated with specific omics features, reflecting the complementary integration of modalities achieved by SMART (Fig. 2f). Moreover, since cells of the same type are often spatially clustered, such as clone populations of T cells[42], we evaluated the overall spatial coherence of the embeddings using Moran's *I*, a measure of spatial autocorrelation. A higher Moran's *I* value indicates a more orderly spatial distribution. However, for certain datasets with complex spatial domain patterns, Moran's I may be relatively low. SMART achieved Moran's I scores approaching 1, demonstrating excellent spatial clustering performance (Fig. 2g).

To quantitatively compare the performance of SMART with existing methods, we evaluated clustering accuracy using several standard metrics: Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Homogeneity (Homo), Mutual Information (MI), FMI (Fowlkes-Mallows Index) and V-Measure. Across all metrics, SMART consistently achieved the highest performance, followed by CellCharter. In contrast, MISO and SNF did not produce comparable results (Fig. 2h).

SMART is a robust model applicable not only to multi-omics data but also to single-omics modalities. We evaluated its performance separately on RNA, ADT, and ATAC data, as well as their combinations, using various metrics (Fig. 2i). Repeated experiments conducted ten times with different random seeds demonstrated that SMART achieves strong performance even when utilizing only one modality, though slightly lower than results obtained with three modalities. To investigate differences in the embeddings produced by SMART between single-omics and multi-omics data, we compared the k-nearest neighbor graphs derived from SMART embeddings in both scenarios. Specifically, we computed embeddings using SMART for each individual modality (SMART-ADT, SMART-RNA, SMART-ATAC) as well as for the integration of all three modalities (SMART-ADT + RNA + ATAC). For each embedding, we constructed a graph using the k-nearest neighbors ($k = 10$) and recorded all edges. By comparing the edges across these graphs, we found that only 12.1% of the edges in the three-omics embedding were unique, which is lower than the proportion of unique edges found in any single-omics embedding (ATAC: 21.2%; ADT: 17.6%; RNA: 16.1%) (Fig. 2j). This indicates that the tri-modal embedding captures complementary features by integrating information from all modalities.

SMART integrates spatial transcriptomic and proteomic data

To evaluate the effectiveness of SMART on real-world data, we applied SMART and several comparison methods to a co-detected transcriptome and proteome dataset from human lymph node section A1, generated using the 10x Visium Spatial Platform. Hematoxylin and eosin (H&E)-based annotations were used as the ground truth (Fig. 3a). The outer region of the tissue is primarily composed of pericapsular adipose tissue and the capsule, while the inner region includes the cortex, medullary sinuses, and medullary cords. In addition to the previously compared methods, we also included scMM[22], TotalVI[25], MultiVI[28], COSMOS[32] and SpaMultiVAE[33] for RNA-protein integration, resulting in a total of thirteen comparison methods used to assess the performance of SMART.

The results with a cluster number of 7 demonstrate that SMART not only successfully identified the cortex (cluster 2) and pericapsular adipose tissue (cluster 3), but also more accurately delineated the medullary sinuses (cluster 1) and medullary cords (cluster 5) compared to existing methods (Fig. 3b). In addition, SMART detected smaller structures such as follicles (cluster 6) and the capsule (cluster 4,7) (Fig. 3b, Supplementary Fig. 7). While most other methods were able to reliably identify the cortex, medullary sinuses, medullary cords and pericapsular adipose region, MEFISTO, SNF, COSMOS and CellCharter fail to separate the medullary sinus from the medullary cords due to over-integration of spatial information. In the UMAP (Uniform Manifold Approximation and Projection) visualization, SMART, PRESENT, MISO and SpatialGlue exhibit well-defined cluster boundaries corresponding to the ground truth labels, demonstrating their effectiveness in identifying distinct tissue regions (Fig. 3c).

We further compared the performance of different methods using supervised evaluation metrics, and the results show that SMART demonstrates a significant advantage in quantitative assessments when the number of clusters is set to 7 (Fig. 3d). Considering that the

https://doi.org/10.1038/s41467-026-70821-5

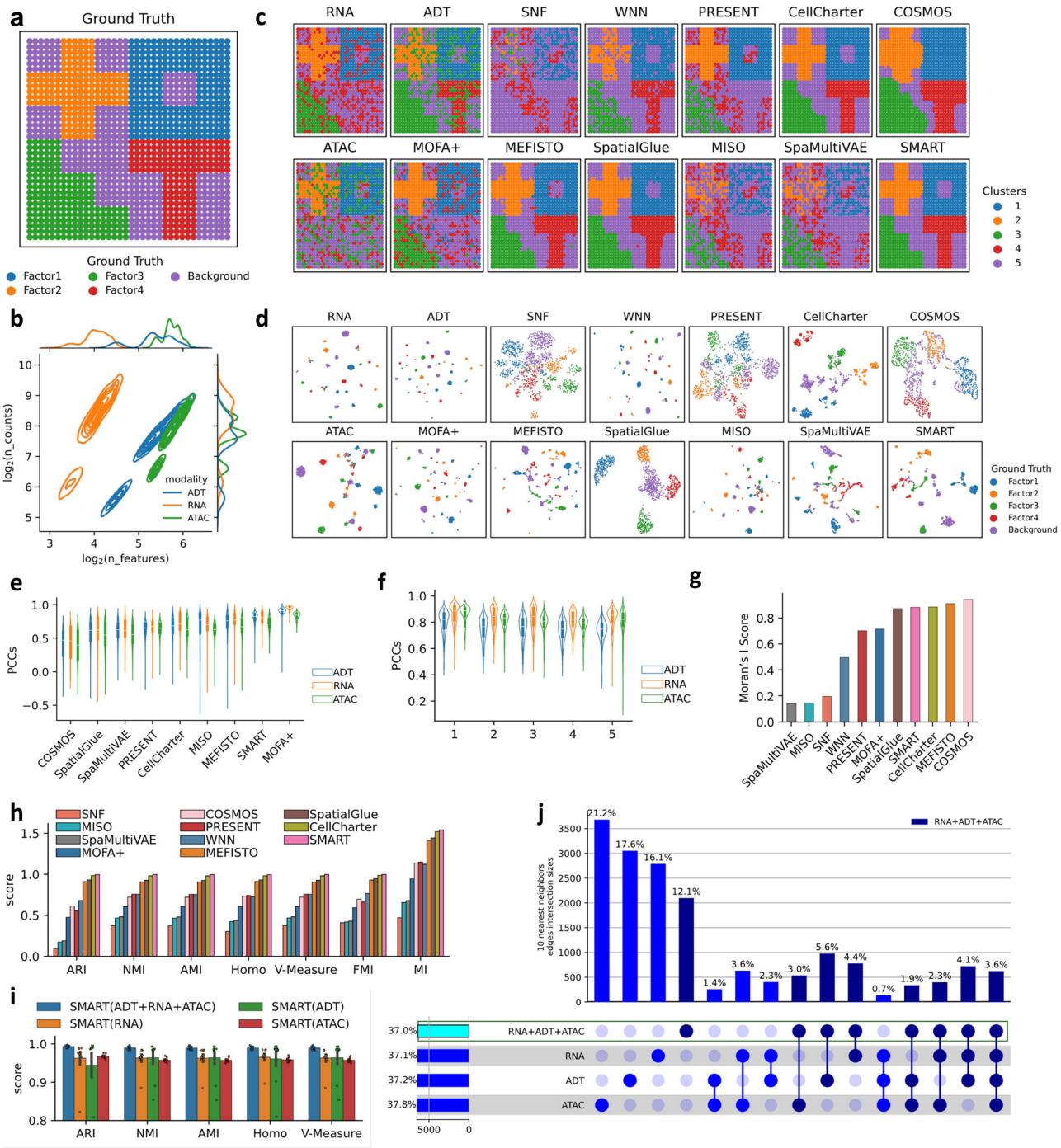

Fig. 2 | Integration of simulated multi-omics data using SMART for accurate spatial region identification. a Ground truth spatial domains of the simulated data. **b** Joint distribution map of expression data across the three omics. **c** Clustering results based on PCA features from RNA, ADT, and ATAC, compared with spatial multi omics integration methods. **d** Umap plots with ground truth based on PCA features from RNA, ADT, and ATAC, compared with spatial multi omics integration methods. **e** Violin plots of Pearson correlation coefficients (PCCs) between distance matrices derived from the integrated embeddings of different methods and those computed from individual omics (n = number of independent distance matrix correlation computations for each method and modality). The central white dot represents the median, the vertical bar represents the interquartile range (25th–75th percentiles), the violin bounds represent the minima and maxima, and the curvature represents the kernel density estimate of the data distribution. **f** Violin plots of PCCs between the distance matrix derived from SMART's integrated embedding and those computed from individual omics across different clusters (n = number of independent distance matrix correlation computations for each cluster and modality). The central white dot represents the median, the vertical bar represents the interquartile range (25th–75th percentiles), the violin bounds represent the minima and maxima, and the curvature represents the kernel density estimate of the data distribution. **g** Bar plot illustrating Moran's I scores for different methods. **h** Bar plot displaying seven evaluation metrics to assess the performance of different methods. **i** Bar plots of five evaluation metrics evaluating the performance of different omics combinations across ten replicate experiments. Data are presented as mean ± standard error of the mean (n = 10 independent replicate experiments). **j** UpSet plot showing the edge intersections of nearest-neighbor graphs constructed from embeddings generated by SMART-uniomics and SMART-triomics. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-70821-5

