SpatialFuser: A Unified Deep Learning Framework for Spatial Multi-Omics Data Integrative Analysis

Wenhao Cai[1,2], Weizhong Li[1,2,*]

[1.] School of Medicine, Sun Yat-sen University, Shenzhen, China

[2.] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

* Corresponding author: Weizhong Li (liweizhong@mail.sysu.edu.cn)

ORCID: Weizhong Li (0000-0002-9003-7733)

Abstract

Recent advances in spatial multi-omics technologies provide unprecedented opportunities to integrate and interpret molecular features within the tissue microenvironment. Here we present SpatialFuser, the first unified deep learning framework for detailed molecular profiling within individual tissue sections and multi-omics integrative analysis of cross-modality spatial data. SpatialFuser offers comprehensive tools and models for accurate spatial interpretation, robust cross-modality integration, and effective cross-slice alignment of spatial epigenomics, transcriptomics, proteomics, and metabolomics. Benchmarking results demonstrate SpatialFuser's superior performance and reliability in spatial domain detection and consecutive slice alignment task compared to existing state-of-the-art methods. Applications to diverse datasets spanning various resolution and omics types further highlight SpatialFuser's ability to capture precise molecular patterns, reveal developmental dynamics, and uncover fine-grained biological variation from complementary perspectives, offering a holistic view of cellular and tissue properties. The SpatialFuser framework is open-source and available at https://github.com/liwz-lab/SpatialFuser.

Introduction

Spatial technologies enable researchers to measure the expression levels and reveal cellular heterogeneity across multiple molecular layers such as transcriptomics, epigenomics, proteomics, and metabolomics at specific spatial locations, and have been widely applied in studies of tissue organization and molecular dynamics[1]. These technologies can be roughly categorized into imaging-based approaches including osmFISH[2], MERFISH[3], CODEX[4], MIBI[5], CosMx SMI[6], and MALDI-MSI[7], and sequencing-based techniques such as 10x Visium[8], BaristaSeq[9], Stereo-seq[10], MAGIC-seq[11], Stereo-CITE-seq[12], Spatial ATAC-seq[13], and Spatial ATAC–RNA-seq[14]. Recent advances also cover AI-powered techniques such as PLATO[15], which leverage computational approaches to overcome current limitations in biotechnology, such as low sequencing depth. With different data modalities conveying unique biological insights from diverse complementary perspectives, we now have an unprecedented opportunity to achieve a comprehensive spatial landscape of cellular profile, thereby enhancing our understanding of key processes such as cell function[16], tissue development[17], and disease progression[18].

Computational methods have been developed to decipher intra-slice heterogeneity from typically spatial transcriptomics data with high-dimension and high-sparsity through spatial domain detection. SpaGCN[19] employs a graph convolutional network to integrate spatial gene expression with coordinate information for spatial domain identification. STAGATE[20] adopts a spatially-aware graph attention autoencoder to capture spatial structures, while SEDR[21] uses a deep variational autoencoder to jointly learn gene expression and spatial relationships. These unimodal methods assume neighboring cells share similar molecular profiles, enabling information propagation only within nearby spatial regions while overlooking important long-range feature dependencies. Meanwhile, the growing expansion of spatial technologies into multi-omics indicates that increasing numbers of tissue slices will be profiled from proteomic, epigenomic, and other molecular perspectives, underlying this is an urgent need for

unified computational tools capable of decoding spatial data from diverse omics modalities.

To achieve a deeper understanding of complex biological systems, integrative analysis of spatial data has become increasingly important. However, heterogeneous datasets often suffer from batch effects or modality biases, resulting in discrepancies in feature structures, expression distributions, and biological semantics, posing challenges to effective data integration and slice alignment. Existing computational tools attempt to address this issue but remain limited. For example, STAligner[22] employs a graph attention autoencoder and triplet adversarial learning to align spatial transcriptomic slices, but it does not extend to multi-omics integration. SLAT[23] uses graph convolutional networks and adversarial learning to reconstruct 3D tissues while correcting batch effects across slices, yet it fails to integrate modalities with weakly correlated features across transcriptomics and proteomics. SpatialGlue[24] applies a dual-attention graph model to jointly analyze spatial multi-omics from the same tissue section but cannot perform cross-slice integration. Those integrative methods are effective for unimodal or same-slice multi-omics data but lack generalizability in heterogeneous multi-omics scenarios. Methodologically, many integrative tools assume consistent feature structures across datasets, requiring dimensionality matching through preprocessing. This can lead to information loss when the intrinsic complexity of raw modalities differs. Some also assume similar cell compositions across slices and adopt rigid integration strategies, increasing the risk of over-alignment. Therefore, new computational tools are needed for spatially aware cross-omics integration that offer better biological interpretability and generalizability, enabling more flexible and robust analysis across diverse and emerging spatial omics modalities.