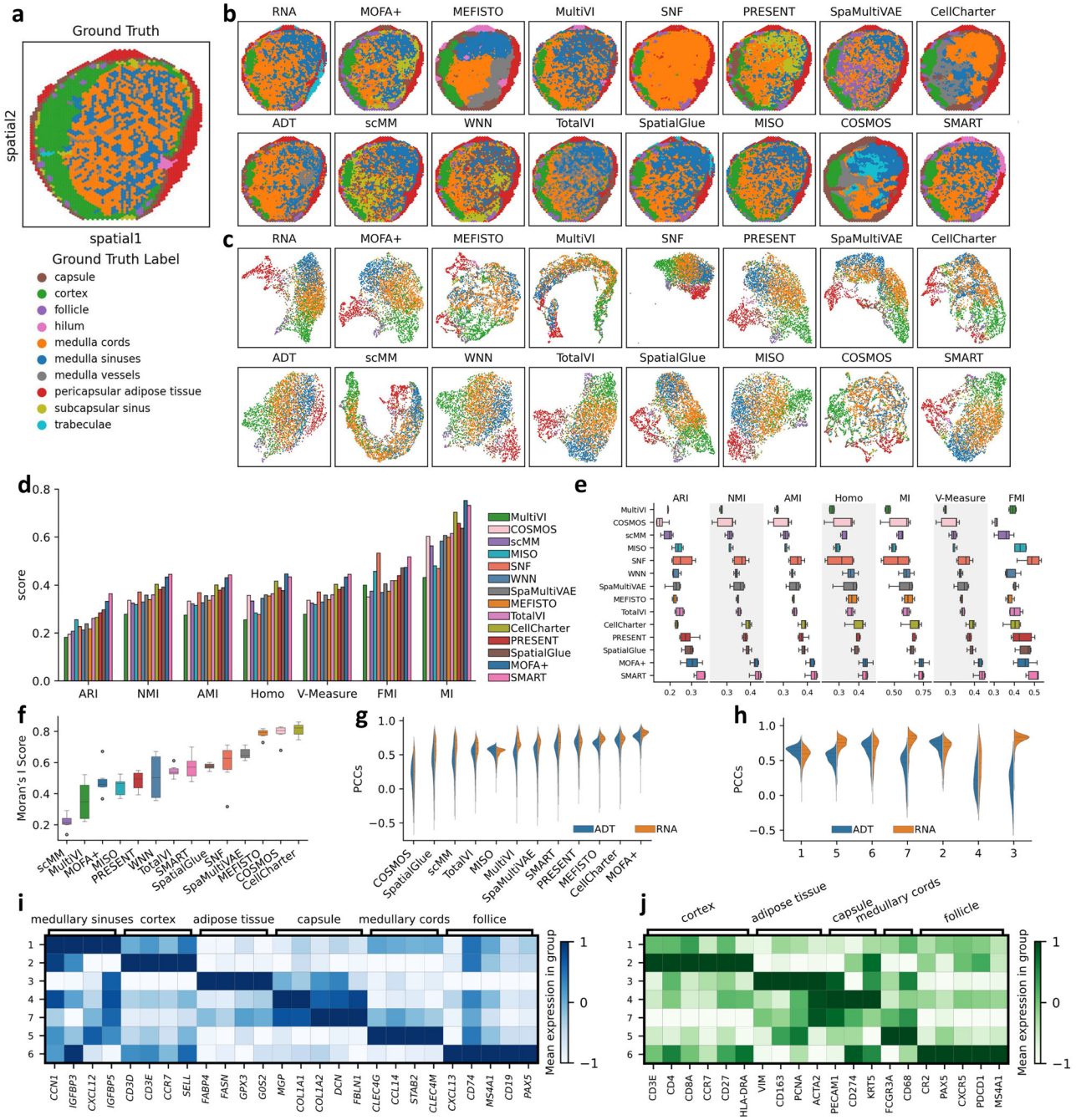

Fig. 3 | Integration of 10X Visium CytAssist Gene and Protein co-profiling multi-omics data human lymph node section A1 using SMART. a Ground truth spatial domains of human lymph node section A1. **b** Clustering results based on PCA features of RNA and ADT modalities, compared with various spatial multi-omics integration methods. **c** UMAP visualizations colored by ground truth, based on PCA features from RNA and ADT, compared with results from different integration methods. **d** Bar plot showing seven evaluation metrics for different methods when the number of clusters is fixed at 7. **e** Box plots of the seven evaluation metrics across different methods, with the number of clusters varying from 5 to 10. In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, whiskers denote 1.5 times the interquartile range ($n = 6$ clustering results). **f** Box plots of Moran's I scores for different methods, with cluster numbers ranging from 5 to 10. In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, whiskers denote 1.5 times the interquartile range, and data points beyond the whiskers represent outliers ($n = 6$ clustering results). **g** Pearson correlation coefficients (PCCs) between distance matrices computed from integrated embeddings of each method and those derived from individual omics. The plot shows the distribution of PCCs (n = number of independent distance matrix correlation computations for each method and modality). **h** Pearson correlation coefficients (PCCs) between the distance matrix computed from SMART's integrated embedding and those derived from each omic across varying cluster numbers. The plot shows the distribution of PCCs (n = number of independent distance matrix correlation computations for each cluster number and modality). **i** RNA expression of region-specific markers in SMART's clusters. **j** Protein expression of region-specific markers in SMART's clusters. Source data are provided as a Source Data file.

optimal number of clusters may vary across datasets, we additionally evaluated the stability of SMART across a range of clustering resolutions (from 5 to 10 clusters). The results indicate that SMART consistently outperforms other methods under all settings, highlighting its robustness across different clustering granularities (Fig. 3e). In the spatial autocorrelation evaluation, SMART exhibits a moderate performance, as its clustering results closely align with the ground truth labels but do not display particularly pronounced spatial patterns (Fig. 3f). Additionally, for each cluster, we calculated the Pearson correlation coefficients between the original transcriptomic or proteomic features and the unified embeddings generated by SMART by comparing the pairwise relationship among spots before and after embeddings. The results show that SMART exhibits a high degree of consistency between the original and embedded representations (Fig. 3g). Specifically, SMART placed greater emphasis on protein features in clusters 1 and 2, while RNA features contributed more prominently in clusters 3, 4, 5, 6 and 7 (Fig. 3h).

To validate the accuracy of the spatial domains identified by SMART, we extracted spatially specific marker genes and proteins for selected regions (Fig. 3h, i). For example, *CD3E* and *CCR7* show high expression in both RNA and ADT modalities within the cortex (Cluster 3), indicating that this region harbors a higher abundance of T cells compared to other spatial domains (Supplementary Fig. 8). Similarly, in the ADT modality, the medullary sinus (Cluster 1) may have been grouped as a distinct class despite the absence of known spatially specific marker proteins, but it can be reliably annotated based on the expression of spatially enriched genes such as *CCN1*. These findings demonstrate that SMART effectively leverages the complementary strengths of transcriptomic and proteomic data to identify spatial domains with high biological relevance.

We further evaluated SMART on an additional sample, human lymph node section D1, and observed that it consistently outperformed other methods. These results demonstrate SMART's strong generalization capability in integrating transcriptomic and proteomic data (Supplementary Figs. 9–11).

SMART integrates spatial transcriptomic and chromatin accessibility data

In addition to integrating spatial transcriptomic and proteomic data, SMART can also fuse spatial transcriptomics with chromatin accessibility. We evaluated SMART's performance on brain datasets from four developing mice at distinct embryonic stages (E11.0, E13.5, E15.5, and E18.5)[2]. These datasets were generated using MISAR-seq, a technique that enables spatially resolved joint profiling of chromatin accessibility and gene expression.

Since TotalVI and SpaMultiVAE is not applicable to epigenomic data, we compared SMART with other representative methods, including scMM, WNN, SNF, MultiVI, PRESENT, MOFA+, MEFISTO, MISO, COSMOS, CellCharter and SpatialGlue. We used the unsupervised clustering results from Jiang et al.[2] as artificial annotations, which closely match the anatomical structures observed in H&E-stained images of brain at different developmental stages. On the E18.5_S1 brain section, the clustering results for ten regions showed that only WNN, SMART and SpatialGlue produced clear boundaries between different brain areas. Both methods successfully delineated regions such as dorsal pallium (DPallm), portions of the diencephalon, and hindbrain. while these region in other methods were disperse (Fig. 4a, c). However, SMART additionally detected the boundary of a small structure, cartilaga-1, which was not as smoothly defined by SpatialGlue (Fig. 4c). In the manually annotated UMAP visualizations of different methods, the embeddings produced by SMART, WNN, and SpatialGlue exhibited relatively clear boundaries between distinct categories (Fig. 4b, c).

In terms of unsupervised evaluation, SMART better preserved the original multi-omic relationships among spatial spots and ranked third in spatial autocorrelation assessment (Fig. 4d), indicating its effectiveness in identifying continuous spatial domains within the brain. Moreover, we computed the correlations between the original transcriptomic or epigenomic features and the embeddings generated by SMART. The results showed that SMART embeddings exhibited strong correlations with the original features (Fig. 4e), and each SMART cluster tended to align more closely with one specific omic modality (Fig. 4f). In the supervised evaluation using quantitative metrics, SMART also achieved the best performance on the E18.5_S1 section (Fig. 4g).

We further evaluated SMART on the remaining three sections: E11.0_S1, E13.5_S1, and E15.5_S1, which contain 5, 7, and 11 clusters, respectively. The clustering results indicate that, compared with other methods, SMART more effectively reconstructs brain structure across developmental stages. For example, in E15.5_S1, SMART accurately distinguishes the dorsal pallium medial (DPallm) and ventral (DPallv) regions, while in E11.0_S1, it clearly identifies the primary brain and mesenchyme. Moreover, SMART outperformed most methods in terms of supervised evaluation metrics, except on E11.0_S1 where it was surpassed by MOFA+ (Fig. 4c, Supplementary Figs. 12–14).

To quantitatively compare the overall performance, we computed supervised clustering metrics for the results of all spatial multi-omics integration methods across the four sections. SMART consistently achieved the highest scores, highlighting its effectiveness in identifying spatial domains and integrating spatial transcriptomic and chromatin accessibility data (Fig. 4h). These findings further confirm the robustness and superiority of SMART in analyzing spatial RNA and ATAC-seq data.

SMART enables efficient integration of spatial multi-omics data

SMART is an efficient model with fewer parameters and lower computational complexity, enabling the analysis of large-scale datasets. To further demonstrate SMART's advantages in computational efficiency and resource utilization for multi-omics integration, we evaluated its performance on one of the largest currently available spatial transcriptomics and epigenomics datasets. This dataset, obtained by CUT&Tag–RNA-seq, co-profiles transcriptomic data and H3K27me3 epigenetic marks in the P22 mouse brain, encompassing 9752 spots, 25,881 genes, and 70,470 peaks.

As the dataset lacks ground-truth labels, we determined the number of clusters to be 12 based on internal clustering validation metrics, including mean FMI, SCI (Silhouette Coefficient), and DBI (Davies-Bouldin Index) (Supplementary Fig. 15a, b). While all methods successfully identified major anatomical layers such as the cerebral cortex (CTX), caudoputamen (CP), and genu of the corpus callosum (ccg), SMART, COSMS, and CellCharter's integration results exhibited fewer noise artifacts and more clearly defined layer boundaries (Fig. 5a and Supplementary Fig. 15c–e). While CellCharter, MEFISTO, and COSMOS also showed relatively distinct cluster boundaries and achieved high Moran's I scores for spatial autocorrelation, they tended to overemphasize spatial smoothness at the expense of preserving true anatomical structures (Supplementary Fig. 15f). In comparative analyses of runtime and memory usage, SMART consistently demonstrated the shortest runtime and lowest memory consumption, whereas MEFISTO required the most computational resources (Fig. 5b).

To further validate the efficiency of SMART, we applied it to a large-scale spatial transcriptomics and proteomics co-profiling dataset, the Stereo-CITE-seq dataset of the mouse spleen. Stereo-CITE-seq is a multi-omics technology that integrates spatial transcriptomics (Stereo-seq) with spatial proteomics (CITE-seq), achieving subcellular spatial resolution of up to 500 nanometers. It enables simultaneous measurement of both mRNA and protein expression levels on a single tissue section. Given the high resolution (500 nm) of the raw spatial data, neighboring detection points are commonly aggregated into

https://doi.org/10.1038/s41467-026-70821-5

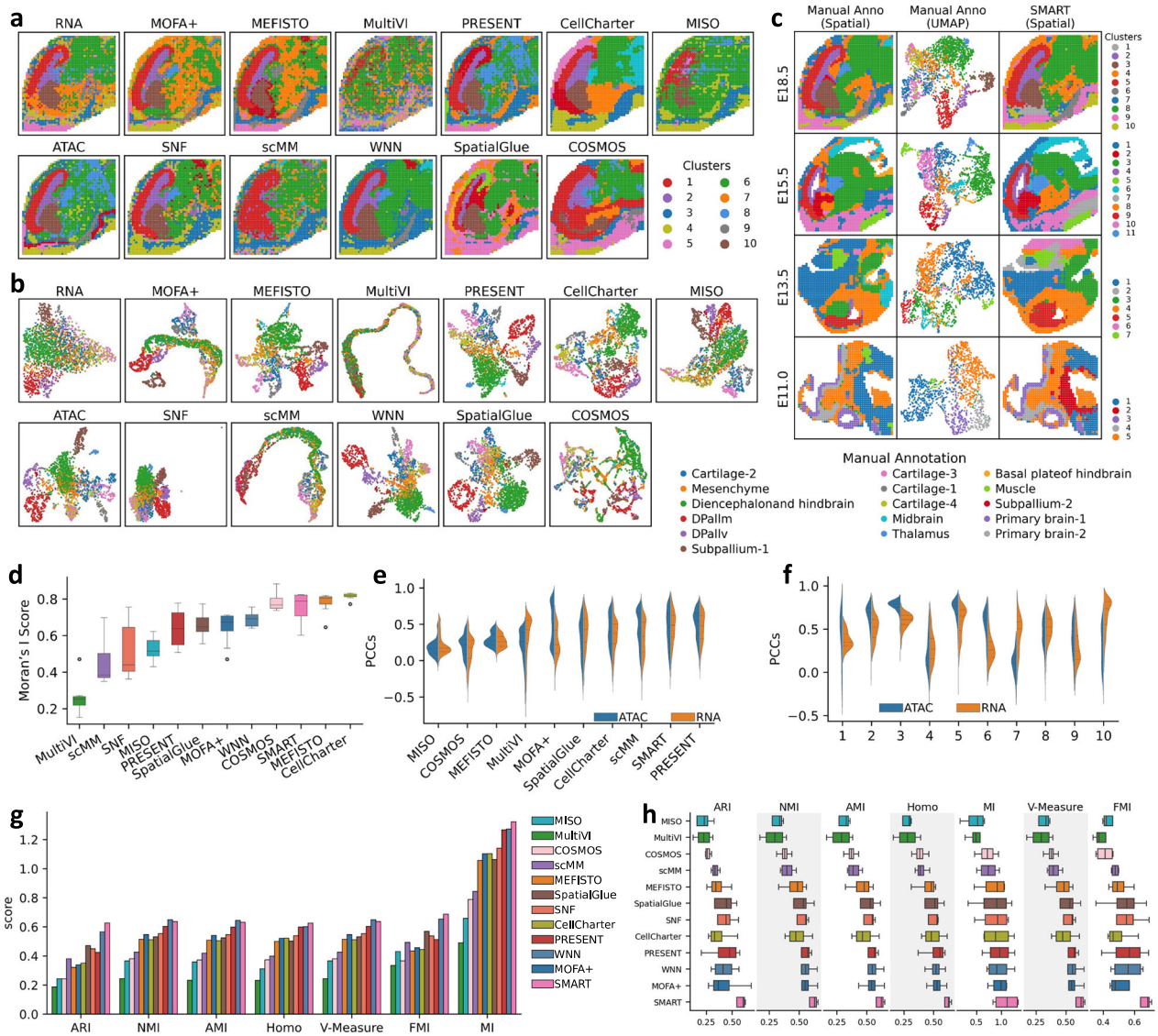

Fig. 4 | SMART integrates MISAR-seq multi-omics data and accurately identifies spatial regions. a Clustering results based on PCA features of RNA and ATAC modalities, compared with other spatial multi-omics integration methods. **b** UMAP visualizations colored by manual annotation, based on PCA features from RNA and ATAC, compared with results from other different integration methods. **c** Manual annotation, UMAP and clustering result integrated by SMART of four sections. **d** Box plots of Moran's I scores for different methods, with cluster numbers ranging from 7 to 14. In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, whiskers denote 1.5 times the interquartile range, and data points beyond the whiskers represent outliers ($n = 8$ clustering results). **e** Pearson correlation coefficients (PCCs) between distance matrices computed from integrated embeddings of each method and those derived from individual omics. The plot shows the distribution of PCCs (n = number of independent distance matrix correlation computations for each method and modality). **f** Pearson correlation coefficients (PCCs) between the distance matrix computed from SMART's integrated embedding and those derived from each omics across varying cluster numbers. The plot shows the distribution of PCCs (n = number of independent distance matrix correlation computations for each cluster number and modality). **g** Bar plot showing seven evaluation metrics for different methods when the number of clusters is fixed at 8. **h** Box plot of seven evaluation metrics to evaluate the performance of combinations of different omics with four sections (E11.0_S1, E13.5_S1, E15.5_S1, E18.5_S1). In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, whiskers denote 1.5 times the interquartile range ($n = 4$ independent datasets). Source data are provided as a Source Data file.

larger spatial units, or bins, to facilitate analysis and improve the signal-to-noise ratio. For instance, Bin10 denotes the aggregation of 10 × 10 neighboring detection points into one unit, covering an area of 5 μm × 5 μm for downstream analysis.

We extracted spatial grid data at six resolution levels from the raw dataset (Bin200 to Bin10), with spatial resolution increasing progressively from 100 μm (Bin200) to 5 μm (Bin10). Correspondingly, the number of spatial locations (spots) increased from 2001 to 756,430. The dataset includes a total of 29,034 genes and 128 ADTs. Considering that most state-of-the-art spatial multi-omics integration methods support GPU acceleration, we compared their GPU memory consumption and training time under a unified GPU environment to evaluate scalability and computational efficiency.

For consistency, we set the number of clusters to 6 across all analyses. SMART maintained strong spatial continuity in its clustering results across different resolutions, with several spatial structures stably identified across adjacent bin sizes. In contrast, clustering results based on RNA or ADT modalities alone showed instability at certain resolutions, with ambiguous cluster boundaries (Fig. 5c). Regarding resource utilization, aside from SpaMultiVAE (which was configured with a batch size of 2000), all methods used a batch size equal to the number of spots, leading to GPU memory usage being

https://doi.org/10.1038/s41467-026-70821-5

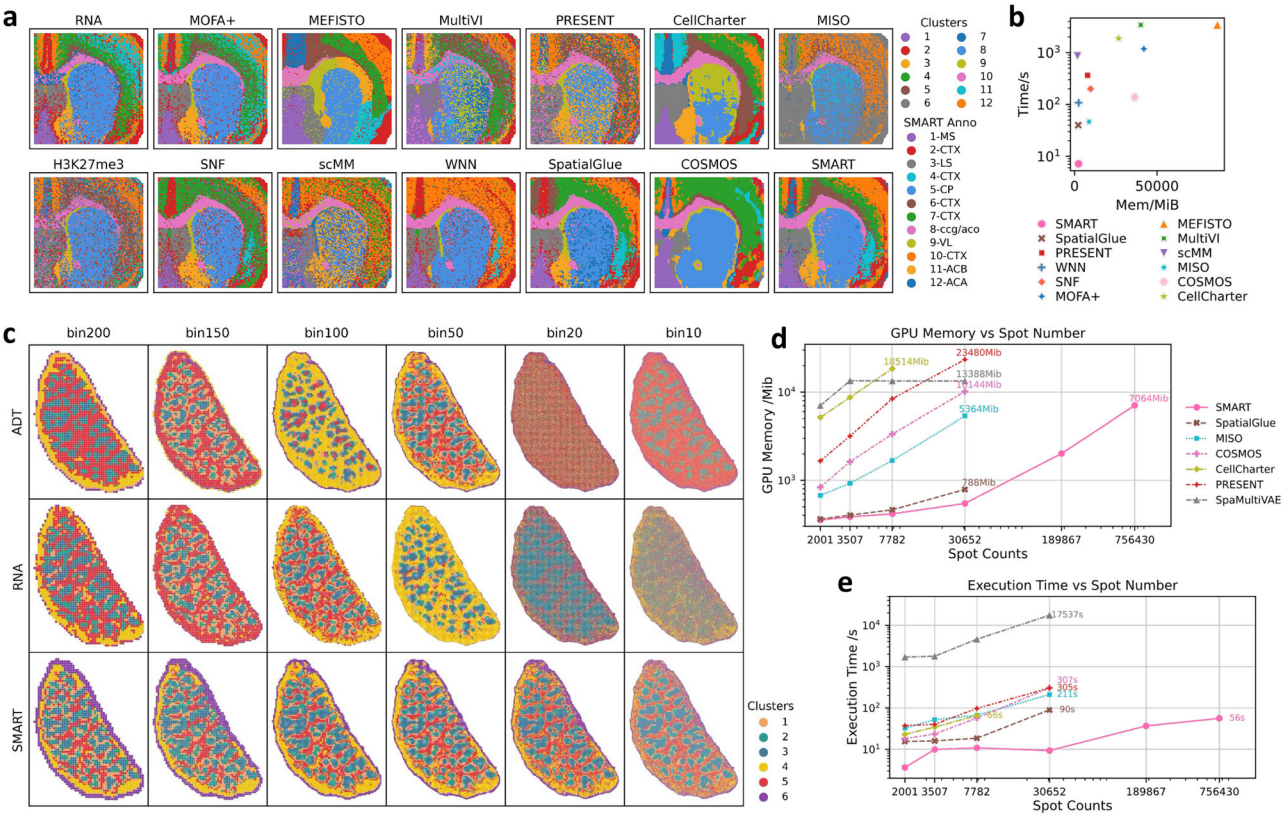

Fig. 5 | SMART efficiently integrates large-scale multi-omics datasets. **a** Clustering results of the P22 mouse brain section using PCA features from RNA or H3K27me3, compared with results from different spatial multi-omics integration methods. **b** Runtime and peak memory usage of various methods on the P22 mouse brain dataset. **c** Clustering results of the mouse spleen at multiple spatial resolutions using PCA features of RNA, ADT, and SMART's integrated embedding. **d** GPU memory usage of different methods across spatial resolutions in the mouse spleen dataset. **e** Runtime comparison of different methods across spatial resolutions in the mouse spleen dataset. Source data are provided as a Source Data file.

directly proportional to the number of spots. Under the same data scale, SMART exhibited the lowest memory usage. Notably, all other methods failed to complete computations on the Bin20 dataset with about 190 K spots, whereas SMART ran smoothly even on the largest Bin10 dataset with over 750 K spots (Fig. 5d). In terms of training time, SMART completed training on the Bin10 dataset with over 750 K spots in only 56 seconds, significantly outperforming all other methods (Fig. 5e). This is attributed to the rapid convergence of the SMART model and the high efficiency of the SAGEConv architecture when handling large-scale graph structures.

In summary, SMART not only demonstrates strong adaptability and stability in spatial multi-omics integration, but also significantly outperforms existing methods in terms of computational efficiency and resource usage, highlighting its broad potential for large-scale spatial data analysis.

SMART enables the integration of multi-omics data across multiple sections

Spatial omics data from biological tissue samples are often derived from multiple tissue sections. To enable multi-omics integration across sections, we developed SMART-MS, an extension of SMART that supports multi-section data integration, batch effect correction, and the construction of cross-section spatial neighbor graphs to align omics features across tissue sections.

We first evaluated the performance of SMART-MS and other methods on a human tonsil dataset. As a critical component of the peripheral lymphoid system, the tonsil plays a central role in mucosal immune defense. It exhibits complex internal architecture and pronounced spatial heterogeneity, characterized by densely arranged lymphocyte clusters. The follicular zone is enriched in B cells and often contains germinal centers; the surrounding mantle zone also consists primarily of B cells; and the interfollicular zone between follicles is a T cell–rich region. The outer regions include epithelial and connective tissue areas, which may contain non-lymphoid cell types such as plasma cells and epithelial-related cells.

The dataset consists of three tissue sections co-profiled with spatial transcriptomics (RNA) and proteomics (ADT) using the 10x Genomics CytAssist Visium platform. UMAP visualization of PCA-reduced features revealed a clear batch effect in the third section, likely due to the fact that the first two are consecutive sections and thus more similar (Supplementary Fig. 16a). In addition, major tissue regions were annotated based on H&E-stained images and used as ground truth for computing supervised metrics.

We compared SMART-MS with several integration methods capable of batch effect removal, including MultiVI, Present-BC, TotalVI, and MOFA+, as well as RNA- or ADT-derived features corrected by Harmony. When setting the number of clusters to 6, we observed that all methods except TotalVI achieved good batch effect removal (Supplementary Fig. 16b). Moreover, while all methods successfully identified the germinal center, MultiVI poorly distinguished the follicular region, and Present-BC and TotalVI failed to accurately detect the epithelial and connective tissue zones. In contrast, SMART-MS accurately identified the follicular and germinal center regions, as well as the outer tonsillar plasma cell/epithelial zones and the interfollicular T cell–enriched regions (Fig. 6a, Supplementary Fig. 16c). The utility of SMART-MS clustering was further validated by the co-expression of