Here, we introduce SpatialFuster, the first unified framework for single-sample deciphering, cross-slice integration, and alignment of spatial multi-omics data, supporting both single-sample and joint analysis across multiple experimental platforms and modalities. SpatialFuster constructs graph representations from raw multi-omics data and employs graph neural networks to learn the low-dimensional

embeddings for capturing spatial heterogeneity across modalities. For integration and alignment, it incorporates flexible coordinate matching strategies to support diverse application scenarios and adopts a relaxed contrastive learning scheme to correct batch effects or modality biases, aligning datasets into a shared spatial reference.

We first benchmarked SpatialFuster on classical 10x Visium datasets to evaluate its quantitative accuracy and robustness in spatial domain identification. As the first universal tool tailored for comprehensive spatial multi-omics analysis, we further validated its cross-platform and cross-modality generalizability on osmFISH and CODEX datasets. In unimodal multi-slice joint analysis tasks, SpatialFuster better conducted adjacent BaristaSeq slices integration and alignment than state-of-art methods reported in previous benchmarks[25], and effectively captured fine-grained developmental dynamics through integrative analysis of multi-developmental-stage Stereo-seq samples. In multi-omics scenarios, SpatialFuser enhances the detection of lowly expressed but epigenetically primed marker genes by integrating spatial ATAC and RNA data, thereby uncovering lineage-primed regulatory programs[26]. This capability highlights the potential of SpatialFuser to infer cell state transitions and transcriptional readiness within spatial tissue architectures. Furthermore, in integrative analyses across multiple samples, SpatialFuser effectively integrates transcriptomic (MAGIC-seq), proteomic (PLATO), and metabolomic (MALDI-MSI) data across varying spatial resolutions, enabling the identification of fine-grained functional regions that are challenging to resolve with low-resolution technologies alone. These results demonstrate the versatility and power of SpatialFuser in decoding complex biological systems through comprehensive spatial multi-omics integration.