https://doi.org/10.1038/s41467-026-70821-5

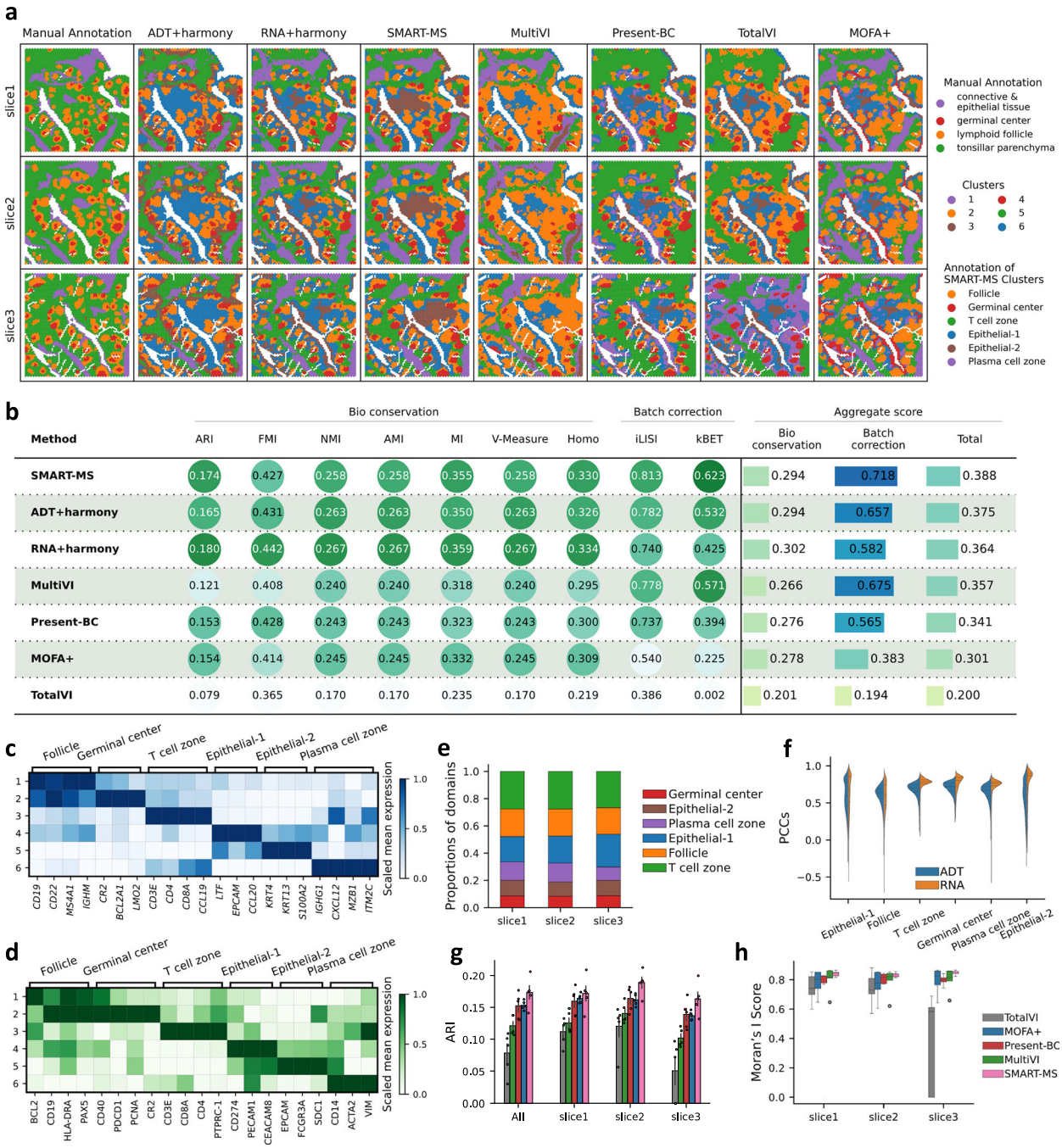

Fig. 6 | SMART-MS integrates multi-omics and multi-section data while effectively removing batch effects. a Clustering results of three human tonsil slices using PCA features of RNA or ADT after Harmony correction, compared with results from different spatial multi-omics integration methods. **b** Biological conservation scores and batch correction scores across different integration methods. Biological conservation performance was evaluated using ARI (adjusted Rand index), FMI (Fowlkes–Mallows index), NMI (normalized mutual information), AMI (adjusted mutual information), MI (mutual information), V-measure, and Homo (homogeneity). Batch correction performance was assessed using iLISI (integration local inverse Simpson's index) and kBET (k-nearest neighbor batch effect test). **c** RNA expression of region-specific marker genes in clusters identified by SMART-MS. **d** Protein expression of region-specific marker proteins in clusters identified by SMART-MS. **e** Bar plot showing the proportion of spatial domains across the three slices. **f** Pearson correlation coefficients (PCCs) between the distance matrix derived from SMART-MS's integrated embedding and those derived from each individual omic across different clusters. The plot shows the distribution of PCCs (n = number of independent distance matrix correlation computations for each cluster and modality). **g** Bar plots of ARI scores on the entire dataset, as well as separately on slice 1, slice 2, and slice 3, with cluster numbers ranging from 4 to 8. Data are presented as mean ± standard error of the mean ($n = 5$ clustering results). **h** Box plots of Moran's I scores on slice 1, slice 2, and slice 3 for different methods, with cluster numbers ranging from 4 to 8. In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, whiskers denote 1.5 times the interquartile range, and data points beyond the whiskers represent outliers ($n = 5$ clustering results). Source data are provided as a Source Data file.

marker genes and proteins, such as *CD19* in the follicular zone, *CR2* in the germinal center, and *CD3E* and *CD4* in T cell–rich areas, across both RNA and ADT modalities (Fig. 6c, d). Furthermore, the proportions of spatial domains were highly consistent across sections (Fig. 6e), and analysis of SMART-MS embeddings revealed modality-specific preferences for each cluster (Fig. 6f).

For quantitative evaluation, we used previously defined supervised metrics for biological signal preservation, and iLISI and kBET scores to assess batch effect removal. SMART-MS achieved the highest overall score among all compared methods and substantially outperformed others in both biological information retention and batch effect correction (Fig. 6b). To control for potential bias from fixed clustering numbers, we also evaluated clustering performance across a range of cluster numbers (4–8). SMART-MS consistently achieved the highest ARI scores both across the three sections collectively and for each individual section (Fig. 6g). In terms of spatial autocorrelation, SMART-MS also attained the highest Moran's I score (Fig. 6h). These results demonstrate that SMART-MS effectively integrates multi-section data and accurately identifies spatial domains.

To further demonstrate the generalizability of SMART, we conducted additional experiments on two mouse spleen sections generated by SPOTS, three high-resolution mouse thymus sections from Stereo-CITE-seq, and three mouse postnatal brain sections from spatial ATAC-RNA-seq. Although these datasets lack manual annotations to serve as ground-truth labels, SMART-MS was still able to identify distinct spatial structures, for example by distinguishing cell types in the mouse spleen (Supplementary Fig. 17). The similarity metric used for comparing the two sections is cosine similarity. Moreover, it consistently outperformed other methods in terms of batch effect removal and spatial autocorrelation (Supplementary Figs. 18–19).

Discussion

SMART is an unsupervised deep learning model based on graph neural networks and metric learning, designed to integrate multiple omics and spatial coordinate information into unified latent representations. These embeddings are optimized to preserve both the modality-specific features of each spot and the spatial similarity between spots, enabling effective spatial multi-omics integration. We evaluated SMART on simulated tri-omics datasets and real dual-omics datasets, combining transcriptomics with either proteomics or chromatin accessibility, generated using established techniques. Benchmarking results demonstrated that SMART outperformed seven non-spatial and two spatial multi-omics integration methods across a range of metrics (Supplementary Table 2). To extend its capability to serial sections, we developed SMART-MS, which reconstructs a more complete molecular atlas by incorporating spatial context beyond individual tissue sections, thereby enhancing both resolution and coverage. The datasets used in this study span five platform technologies, include between 1000 and nearly 20,000 spots, and cover spatial resolutions from nanometer to micrometer scale, highlighting the versatility, scalability, and efficiency of SMART.

To incorporate spatial information, SMART builds a k-nearest neighbor graph based on spatial coordinates. Both message propagation and node aggregation in the graph neural network are inherently constrained by the spatial topology, ensuring that spatial information is integrated from the beginning of training. Moreover, SAGEConv, which aggregates information from spatial neighbors at each layer, allows node embeddings to incorporate expression features from nearby spatial locations, effectively imposing an implicit spatial constraint. This design aligns with mainstream spatial transcriptomics methods such as spaGCN[7], STAGATE[9], and STAligner[43]. To avoid overfitting to expression features and neglecting spatial structure, we introduced an early stopping mechanism based on the slope of the loss curve. Training halts automatically when the change in reconstruction or triplet loss stabilizes, helping preserve spatial integrity in the learned embeddings.

In SMART, the trade-off between spot specificity and spatial structure modeling can be adjusted via the triplet loss weight and the reconstruction weight in the spatial graph. The triplet loss emphasizes similarity among spots of the same type, regardless of their spatial distance. In contrast, the reconstruction loss, computed through a GraphSAGE-based neural network that incorporates spatial graphs during message passing, constrains the embeddings to reflect the underlying spatial structure. SMART introduces a tunable parameter $\lambda \in [0,1]$ to control the trade-off between spatial continuity and expression specificity. Increasing λ emphasizes spatial locality, while decreasing it prioritizes expression-based spot type separation. Moreover, the number of neighbors k in the spatial graph also affects model behavior. It significantly influences spatial structure modeling. For example, when $k = 0$, the model performs well in identifying spot types but poorly in capturing spatial domains (Supplementary Fig. 5).

To validate the design of SMART's framework, we conducted a series of ablation studies and comparison. We replaced the default GraphSAGEConv with different graph encoders, GCNConv, GATConv, and GraphConv, keeping other parameters constant. The GraphSAGEConv encoder consistently demonstrated better stability, faster convergence, and higher clustering accuracy (Supplementary Fig. 1c–e). To construct negative samples for triplet loss, we selected the top 60% most distant nodes (measured in Euclidean distance) as negatives. A sensitivity analysis across different thresholds (20%, 40%, 60%, and 80%) revealed that 60% yields the most stable and accurate results across multiple datasets (Supplementary Fig. 2). This design also aligns well with the concept of "semi-hard negatives" commonly used in triplet loss strategies. We also studied the effect of the margin parameter (τ) in the triplet loss, comparing $\tau \in \{0.1, 0.2, 0.5, 1.0, 2.0\}$, and found that $\tau = 0.5$ provided robust performance across datasets (Supplementary Fig. 4a), which we adopt as the default value in SMART. For the construction of anchor-positive pairs, we employed a mutual nearest neighbors (MNN) strategy, in which we select node pairs that are among each other's top-n nearest neighbors. To evaluate the effect of this design, we conducted an ablation study across different n values, and analyzed their impact on training stability and representation quality (Supplementary Fig. 3g). To quantify the effect of the triplet loss, we conducted experiments under three conditions: using only the reconstruction loss, using only the triplet loss, and using both losses together. We evaluated the spatial clustering consistency using supervised metrics across multiple real datasets. The removal of the triplet loss notably degraded the model's ability to discriminate expression structures, especially in tissues with complex organization such as the human lymph node (Supplementary Fig. 4c). This supports the conclusion that the triplet loss plays a key role in enhancing semantic separation in the latent embedding space.

SMART integrates spatial information using graph neural networks and adopts GraphSAGE (Graph Sample and Aggregate), which efficiently generates node embeddings via neighbor sampling, making it particularly suitable for large-scale or sparse graphs. To integrate multi-omics data, SMART employs a modular stacking design, enabling it to handle datasets with any number of omics layers. This modular design, which processes each omics layer independently, offers both scalability and flexibility for accommodating emerging technologies capable of profiling multiple omics simultaneously. Leveraging both the GraphSAGE architecture and modular omics integration, SMART demonstrates shorter training times compared to methods such as PRESENT and SpatialGlue. Across diverse datasets, SMART converges reliably within 100–300 epochs. For instance, on the largest transcriptomic-epigenomic spatial dataset (CUT&Tag–RNA-seq mouse brain, 9752 spots), SMART achieved the best performance with the shortest training time and minimal memory usage. For instance, on the

largest transcriptomic-epigenomic spatial dataset (CUT&Tag–RNA-seq mouse brain, 9752 spots), SMART achieved the best performance with the shortest training time and minimal memory usage. Similarly, on Stereo-CITE-seq datasets with different spatial resolutions, SMART also achieved optimal training time and memory efficiency. These results highlight SMART's computational efficiency and its suitability for large-scale spatial omics applications. Moreover, SMART still performs well on tissue types with high spatial heterogeneity. To further validate this, we conducted additional experiments using tumor datasets, including the 10X breast cancer and glioblastoma datasets[44,45]. However, the lack of manual annotations in these datasets limited their applicability for fully quantitative evaluation. Our results show that SMART, when applied to multi-omics data (RNA and protein), can successfully identify continuous spatial domains even in highly heterogeneous tissues (Supplementary Figs. 20–23). Compared to using single-omics data alone, the integration of multi-omics data significantly enhances the ability to capture spatial structure. Of note, although SMART is primarily designed for spot-based data, it is applicable to single-cell–segmented data. Here, spot-based data refer to datasets where each spatial unit represents a fixed-size region that is neatly arranged and may aggregate signals from multiple cells (Supplementary Notes), whereas single-cell-segmented data refer to datasets in which each spatial unit corresponds to an individual biological cell obtained via image-guided segmentation or imaging-based technologies (Supplementary Notes). In spot-based analyses, nodes correspond to spots or bins, spatial graphs are built based on the coordinates of spatial units using a k-nearest-neighbor (kNN) strategy. In single-cell-segmented analyses, nodes correspond to individual cells, with cell centroids used as spatial coordinates. kNN or radius of one cell distance to select the nearest cells around can both be used in spatial graph construction (Supplementary Fig. 24). Because SMART formulates spatial multi-omics data as a graph over spatial units, replacing spots with cells only changes the granularity of graph nodes without altering the model architecture. Apart from the substitution of spatial units, other model components, loss functions, and hyperparameters remain unchanged for single-cell-segmented data.

SMART was originally designed for spot-level spatial data as the primary application scenario. However, it is also applicable to single-cell-segmented data by simply replacing the meaning of nodes from spots to cells without modifying the network architecture or optimization strategy. An additional experiment on cell-segmented data derived from the same Stereo-CITE-seq dataset yields similar results with those obtained from spot-based data (Supplementary Fig. 25). When the same data source can simultaneously generate spot-level data and single-cell segmented data, SMART does not exhibit a preference for any particular input form in our experimental results (Supplementary Fig. 25). However, different input data quality may influence the model's performance. For example, in our Stereo-CITE-seq experiment (Fig. 5c), settings with a bin size between 20 and 100 yielded better performance compared to finer or coarser spatial divisions. For single-cell segmented data, spatial granularity is no longer a primary factor, but model performance may be affected by the accuracy of the cell segmentation algorithm. For both input types, the differences arising from the quality and characteristics of the input data itself are likely much higher than those induced by the model.

For better usage of SMART, we would like to highlight three typical situations where SMART is preferred: (1) SMART is preferred for spatially discrete (non-contiguous) niches of the same type, such as follicles in spleen. It is because SMART not only performs neighborhood aggregation to capture spatially adjacent spots/cells but also incorporates a triplet loss to detect similarities among distant spots/cells, while CellCharter only conducts neighborhood aggregation. (2) Multi-section spatial multi-omics data. For multiple sections, SMART constructs cross-section triplets, which enables better capture of relationships across sections. (3) For large-scale datasets. Owing to its streamlined architectural design, SMART can reduce both computation time and memory consumption (Fig. 5b, d, e).

Despite its strong performance in identifying spatial domains and discrete tissue structures, SMART also has limitations. SMART prioritizes feature-driven discrimination of spatial units and the accurate identification of domain spots, rather than explicitly enforcing strong spatial smoothness. As a result, in tissues where spatial domains vary continuously with gradual transitions, SMART may yield embeddings with unclear boundaries compared with methods that impose stronger spatial continuity priors. This behavior reflects an inherent trade-off between preserving molecular heterogeneity and enforcing spatial smoothness. Moreover, while the modality-independent network architecture of SMART offers notable advantages in computational efficiency and scalability, the current lack of explicit modeling of inter-modality relationships may limit SMART's ability to capture complex biological interplay across omics layers. SMART does not yet support integration across sections with completely non-overlapping modalities (e.g., RNA in one section and protein in another with no shared spots) or with different omics compositions (e.g., integrating an RNA + ATAC section and an RNA+protein section). In the near future, we plan to incorporate modules that model inter-modality interactions while preserving spatial domain identification. As cross-modality interactions are critical for understanding molecular processes, enhancing SMART in this way would improve integration accuracy, interpretability, and support both modality-specific and shared domain discovery.

Methods

Ethics statement

This study exclusively used publicly available datasets. No new human or animal samples were collected, and no experiments involving human participants or animals were performed by the authors. Therefore, ethical approval was not required for this study. Sex and/or gender was not considered in the study design or analysis, as this work focuses on computational method development rather than biological comparisons between sexes or genders, and sex- or gender-disaggregated information was not consistently available across the public datasets used.

Data

Simulated data were generated using the method outlined by Townes et al.[41], which is designed to produce synthetic datasets with specific spatial structures, facilitating the validation of the model's ability to identify and process spatial features effectively. The simulation generated multivariate counts with spatial correlation patterns. We use the ggblocks simulation which is based on the Indian Buffet Process, where each latent factor represents a typical spatial pattern defined over a 36×36 grid of locations ($N = 1296$ spatial locations). For the RNA and ATAC modalities, we generated spatial expression matrices of size 1296 cells × 1000 genes (peaks) characterized by four factors and following a zero-inflated negative binomial distribution (ZINB). For the protein modality, we generated a spatial expression matrix of size 1296 cells × 100 proteins characterized by four factors and following a negative binomial distribution (NB). The negative binomial distribution introduces randomness by modeling the variability of discrete events. This variability is considered noise, representing the uncertain portion of the data and simulating the random fluctuations observed in real-world data.

10x Genomics CytAssist Visium datasets were generated using the 10x Genomics CytAssist Visium platform, which enables the joint profiling of transcriptome and protein expression. They include the human lymph node dataset used in SMART for spatial transcriptomics and proteomics integration, and the three-section tonsil dataset used in SMART-MS for multi-section integration. This platform integrates spatial whole-transcriptome sequencing with multiplexed protein

detection on a single tissue section via next-generation sequencing (NGS), with a spatial resolution of 55 μm, a spot spacing of 100 μm, and an imaging area of 6.5 mm × 6.5 mm. The lymph node dataset consists of two tissue sections: Human Lymph Node A1 (3484 spots) and D1 (3359 spots), each containing expression profiles of 18,085 genes and 31 proteins. The tonsil dataset includes three sections with 4326, 4519, and 4521 spots, respectively. Each section contains 18,085 genes; the first two sections include 31 proteins, while the third section includes 35 proteins.

MISAR-seq mouse brain datasets were generated by measuring spatial transcriptome and chromatin accessibility in the mouse brain using MISAR-seq[2], a microfluidic indexing–based spatial assay for transposase-accessible chromatin and RNA sequencing. MISAR-seq enables spatially resolved joint profiling of chromatin accessibility and gene expression. In this dataset, the tissue is positioned on a microfluidic chip, where a grid-based spatial barcoding system assigns unique DNA barcodes across 2500 grids, each 50 × 50 μm, covering approximately 5–20 cells per grid. This dataset includes eight sections of the mouse fetal brain at embryonic days, i.e., E11.0, E13.5, E15.5, and E18.5. The number of grids covered by the sections ranges from 1263 to 2248, with a total of 32,285 sequenced genes and peak counts varying from 60,112 to 143,879.

Spatial Epigenome–Transcriptome mouse brain datasets originate from the method proposed by Zhang et al.[1], including spatial ATAC–RNA-seq and CUT&Tag–RNA-seq, which enable simultaneous detection of chromatin accessibility or histone modifications (e.g., H3K27me3, H3K27ac, or H3K4me3) together with gene expression within a single tissue section at near single-cell resolution. In the P22 mouse brain coronal sections, a 100 × 100 barcode grid (pixel size: 20 μm) was used to cover the entire hemisphere region, with most pixels capturing 1–3 cells. The RNA + H3K27me3 dataset used to evaluate the scalability of SMART contains 9752 spots, 70,470 peaks, and 25,881 genes. Meanwhile, the RNA + ATAC dataset used to assess multi-section integration with SMART-MS includes three sections with 2372, 2497, and 9215 spots, respectively, along with 161,514 peaks and 18,107 genes.

STARmap-RIBOmap-Align datasets were generated by applying STARmap, a high-resolution spatial transcriptomics technology developed by Wang et al.[46], and RIBOmap, a three-dimensional in situ spatial translatomics method introduced by Zeng et al.[47], to consecutive mouse brain tissue sections, enabling joint measurement of transcriptomic and translatomic profiles for 5,413 genes at single-cell and molecular resolution. The STARmap section contained 59,165 spots, and the RIBOmap section contained 58,692 spots. Using the CAST projection algorithm introduced by Tang[48], spots from the STARmap section were aligned to those in the RIBOmap section, resulting in dual-omics data for 58,692 spots.

The SPOTS mouse spleen dataset was generated using Spatial PrOtein and Transcriptome Sequencing (SPOTS)[4] to profile spatial transcriptomic and proteomic data for assessing multi-section integration with SMART-MS. Similar to 10X Human Lymph Node dataset, SPOTS also offers a spatial resolution of 55 μm, with a distance of 100 μm between spots, covering an imaging area of 6.5 mm × 6.5 mm. This dataset includes two sections, SPOTS Spleen Replicate 1 and SPOTS Spleen Replicate 2, containing 2647 and 2759 spots, respectively, with each section capturing the expression of 32,285 genes and 21 antibody-derived tags (ADTs).

Stereo-CITE-seq datasets were generated using the method proposed by Liao et al.[6], which combines CITE-seq with Stereo-seq to enable simultaneous profiling of the transcriptome and proteome on the same tissue section at subcellular spatial resolution (0.5 μm) with high reproducibility and precision. This technique achieves a spatial resolution of up to 0.22 μm, with a spot spacing of 0.5 μm and an imaging area of 200 mm^2. The mouse spleen dataset used to evaluate the scalability of SMART was processed into multiple resolution levels, including Bin10 (5 μm), Bin20 (10 μm), Bin50 (25 μm), Bin100 (50 μm), Bin150 (75 μm), and Bin200 (100 μm), comprising 756,430; 189,867; 30,652; 7,782; 3507; and 2001 spots, respectively. In addition to spot-level representations generated by spatial binning, we also used the image-guided cell segmentation results provided by the Stereo-CITE-seq pipeline (cellbin.gef), in which RNA and protein signals are aggregated at the single-cell level. These data were used to evaluate SMART's applicability to single-cell–segmented spatial data. This dataset includes expression data for 29,034 genes and 128 ADTs. For evaluating multi-section integration with SMART-MS, we used three mouse thymus sections at Bin50 resolution (25 μm), which contain 4253, 4646, and 4228 spots, respectively. Each section captures expression profiles for 23,221 to 23,960 genes and 19 ADTs.

Data preprocessing

Spatial transcriptomic data were preprocessed by first removing genes expressed in fewer than 10 spots to minimize the impact of low-abundance signals. The top 3000 or 5000 highly variable genes were then selected to retain features contributing most to data variability. Expression counts were normalized by scaling the total expression in each spot to 10,000 to account for differences in sequencing depth, followed by logarithmic transformation to reduce the influence of highly expressed genes and approximate a normal distribution. The data were subsequently scaled to zero mean and unit variance.

Spatial chromatin accessibility data were preprocessed by first applying a TF–IDF transformation to accessible chromatin regions, with a scaling factor of 10,000 to downweight commonly accessible regions and emphasize features with higher variability across spots. Peak counts within each spot were then normalized by scaling the total signal to 10,000 to correct for differences in sequencing depth. A logarithmic transformation was subsequently applied to stabilize variance and reduce the influence of highly accessible regions, preparing the data for downstream analyses.

Spatial protein data were preprocessed by first applying a centered log-ratio transformation to remove scale differences across proteins. The data were then standardized to zero mean and unit variance to ensure that all protein features were on a comparable scale for downstream analyses.

The SMART framework

SMART is a deep learning model that leverages graph neural networks and metric learning to integrate spatially-resolved multi-modal omics into unified embeddings. For each modality, SMART extracts principal components (PCs) using principal component analysis (PCA) and constructs spatial neighboring graphs with PCs as node features. Spatial graphs are built based on the coordinates of spatial units using a k-nearest-neighbor (kNN) strategy. In spot-based analyses, nodes correspond to spots or bins, whereas in single-cell–segmented analyses, nodes correspond to individual cells, with cell centroids used as spatial coordinates. For single-cell–segmented data, radius within one cell distance to select the nearest cells around can also be used in spatial graph construction. Then, the model employs SAGEConv (Sample and Aggregate Convolution) to aggregate both the spatial information and the PCs from a specific modal omics, thereby synthesizing a integrate representation of the omics. Ultimately, these representations are combined into a unified intermodal embedding.

However, using spatial neighboring graphs for aggregation may neglect the feature correlations even in long distance. Therefore, we further adjust the graph aggregation using metric learning with triplet loss to ensure the mutual correlation between spots or bins in the original omics. The loss functions of reconstruction to the original omics and the triplet loss are crucial to guarantee the representation of the original omic intact after embedding with spatial information.

In a spatial multi-omics dataset featuring three distinct omics modalities, the inputs to SMART consist of characteristics for each

modality $\mathbf{X}_1 \in \mathbb{R}^{m\times n_1}$, $\mathbf{X}_2 \in \mathbb{R}^{m\times n_2}$, and $\mathbf{X}_3 \in \mathbb{R}^{m\times n_3}$ and the spatial coordinates $S \in \mathbb{R}^{m\times 2}$. Here m represents the number of spots in the tissue section, while n_1, n_2, and n_3 correspond to the number of features in the three omics modality, for example, the number of genes in transcriptome. In the simulated spatial multi-omics data, $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ represent the features of transcriptome, proteome, and epigenome. The output of SMART is a unified embedded representation $\mathbf{Z}\in\mathbb{R}^{N\times f}$ in the low-dimensional latent space, which integrates information from multiple spatial omics. The parameter f is denoted as the dimension in the latent space. If there are two omics, the inputs become $\mathbf{X}_1 \in \mathbb{R}^{m\times n_1}$, $\mathbf{X}_2 \in \mathbb{R}^{m\times n_2}$ and the rest of the operation is the same. Overall, the main components of SMART can be categorized into five modules: principal component analysis, spatial neighboring graph construction, SAGEConv encoder, triplet construction, and SAGEConv decoder. Each module is elaborated in the following sections.