Results

The overview architecture of SpatialFuser

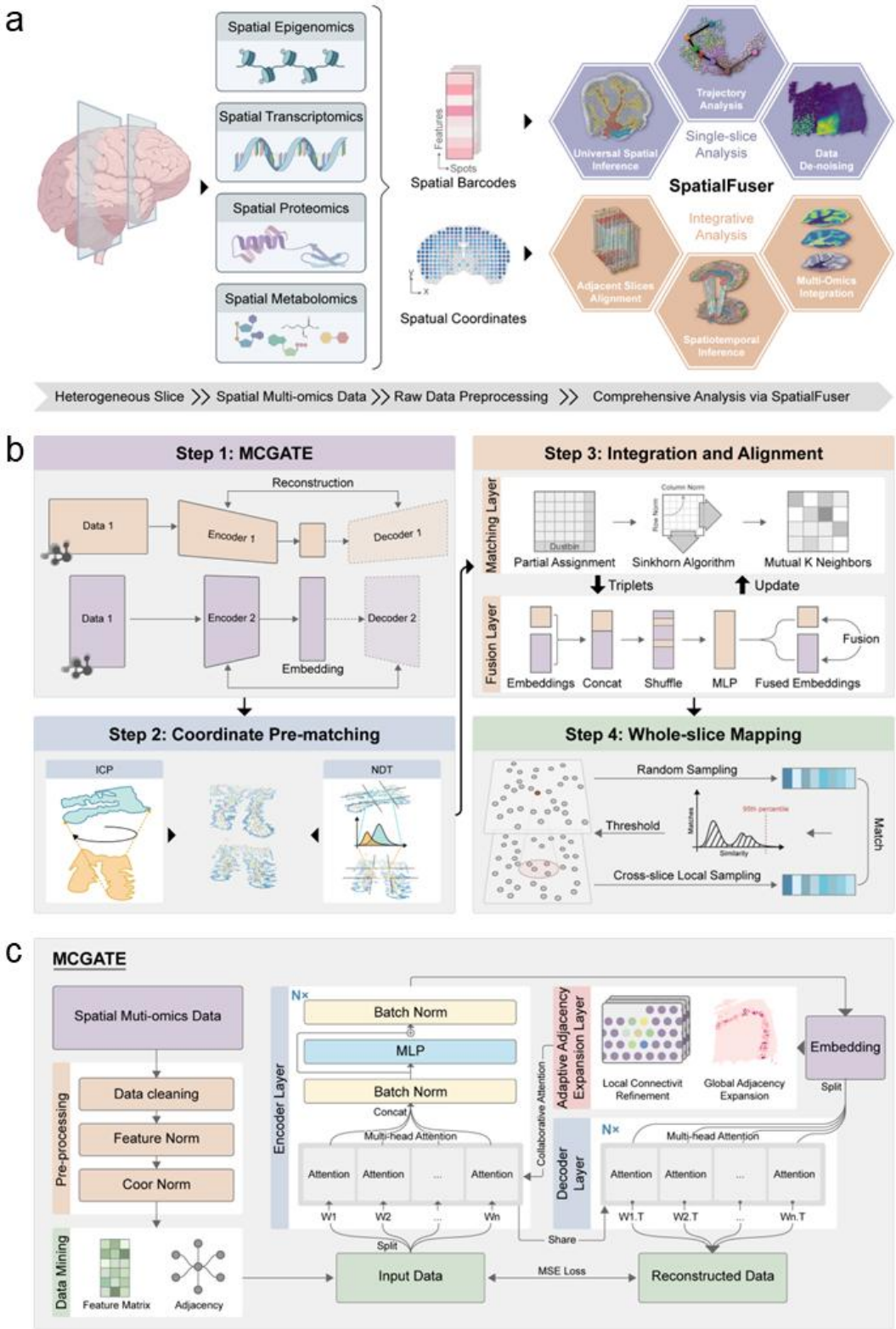


Fig. 1. A unified deep learning framework for single-sample and cross-slice spatial multi-omics analysis. a, Overview of SpatialFuser. SpatialFuser integrates molecular features and spatial coordinates from spatial omics data, supporting both fine-grained single-slice profiling and cross-sample analysis across modalities such as epigenomics, transcriptomics, proteomics, and metabolomics. **b,** Schematic representation of the

SpatialFuser framework. The analysis workflow includes: MCGATE-based embedding learning, coordinate pre-matching for rigid slice registration, dual matching and fusion layers for multimodal integration, and whole-slice spot mapping for downstream integrative analysis. **c,** Network architecture of MCGATE. MCGATE represents spatial omics data as a graph by constructing a spatial adjacency network from coordinates. It applies multi-head collaborative graph attention to learn modality-specific embeddings via a reconstruction task, guided by a spatial graph for local attention and a dynamically updated feature similarity graph for long-range attention. Designed to be modular and scalable, MCGATE can be applied across diverse spatial omics modalities and platforms.

SpatialFuser takes molecular features and spatial coordinates from spatial omics data as input, enabling both fine-grained single-slice analysis and cross-sample integrative analysis across modalities including epigenomics, transcriptomics, proteomics, and metabolomics (Fig. 1a). SpatialFuser is a unified framework designed to support diverse tasks, including accurate spatial interpretation, robust cross-modality integration, and effective cross-slice alignment (Fig. 1b). It first represents the input as graph-structured data by constructing a spatial adjacency network based on spatial coordinates, and employs a Multi-head Collaborative Graph Attention auToEncoder (MCGATE) to learn modality-specific embeddings through a reconstruction task (Fig. 1c). In the encoder, MCGATE uses the predefined spatial adjacency graph to guide local attention aggregation, while also constructing a feature similarity graph in the low-dimensional latent space to capture long-range feature dependencies. This similarity graph is fed back into the encoder to dynamically refine local attention weights and establish long-range attention. The collaborative attention mechanism enables the integration of multi-scale spatial context and semantic information, allowing SpatialFuser to more effectively learn spatial molecular patterns. As the first general embedding model for spatial omics data, MCGATE is designed to be modular and scalable, enabling SpatialFuser to be broadly applicable across diverse modalities and experimental platforms.

An essential initial step for the alignment of spatial omics data is the pre-matching of coordinates across tissue slices[27]. To accommodate diverse spatial distributions, SpatialFuser models spatial expression data as 2D point clouds and offers two

geometric registration algorithms: Iterative Closest Point (ICP) and Normal Distributions Transform (NDT) (Fig. 1b Step 2). ICP is well suited for datasets with regularly spaced spots and clear structural grids (e.g., 10x Visium), while NDT performs better on datasets with densely packed, irregularly distributed spots and indistinct boundaries (e.g., CODEX and Stereo-seq). This flexible pre-matching strategy provides robust initialization for subsequent fine-grained alignment.