Principal component analysis

We first apply PCA to the preprocessed data from each modality: $\mathbf{X}_1 \in \mathbb{R}^{m\times n_1}$, $\mathbf{X}_2 \in \mathbb{R}^{m\times n_2}$, and $\mathbf{X}_3 \in \mathbb{R}^{m\times n_3}$, in order to reduce the dimensionality of the data to d_1, d_2 and d_3 dimensions, respectively. Specifically, after applying PCA, we obtain the dimensionality-reduced feature matrices: $\mathbf{X}_1 \in \mathbb{R}^{m\times d_1}$, $\mathbf{X}_2 \in \mathbb{R}^{m\times d_2}$ and $\mathbf{X}_3 \in \mathbb{R}^{m\times d_3}$. By retaining the most representative principal components, we achieve the dimensionality reduction. Typically, the dimensionality of the transcriptomics and chromatin accessibility modalities is reduced to 30, while the dimensionality of the proteomics modality is determined based on the number of ADTs.

Spatial neighboring graph construction

To learn the spatial coordination of omics, we constructed a spatial neighboring graph with each spot or bin as node and principal components as node feature. Each node (spot or bin) was connected to its k nearest neighbors according to the Euclidean distance between spots on the coordinates. Consequently, the spatially-resolved omics was converted into an undirected spatial neighbor graph $G=(V,E)$, where V represents m spots and E represents the set of connecting edges among the spots in their k nearest neighbors. The number of neighbors k was set to 4 for datasets with a quadrilateral spatial layout (e.g., SPOTS, MISAR-seq, and Stereo-CITE-seq), and to 6 for 10X Visium data with a hexagonal lattice. We denoted $\mathbf{A}\in\mathbb{R}^{m\times m}$ as the adjacency matrix of the graph G and $\mathbf{A}_{ij}=1$ when there is an edge between node i and node j, otherwise $\mathbf{A}_{ij}=0$.

SAGEConv encoder and multimodal feature integration

SAGEConv (Sample and Aggregate Convolution) is a convolution layer designed to efficiently capture the graph structure and learn high-quality representation for graph features, particularly well-suited for large-scale graph data in spatial multi-omics. It samples a subset of the neighboring nodes and aggregates their features, thus reducing computational complexity. SAGEConv employs various aggregation functions and is stacked in multiple layers to enhance node representation. For omics $i(i\in\{1,2,3\})$, we use the principal components $\widetilde{\mathbf{X}}_i$ as input to SAGEConv encoder to learn the graph-specific representation $\mathbf{H}_i$ of each node. $\mathbf{h}_{iv}^{l}$ represents node v's representation of the original feature of omics i through the l-th ($l\in\{1,2,\cdots,L\}$) layer of SAGEConv encoder. The formula is as follows:

$$\mathrm{h}_{i\mathcal{N}(v)}^{l}=f(\mathcal{N}(v)) \tag{1}$$

$$\mathrm{h}_{iv}^{l}=g(\sigma\left(\mathrm{W}_i^{l}\cdot[\mathrm{h}_{iv}^{l-1};\mathrm{h}_{i\mathcal{N}(v)}^{l}]\right)) \tag{2}$$

where $f(x)$ is an aggregator function and we use the mean function as the aggregate function which formula is as follows:

$$f(\mathcal{N}(v))=\frac{1}{|\mathcal{N}(v)|}\sum_{u\in\mathcal{N}(v)}\mathrm{h}_u^{(l-1)} \tag{3}$$

$\mathcal{N}(v)$ represents the set of neighbors of node v in the spatial neighbor graph $G=(V,E)$. W_i denote a trainable weight matrix and $\sigma(x)=\max(0,x)$ is a nonlinear activation function. The normalization function $g(x)=\frac{x}{||x||_2}$ applies L2 normalization to the updated node embeddings, ensuring that the feature vector for each node has a unit Euclidean norm.

In the output of omics i embeddings $\mathbf{h}_{iv}^{L}\in\mathbb{R}^{f}$, f is the dimension of latent embedded representation. By default, f is set to 64. We use the fully connected layer to integrate the SAGEConv output of different omics. Assuming integrating three omics, we first perform concatenation, i.e., $\mathbf{h}_v=[\mathbf{h}_{1v}^{L};\mathbf{h}_{2v}^{L};\mathbf{h}_{3v}^{L}]\in\mathbb{R}^{3f}$, and obtain latent embedded representation $\mathbf{Z}\in\mathbb{R}^{f}$ that integrates all modalities using fully connected layer

$$\mathrm{Z}=\mathrm{W}\cdot\mathrm{h}_v+\mathrm{b} \tag{4}$$

where $\mathbf{W}$ denotes a trainable weight matrix that transforms the input into another feature space, and $\mathbf{b}$ denotes a trainable bias vector. The number of GraphSAGE encoder layers was set to two based on the results of an ablation study (Supplementary Fig. 1a).

SAGEConv decoder

SAGEConv decoder is designed to reconstruct the principal components of different omics from the latent embedded representation. For omics $i(i\in\{1,2,3\})$, we the use final latent embedded representation $\mathbf{Z}$ as input to SAGEConv decoder to reconstruct the graph-specific representation $\widetilde{\mathbf{h}}_i$ of each node. The decoder is composed of multiple stacked layers to enhance reconstruction capabilities. $\widetilde{\mathbf{h}}_{iv}^{l}$ is denoted as the representation of node v reconstruction feature of omics i through the l-th ($l\in\{1,2,\cdots,L\}$) layer of SAGEConv decoder. In particular, $\widetilde{\mathbf{h}}_{iv}^{0}=\mathbf{Z}$. The formula is as follows:

$$\widetilde{\mathrm{h}}_{i\mathcal{N}(v)}^{l}=f(\mathcal{N}(v)) \tag{5}$$

$$\widetilde{\mathrm{h}}_{iv}^{l}=g(\sigma\left(\mathrm{W}_i^{l}\cdot[\widetilde{\mathrm{h}}_{iv}^{l-1};\widetilde{\mathrm{h}}_{i\mathcal{N}(v)}^{l}]\right)) \tag{6}$$

where $f(x)$ is an aggregator function as same as formula (3). $\mathbf{W}_i$ denote a trainable weight matrix and $\sigma(x)=\max(0,x)$ is a nonlinear activation function. The normalization function $g(x)=\frac{x}{||x||_2}$ applies L2 normalization to the updated node embeddings, ensuring that the feature vector for each node has a unit Euclidean norm. Finally, we get the reconstructed feature $\widetilde{\mathbf{h}}_{iv}^{L}\in\mathbb{R}^{d_i}$ of omics i. The number of GraphSAGE decoder layers was set to two based on the results of an ablation study (Supplementary Fig. 1a).

Reconstruction loss

To guarantee the latent representation appropriately encode and preserve the information of expression features from all the omics, we enforced that different omic features can be recover through the SAGEConv decoder. We applied the reconstruction loss to maximize the similarity between the output of the decoder and the principal components of each omics. Assuming $\widetilde{\mathrm{x}}_{i_v}\in\mathbb{R}^{d_i}$ is principal components of omics i for node v, which is the input to the SAGEConv encoder, the reconstruction feature output by the SAGEConv decoder

is $\widetilde{\mathbf{h}}_{iv}^{L} \in \mathbb{R}^{d_i}$. The target is to minimize the difference between $\widetilde{\mathbf{h}}_{iv}^{L}$ and $\widetilde{\mathbf{x}}_{i_v}$. Therefore, objective function of the reconstruction loss is:

$$\mathscr{L}_{\text{recon}} = \sum_{i=1}^{I} \alpha_i \sum_{v=1}^{N} \left\| \widetilde{\mathrm{x}}_{i_v} - \widetilde{\mathrm{h}}_{iv}^{L} \right\|_F^2 \tag{7}$$

where I represents the number of omics, α_i is the weight factor used to adjust the contribution of modality i, and N represents the number of spots in the section, also referring to the number of nodes in the spatial neighboring graph. SMART allows the adjustment of α to meet possible requirements on depending more on one or two omics. In our experiments, we consistently set α to 1. During training, the model will try to minimize the reconstruction loss $\mathscr{L}_{\text{recon}}$.

Triplet construction

The spatial neighboring graph is only based on spatial structure and neglect the relationships among spots (for example, spots in the same type) in relatively long distance. Therefore, we adjust the weight in the graph aggregation using metric learning based on the relationship of omic features. We construct triplets, including anchor points, positive points, and negative points among omic features, and regulate the graph aggregation using triplet loss.

To construct the triplets according to omic feature relationships, we calculate the Euclidean distance between the principal components of each omics. For each spot, we select the top k nearest neighbors to form the nearest neighboring set (where k is set to 3 by default). If spot j and spot i are mutually in the nearest neighbor set of each other, then spot i and j serve as the anchor sample a_i and positive sample p_i. The negative sample n_1 is chosen from the $m \times$ r farthest set of spot i, where m is the number of the total spots r is a ratio set to 0.6 by default. For each omics, a set of triplets $T^{tri} = \{(a_1, p_1, n_1), (a_2, p_2, n_2), \cdots\cdots, (a_s, p_s, n_s)\}$, s represents the number of triples in the triplet set in this modality.

Triplet loss

To enable the adjustment of graph aggregation and unified representation according to features, we apply triplet loss in metric learning, specifically using the features' relationship represented by triplet. Based on the triplet set $T_i^{tri} = \{(a_1, p_1, n_1), (a_2, p_2, n_2), \cdots\cdots, (a_{s_i}, p_{s_i}, n_{s_i})\}$ in omics i, we try to ensure that the representation of anchor spots is similar to the positive spots and different from the negative spots at the same time. The objective function of triplet loss is given by the following equation:

$$\mathscr{L}_{\text{triplet}} = \sum_{i=1}^{I} \alpha_i \cdot \frac{1}{s_{\text{i}}} \sum_{(a,p,\text{n}) \in T_i^{tri}}^{s_{\text{i}}} max\left(\left\| \mathrm{Z}_a - \mathrm{Z}_p \right\|_F^2 - \left\| \mathrm{Z}_a - \mathrm{Z}_n \right\|_F^2 + \tau, 0 \right) \tag{8}$$

where I represents the number of modalities, α_i is the weight factor used to adjust the contribution of modality i, $\mathbf{Z}$ is the latent embedded representation, and τ (default 1.0) is the margin used to enforce the distance between positive and negative pairs. SMART allows the adjustment of α to meet possible requirements on depending more on one or two omics. In our experiments, we consistently set α to 1. During training, the object is to minimize the triplet loss $\mathscr{L}_{\text{triplet}}$.

Model training of SMART

To obtain a better unified representation of spatial multi-omics, SMART includes spatial information and multi-omic relationships using the reconstruction loss and triplet loss. The model is trained by minimizing overall loss function as follows:

$$\mathscr{L} = \lambda \mathscr{L}_{\text{recon}} + (1 - \lambda) \mathscr{L}_{\text{triplet}} \tag{9}$$

where $\mathscr{L}_{\text{recon}}$ is the reconstruction loss described in Eq. (7) and $\mathscr{L}_{\text{triplet}}$ is the triplet loss described in Eq. (8), λ represents the weighting coefficient for the reconstruction loss in the overall loss function. It takes a value between 0 and 1, and is typically set to 0.5, assigning equal importance to the reconstruction loss and the triplet loss.

The model was trained using the Adam optimizer, with training parameters adjusted for different datasets. However, the batch size was consistently set to one graph. To prevent the model from overfitting expression features while neglecting spatial structure, we implemented an early stopping strategy based on the slope of the loss functions. Specifically, training was automatically terminated when the rate of change (slope) of either the reconstruction loss or the triplet loss fell below a predefined threshold of 0.0001. This approach ensures that the model does not excessively optimize expression features at the expense of spatial structure in pursuit of lower reconstruction loss, nor does it overemphasize similarity in expression values via the triplet loss. The training process was conducted on a single NVIDIA RTX 3090Ti GPU using the PyTorch and PyTorch Geometric frameworks.