Alignment and integration are not strictly independent tasks[25]. To support cross-slice analysis, SpatialFuser jointly addresses these two tasks through an iteratively trained dual-layer architecture consisting of a matching layer based on the Sinkhorn algorithm and a fusion layer grounded in contrastive learning (Fig. 1b Step 3). In the matching layer, the alignment problem is formulated as a graph matching task and solved as an optimal transport problem via the Sinkhorn algorithm. To better support cross-resolution alignment and prevent over-alignment, a dustbin channel mechanism[28] is incorporated, allowing unmatched or ambiguous points to be filtered out. In the fusion layer, SpatialFuser employs a contrastive learning strategy, where high-confidence matching pairs from the matching layer serve as anchor-positive pairs to guide the correction of batch effects and modality biases, and randomly sampled anchor-negative pairs from non-neighbouring regions within the same slice to preserve the model's sensitivity to true biological variation. SpatialFuser introduces an absolute distance constraint into the triplet loss to enhance model stability and convergence[29], ensuring accurate and consistent mapping of spatial data into a shared embedding space. Finally, SpatialFuser conducts rigorous quality control to achieve whole-slice spot mapping (Fig. 1b Step 3), providing a reliable reference for downstream integrative analyses.

In brief, SpatialFuser consists of four key components: MCGATE, a coordinate pre-matching module, a matching layer based on the Sinkhorn algorithm, and a fusion layer driven by contrastive learning. Each component can operate independently or seamlessly integrate into a unified workflow, providing a comprehensive and

streamlined solution for downstream analysis of spatial multi-omics data across diverse experimental platforms and resolutions.

Accurate spatial inference across diverse experimental platforms and modalities via SpatialFusor

SpatialFuser proposes a unified tool for spatial multi-omics inference, enabling accurate detection of spatial distributions of tissue domains or cell types. To evaluate the performance of SpatialFusor against existing spatial domain detection algorithms, we conducted a benchmark work using the classic 10x Visium[8] human dorsolateral prefrontal cortex (DLPFC) dataset[30], which has well-defined morphological boundaries and reliable manual annotations (Fig. 2a & Supplementary Fig. 1). The original labels were used as ground truth, and five evaluation metrics were employed: Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Homogeneity, Completeness, and V-Measure. We first compared SpatialFusor with other state-of-the-art spatial domain detection methods based on the Mclust[31] clustering strategy, including SpaGCN[19], STAGATE[20], SEDR[21], and STAligner[22] (Fig. 2b). On average across the 12 DLPFC slices, SpatialFusor achieved the highest accuracy (mean = 0.625; max = 0.809; min = 0.533), followed by STAligner (mean = 0.556; max = 0.675; min = 0.469). Notably, although SpaGCN integrates high-resolution histological images during training, its performance was substantially lower than that of other methods relying solely on molecular and spatial information (mean ARI = 0.422; max = 0.541; min = 0.225), suggesting that incorporating histological data may provide limited benefit for spatial domain detection in this context. To further evaluate SpatialFusor's advantages in downstream analysis, we also benchmarked it using commonly used clustering methods in omics data analysis, including Leiden[32] (Supplementary Fig. 2a) and Louvain[33] (Supplementary Fig. 2b). These results demonstrate that SpatialFusor enables accurate spatial domain identification and consistently outperforms existing state-of-the-art tools across diverse clustering contexts.