The SMART-MS framework and training

To integrate multi-sections spatial multi-omics data and map the cellular features from different sections into a common latent space, capturing the biological heterogeneity shared across tissue sections, we developed the SMART-MS model, specifically designed for multi-sections spatial multi-omics data. SMART-MS follows the design principles of SMART, with the addition of modules for multi-sections data integration, batch effect correction, and cross-section spatial neighbor graph construction. The training process remains consistent with that of SMART.

Multi-sections data integration

In a spatial multi-omics dataset, assuming there are t sections of multi-omics data, for section i and modality j, the raw input data is $\mathbf{X}_j^i \in \mathbb{R}^{m_i \times n_j^i}$. Here m_i represents the number of spots in the tissue section i, while n_j^i corresponds to the eigenvalues of modality j in the tissue section i. The final matrix obtained by integrating all sections for modality j can be represented as $\mathbf{X}_j \in \mathbb{R}^{(\sum_{i=1}^{t} m_i) \times \bigcap_{i=1}^{t} \mathrm{n}_j^i}$, where $\sum_{i=1}^{t} m_i$ represents the sum of spots in all sections and $\bigcap_{i=1}^{t} \mathrm{n}_j^i$ represents the number of intersections of features across sections. If modality j is RNA or protein, the intersection of features typically refers to the intersection of genes or ADTs. If modality j is ATAC, the intersection operation $\cap$ identifies overlapping regions that are accessible in the chromatin in both datasets (i.e., peaks that appear in both datasets). Specifically, the intersection operation detects portions of peak regions from two ATAC-seq datasets that overlap, meaning that the peaks in both datasets cover the same genomic region.

Batch effect correction

For modality j, we apply PCA to the preprocessed data from multiple sections integration for dimensionality reduction, resulting in batch-effect-corrected input data $\widetilde{\mathbf{X}}_j \in \mathbb{R}^{(\sum_{i=1}^{t} m_i) \times p_j}$, where p_j is the dimension after dimensionality reduction. Next, we apply Harmony[49] to remove batch effects from the input data. Harmony removes batch effects by performing low-dimensional embedding and batch alignment, eliminating differences between batches while preserving the biological variation in the data. It calculates batch differences and iteratively adjusts the spatial representation of samples, enabling samples from different batches to be reasonably aligned in a unified latent space, thereby improving the accuracy of downstream analyses.

Finally, the batch-effect-corrected feature matrix is denoted as $\mathbf{H}_j \in \mathbb{R}^{(\sum_{i=1}^{t} m_i) \times p_j}$.

Triplet construction in multi sections

In the multi-section integration task, to more effectively align omic features across different tissue sections, we extended the construction of triplets to a cross-section setting. Specifically, each triplet consists of an anchor and a negative sample from the same section, while the positive sample is selected from a different section.

The triplet construction proceeds as follows: we first enumerate all possible section pairs and compute the Euclidean distances between spots based on low-dimensional features obtained after Harmony-based batch correction. For a spot i from section 1, which serves as the anchor a, we identify the farthest $N_1 \times r$ spots within section 1 as negative sample candidates, where N_1 denotes the total number of spots in section 1 and r is a predefined ratio (default: 0.6). One spot is then randomly sampled from this candidate pool as the negative sample n. Concurrently, in section 2, if a spot j is among the top k nearest neighbors of spot i in the cross-section feature space, and i is also among the top k nearest neighbors of j, then j is selected as the positive sample p of a.

Assuming there are b sections in total, for each omics, the complete triplet set for this omics can be expressed as the union over all unordered section pairs:

$$T^{tri} = \bigcup_{1 \le x < y \le b} T^{x,y} \tag{10}$$

where $T^{x,y}$ denotes the number of triplets generated from section pair x and y.

Cross-section spatial neighbor graph construction

For each section i across multiple sections, we construct the spatial neighbor graph $\mathbf{A}_i \in \mathbb{R}^{m_i \times m_i}$ for section i using the k-nearest neighbor (kNN) algorithm, just as in SMART for constructing single-section spatial neighbor graph. Finally, the cross-modal spatial neighbor graph can be represented by the following formula:

$$\mathrm{A} = \mathrm{diag}\left(\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_t\right) \tag{11}$$

where t represents the number of sections.

After obtaining the spatial neighbor graph $\mathbf{A}$ for multi-sections integration and the feature matrix $\mathbf{H}_j$ for modality $\mathbf{H}$, the subsequent steps are the same as those in SMART.

Evaluation metrics

To evaluate the model's data integration and spatial recognition capabilities, we utilized seven supervised metrics (ARI (adjusted rand index), AMI (adjusted mutual information), NMI (normalized mutual information), Homo (homogeneity), MI (mutual information), V-Measure and FMI (Fowlkes-Mallows Index)) along with one unsupervised metric, the Moran's I score. To evaluate the batch effect removal capability of SMART-MS, we employed two metrics: iLISI[49] and kBET[50].

ARI is a measure used to evaluate the similarity between two clustering results by adjusting for the chance grouping of elements. The ARI is calculated using the formula:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E[RI]}}{\max(\mathrm{RI}) - \mathrm{E[RI]}} \tag{12}$$

where RI is the Rand Index, $E[RI]$ is the expected Rand Index for random clustering, and $\max(RI)$ is the maximum possible Rand Index. If U and V are two clustering results, Rand Index can be obtained from the following formula:

$$RI = \frac{TP + TN}{\binom{n}{2}} \tag{13}$$

where True Positives (TP) represents the number of pairs of elements that are in the same cluster in both U and V, True Negatives (TN) represents the number of pairs of elements that are in different clusters in both U and V. n represents the total number of elements in the dataset and $(0exn2)$ represents the total number of pairs of elements in the dataset.

Mutual Information (MI) is a measure from information theory that quantifies the amount of information gained about one random variable through the knowledge of another random variable. If U and V are two clustering results, the Mutual Information between clustering results U and V is given as:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{\left|U_i \cap V_j\right|}{N} \log \frac{N\left|U_i \cap V_j\right|}{\left|U_i\right|\left|V_j\right|} \tag{14}$$

Where $|U_i|$ is the number of the samples in cluster U_i, $|V_j|$ is the number of the samples in cluster V_j and N represents the total number of elements in the dataset.

Normalized Mutual Information (NMI) is a normalization of the Mutual Information (MI) score to scale the results between 0 and 1. Normalized Mutual Information between clustering results U and V is given as:

$$NMI = \frac{MI(U,V)}{avg(H(U), H(V))} \tag{15}$$

where $H(U)$ is entropy of clusters U, measuring the uncertainty or randomness within cluster, which can be obtained from the following formula:

$$H(U) = -\sum_{i=1}^{k} \frac{|U_i|}{N} \log \frac{|U_i|}{N} \tag{16}$$

where k represents the number of clusters in U, $|U_i|$ is the number of the samples in cluster U_i and N represents the total number of elements in the dataset. The same applies to $H(V)$.

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. Adjusted Mutual Information between clustering results U and V is given as:

$$AMI = \frac{MI(U,V) - E[MI(U,V)]}{avg(H(U), H(V)) - E[MI(U,V)]} \tag{17}$$

where $E[MI(U,V)]$ is the expected Mutual Information, which represents the average MI one would expect from random clustering, can be obtained from the following formula:

$$E[MI(U,V)] = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \sum_{n_{ij}=(a_i+b_i-N)^+}^{\min(a_i b_j)} \frac{n_{ij}}{N} \log\left(\frac{N.n_{ij}}{a_i b_j}\right) \frac{a_i! b_j! (N-a_i)! \left(N-b_j\right)!}{N! n_{ij}! \left(a_i - n_{ij}\right)! \left(b_j - n_{ij}\right)! \left(N - a_i - b_j + n_{ij}\right)!} \tag{18}$$

where $a_i = |U_i|$ is the number of the samples in cluster U_i, $b_i = |V_j|$ is the number of the samples in cluster V_j, $(a_i + b_i - N)^+ = \max(1, a_i + b_i - N)$ is the maximum of 1 and $a_i + b_i - N$ and N represents the total number of elements in the dataset.

Homogeneity (Homo) is a clustering evaluation metric, which measures how well a clustering assigns all data points in a true class to the same cluster, is defined as follows:

$$\text{Homogeneity} = 1 - \frac{H(C|K)}{H(C)} \tag{19}$$

where $H(C,|,K)$ is the conditional entropy of the true class labels C given the predicted clusters K, can be obtained from the following formula:

$$H(C|K) = -\sum_{k=1}^{|K|} P(k) \sum_{c=1}^{|C|} P(c|k) \log P(c|k) \tag{20}$$

where $|K|$ is the number of clusters, $|C|$ is the number of true classes, $P(k) = \frac{n_k}{n}$ is the probability of a point being in cluster k with n the total number of samples and n_k the number of samples belonging to cluster k, and $P(c|k) = \frac{n_{c,k}}{n_k}$ is the conditional probability of class c given cluster k with $n_{c,k}$ the number of samples from class c assigned to cluster k. $H(C)$ is entropy of the ground truth labels C and is given by:

$$H(C) = -\sum_{c=1}^{|C|} P(c) \log P(c) \tag{21}$$

where $|C|$ is the number of true classes, $P(c) = \frac{n_c}{n}$ is the probability of a point being in cluster c with n the total number of samples and n_c the number of samples belonging to class c.

V-measure is a clustering evaluation metric that balances two aspects of clustering quality: homogeneity and completeness. The V-measure is the harmonic mean of homogeneity and completeness, is defined as follows:

$$\text{V} - measure = 2 \times \frac{homogeneity \times completeness}{homogeneity + completeness} \tag{22}$$

where $completeness = 1 - \frac{H(K|C)}{H(K)}$ is a clustering metric that evaluates whether all members of the same true class are assigned to the same cluster.

The Fowlkes-Mallows Index (FMI) is a metric used to evaluate the similarity between two clustering results by measuring the geometric mean of precision and recall between the predicted and ground truth clusterings. The FMI is defined as:

$$FMI = \sqrt{\frac{TP}{TP + zFP} \cdot \frac{TP}{TP + FN}} \tag{23}$$

where TP (True Positives) denotes the number of pairs of elements that are in the same cluster in both the predicted clustering C and the ground truth clustering G, FP (False Positives) denotes the number of pairs that are in the same cluster in C but in different clusters in G and FN (False Negatives) denotes the number of pairs that are in different clusters in C but in the same cluster in G.

Moran's I score is a statistic used to measure spatial autocorrelation, indicating how similar or dissimilar values are among nearby spatial units and is given as:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \bar{x}) \left(x_j - \bar{x}\right)}{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}\right) \sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{24}$$

where N is the total number of spatial units, x_i is the value of the variable at location i, $\bar{x}$ is the mean of the variable x over all locations, w_{ij} is the spatial weight between locations i and j, indicating spatial proximity or adjacency (often binary, where $w_{ij} = 1$ if i and j are neighbors, otherwise 0) and the summations $\sum_{i=1}^{N} \sum_{j=1}^{N}$ are taken over all pairs of locations.

Integration Local Inverse Simpson's Index (iLISI) is a commonly used metric for evaluating the quality of multi-batch data integration, measuring the degree of mixing of different batches within each spot's local neighborhood. By calculating the diversity of batch labels among neighbors, iLISI reflects the extent of batch effect removal. The metric is typically normalized between 0 and 1, where 0 indicates complete batch separation and 1 indicates perfect batch mixing. To accommodate integration methods based on graph structures, Graph iLISI[51] improves the neighborhood definition by using shortest path distances within the graph instead of traditional Euclidean distances, avoiding biases caused by embedding spaces and providing a more accurate assessment of batch mixing in graph-structured data.

K-nearest neighbor Batch Effect Test (kBET) is an important metric for evaluating the effectiveness of multi-batch data integration. Its core idea is to assess whether the batch label distribution within the k-nearest neighbors (kNN) of each spot matches the global batch distribution using a chi-squared test, thereby determining if batch effects have been effectively removed. The kBET score is calculated as the average rejection rate of all these tests; a lower raw score indicates poorer batch mixing. For standardized interpretation, the score is often normalized between 0 and 1, where higher values indicate better batch mixing. kBET can be applied to integration methods based on embeddings and can also be extended to graph structures, with adaptations such as connected component analysis and neighbor number adjustments enhancing its applicability.

Modality correlation analysis

Given that $\mathbf{X}^{pca} \in \mathbb{R}^{m \times d}$ represents the PCA-reduced features of a modality, where m denotes the number of spots in the tissue section, p denotes the reduced feature dimensionality after PCA. The aggregated features from each modality using a certain method are represented as $\mathbf{X}^{lat} \in \mathbb{R}^{m \times f}$, where f represents the dimensionality of the aggregated features. The distance matrices D^{pca} and D^{lat} for the two types of features, $\mathbf{X}^{pca} \in \mathbb{R}^{m \times p}$ (PCA-reduced features) and $\mathbf{X}^{lat} \in \mathbb{R}^{m \times f}$ (aggregated features), can be calculated as follows:

$$D^{pca}{}_{ij} = \sqrt{\sum_{k=1}^{p} \left(\mathrm{X}^{pca}{}_{i,k} - \mathrm{X}^{pca}{}_{j,k}\right)^2} \tag{25}$$

$$D^{lat}{}_{ij} = \sqrt{\sum_{k=1}^{f} \left(\mathrm{X}^{lat}{}_{i,k} - \mathrm{X}^{lat}{}_{j,k}\right)^2} \tag{26}$$

where $D^{pca}{}_{ij} \in \mathbb{R}^{m \times m}$ and $D^{lat}{}_{ij} \in \mathbb{R}^{m \times \mathrm{m}}$ represent the Euclidean distances between spots i and j in the PCA-reduced and aggregated feature spaces, respectively. The Pearson correlation coefficient P_i of spot i between the modality-specific feature and the aggregated feature can be computed as follows:

$$P_i = \frac{\sum_j \left(D^{pca}{}_{ij} - \overline{D^{pca}{}_i}\right)\left(D^{lat}{}_{ij} - \overline{D^{lat}{}_i}\right)}{\sqrt{\sum_j \left(D^{pca}{}_{ij} - \overline{D^{pca}{}_i}\right)^2} \cdot \sqrt{\sum_j \left(D^{lat}{}_{ij} - \overline{D^{lat}{}_i}\right)^2}} \tag{27}$$

For all spots, the correlation vector $P = \{P_1, P_2, \cdots, P_m\}^T$ represents the correlation coefficients between the modality-specific features and the aggregated features.

Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment. Statistical analyses and

reproducibility details are described in the Methods. For batch effect assessment, kBET was computed using a chi-squared test with a predefined significance threshold of $\alpha = 0.05$. All computational experiments were conducted on a single NVIDIA RTX 3090 Ti GPU. Code and data availability statements are provided to ensure reproducibility.

Algorithms availability. The benchmark algorithms used in this study are publicly available.

WNN: https://github.com/dylkot/pyWNN
SpatialGlue: https://github.com/JinmiaoChenLab/SpatialGlue
CellCharter: https://github.com/CSOgroup/cellcharter
SpaMultiVAE: https://github.com/ttgump/spaVAE
COSMOS: https://github.com/Lin-Xu-lab/COSMOS
MISO: https://github.com/kpcoleman/miso
scMM: https://github.com/kodaim1115/scMM
PRESENT: https://github.com/lizhen18THU/PRESENT
MEFISTO: https://biofam.github.io/MOFA2/MEFISTO
MOFA + : https://doi.org/10.5281/zenodo.3735162

MultiVI and totalVI were implemented using the scVI tools v1.1.6: https://scvi-tools.org/

SNF was applied using muon v0.1.6: https://github.com/scverse/muon

WNN is implemented using Seurat v4: https://github.com/satijalab/seurat

Detailed implementation settings for all benchmark algorithms are provided in the Supplementary Information. All data analyses were conducted using Python v3.9, with PyTorch v2.4.1, torch-geometric v2.3.0, scikit-learn v1.5.1, and scanpy v1.10.2. The SMART software developed in this study is publicly available at https://github.com/Xubin-s-Lab/SMART-main.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study were obtained from publicly available repositories. The 10X Visium Human Lymph Node data can be accessed from the Gene Expression Omnibus (GEO) with accession code GSE263617. The MISAR-seq mouse brain data can be accessed from National Genomics Data Center with accession number OEP003285. The Stereo-CITE-seq data can be accessed from BGI STOmics Cloud (https://cloud.stomics.tech/). The spatial CUT&Tag–RNA-seq and spatial ATAC–RNA-seq mouse brain data can be accessed at GEO with accession code GSE205055 or UCSC Cell and Genome Browser (https://brain-spatial-omics.cells.ucsc.edu). The STARmap and RIBOmap mouse brain data can be accessed from Zenodo (https://zenodo.org/record/8041114) or Single Cell Portal (SCP) (https://singlecell.broadinstitute.org/single_cell/study/SCP1835). The SPOTS mouse spleen data can be accessed at GEO with accession code GSE198353. The 10X Visium Human Tonsil data can be accessed from https://zenodo.org/records/12654113/preview/data_imputation.zip?include_deleted=0#tree_item0. The data used as input to the methods tested in this study have been uploaded to Zenodo and are freely available at https://doi.org/10.5281/zenodo.17093158[52]. Source data are provided with this paper.

Code availability

The SMART toolkit is accessible at https://github.com/Xubin-s-Lab/SMART-main. The tutorial for implementing SMART to analyze spatially multi-omics data is available at: https://smart-tutorials.readthedocs.io/en/latest/.The Jupyter notebooks for reproducing the results in this paper are available at https://github.com/Xubin-s-Lab/SMART-main/tree/SMART-reproduce. The GitHub repository was linked to Zenodo with the https://doi.org/10.5281/zenodo.18538147[53].

References

1. Zhang, D. et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113–122 (2023).
2. Jiang, F. et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. *Nat. Methods* **20**, 1048–1057 (2023).
3. Vickovic, S. et al. SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nat. Commun.* **13**, 795 (2022).
4. Ben-Chetrit, N. et al. Integration of whole transcriptome spatial profiling with protein markers. *Nat. Biotechnol.* **41**, 788–793 (2023).
5. Merritt, C. R. et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* **38**, 586–599 (2020).
6. Liao, S. et al. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. Preprint at https://doi.org/10.1101/2023.04.28.538364 (2023).
7. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
8. Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput Sci.* **2**, 399–408 (2022).
9. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
10. Long, Y. et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat. Commun.* **14**, 1155 (2023).
11. Zong, Y. et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. Preprint at https://doi.org/10.1101/2022.01.14.476408 (2022).
12. Xu, C. et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res.* **50**, e131 (2022).
13. Ren, H., Walker, B. L., Cang, Z. & Nie, Q. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat. Commun.* **13**, 4076 (2022).
14. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (PMLR, 2021).
15. Yao, W., Liu, C., Yin, K., Cheung, W. K. & Qin, J. Addressing asynchronicity in clinical multimodal fusion via individualized chest X-ray generation. *Adv. Neural Inf. Process. Syst.* **37**, 29001–29028 (2024).
16. Jaume, G. et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11579–11590. https://doi.org/10.1109/CVPR52733.2024.01100 (2024).
17. Song, Q., Gong, T., Gao, S., Zhou, H. & Li, J. QUEST: quadruple multimodal contrastive learning with constraints and self-penalization. *Adv. Neural Inf. Process. Syst.* **37**, 28889–28919 (2024).
18. Yang, Y., Wan, F., Jiang, Q.-Y. & Xu, Y. Facilitating multimodal classification via dynamically learning modality gap. *Adv. Neural Inf. Process. Syst.* **37**, 62108–62122 (2024).
19. Zhang, H., Patel, V. M. & Chellappa, R. Hierarchical multimodal metric learning for multimodal classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2925–2933. https://doi.org/10.1109/CVPR.2017.312 (IEEE, Honolulu, HI, 2017).
20. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

21. Ghazanfar, S., Guibentif, C. & Marioni, J. C. Stabilized mosaic single-cell data integration using unshared features. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01766-z (2023).
22. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep. Methods* **1**, 100071 (2021).
23. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
24. Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics* **36**, 4137–4143 (2020).
25. Gayoso, A. et al. Joint probabilistic modeling of single-cell multiomic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
26. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
27. Velten, B. et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
28. Ashuach, T. et al. MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).
29. Long, Y. et al. Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nat. Methods* https://doi.org/10.1038/s41592-024-02316-4 (2024).
30. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nat. Genet.* **56**, 74–84 (2024).
31. Coleman, K. et al. Resolving tissue complexity by multimodal spatial omics modeling with MISO. *Nat. Methods* **22**, 530–538 (2025).
32. Zhou, Y. et al. Cooperative integration of spatially resolved multi-omics data with COSMOS. *Nat. Commun.* **16**, 27 (2025).
33. Tian, T., Zhang, J., Lin, X., Wei, Z. & Hakonarson, H. Dependency-aware deep generative models for multitasking analysis of spatial omics data. *Nat. Methods* **21**, 1501–1513 (2024).
34. Li, Z. et al. Cross-modality representation and multi-sample integration of spatially resolved omics data. Preprint at https://doi.org/10.1101/2024.06.10.598155 (2024).
35. Miyamoto, Y. & Ishii, M. Spatial diversity of *in vivo* tissue immunity. *Int. Immunol.* dxae051. https://doi.org/10.1093/intimm/dxae051 (2024).
36. Ombrato, L. et al. Generation of neighbor-labeling cells to study intercellular interactions in vivo. *Nat. Protoc.* **16**, 872–892 (2021).
37. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **30** (2017).
38. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *Proc. Int. Conf. Learn. Representat.* Vol. 4, 2713–2726 (2017).
39. Veličković, P. et al. Graph attention networks. In *Proc. Int. Conf. Learn. Representat.* (2018).
40. Morris, C. et al. Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks. *arXiv.org* https://arxiv.org/abs/1810.02244v5.
41. Townes, F. W. & Engelhardt, B. E. Nonnegative spatial factorization applied to spatial genomics. *Nat. Methods* **20**, 229–238 (2023).
42. Hong, S. B., Shin, Y.-W., Hong, J. B., Lee, S. K. & Han, B. Exploration of shared features of B cell receptor and T cell receptor repertoires reveals distinct clonotype clusters. *Front. Immunol.* **13**, 1006136 (2022).
43. Zhou, X., Dong, K. & Zhang, S. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nat. Comput. Sci.* **3**, 894–906 (2023).
44. Visium CytAssist Gene and Protein Expression Library of Human Breast Cancer, IF, 6.5mm (FFPE). *10x Genomics* https://www.10xgenomics.com/datasets/gene-and-protein-expression-library-of-human-breast-cancer-cytassist-ffpe-2-standard.
45. Visium CytAssist Gene and Protein Expression Library of Human Glioblastoma, IF, 11mm (FFPE). *10x Genomics* https://www.10xgenomics.com/datasets/gene-and-protein-expression-library-of-human-glioblastoma-cytassist-ffpe-2-standard.
46. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
47. Zeng, H. et al. Spatially resolved single-cell translatomics at molecular resolution. *Science* **380**, eadd3067 (2023).
48. Tang, Z. et al. Search and match across spatial omics samples at single-cell resolution. *Nat. Methods* **21**, 1818–1829 (2024).
49. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
50. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
51. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
52. Zheng, X. Datasets used in SMART's experiments [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17093158 (2025).
53. Zheng, X. SMART: spatial multi-omic aggregation using graph neural networks and metric learning. Zenodo. https://doi.org/10.5281/zenodo.18538147 (2026).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 32300554 to X.Z. and Grant No. 62176164 to Z.D.) We would like to acknowledge the support from Dongguan Key Laboratory for AI and Dynamical Systems and Institute of Artificial Intelligence at Great Bay University. The computational resources are provided by the Songshan Lake High Performance Computing Center (SSL-HPC) at Great Bay University.

Author contributions

X.Z., Z.D. and J.C. supervised the project and wrote manuscript. X.Z., Z.D. and Q.C. conceived the idea and designed the experiment. X.Z. and Q.C. analyzed the data and performed experiments. Q.C. collected the data. W.H. validated experimental results, developed package documentation, and helped revised the manuscript. All authors discussed the results and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70821-5.

Correspondence and requests for materials should be addressed to Jinmiao Chen or Xubin Zheng.

Peer review information *Nature Communications* thanks Joseph Beechem, Lin Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026