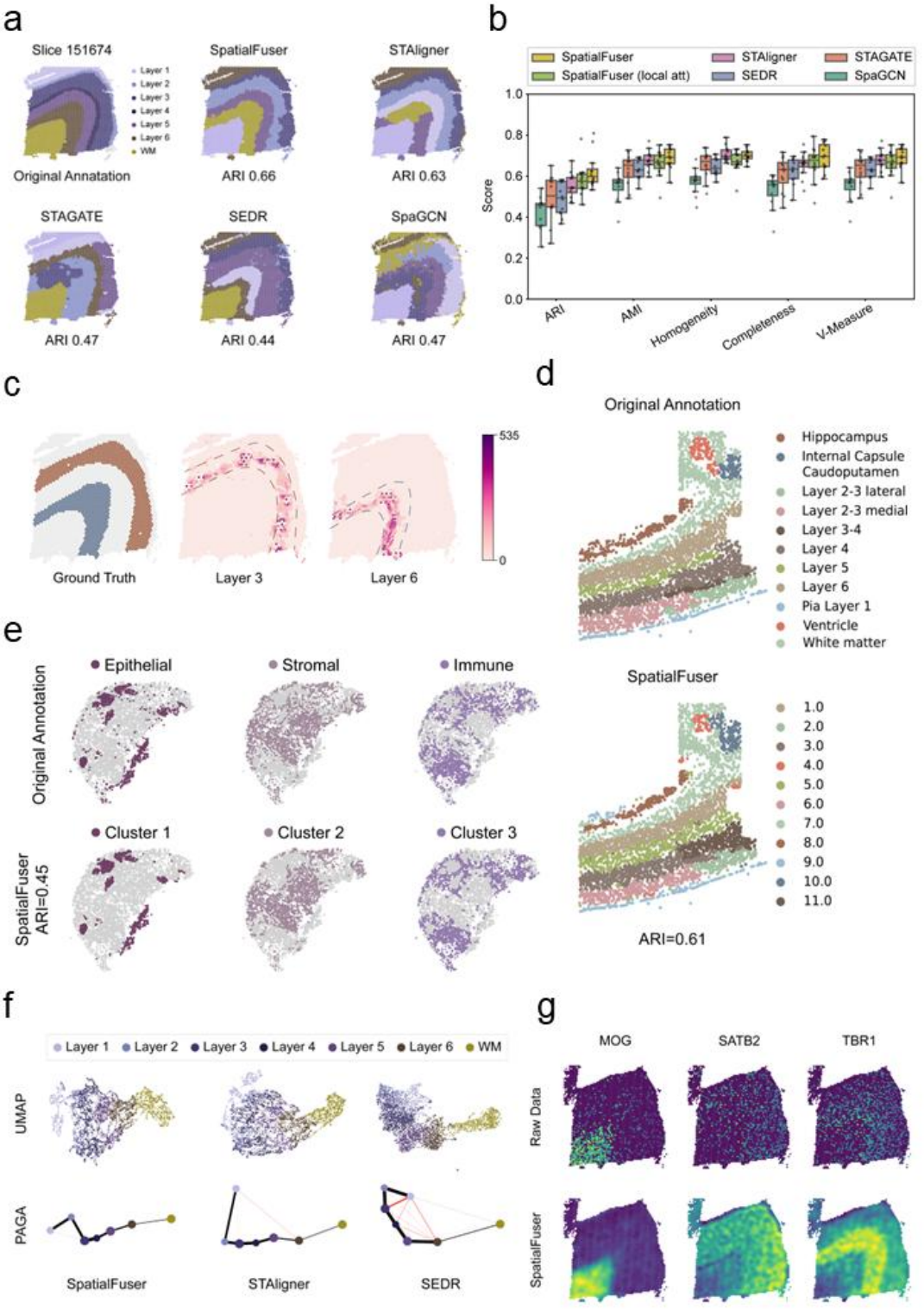


Fig. 2. SpatialFuser accurately resolves tissue domains across diverse experimental platforms and modalities. a, Comparison of spatial domains identified by SpaGCN, STAGATE, SEDR, STAligner, and SpatialFuser for DLPFC slice-151674. **b,** Boxplots of five evaluation metrics (ARI, AMI, Homogeneity, Completeness, V-measure) for Mclust clustering results across 12 DLPFC sections. **c,** Long-range feature propagation paths of ten

randomly selected spots from cortical layers 3 and 6 during MCGATE training on DLPFC slice-151676. **d,** Spatial domains identified by Mclust clustering on low-dimensional embeddings generated by SpatialFuser in the osmFISH somatosensory cortex dataset. **e,** Distributions of epithelial, stromal, and immune cells identified by Mclust clustering on low-dimensional embeddings generated by SpatialFuser for slice-210308_TMA2_reg6 of the CODEX muscle-invasive bladder cancer dataset. **f,** UMAP visualizations and PAGA graphs generated from SEDR, STAligner, and SpatialFuser embeddings for DLPFC section-151675. Red lines highlight incorrect trajectory inferences. **g,** SpatialFuser enhances spatial patterns of layer-enriched and marker genes in the DLPFC dataset. Raw and SpatialFuser-denoised spatial expression patterns are shown for MOG, SATB2, and TBR1 in DLPFC section-151675. ARI, adjusted Rand index; AMI, adjusted mutual information; DLPFC, dorsolateral prefrontal cortex; MCGATE, multi-head collaborative graph attention autoencoder.

We conducted an ablation study to evaluate the effectiveness of the collaborative attention mechanism. The results demonstrate that the long-range attention significantly enhances the SpatialFusor's ability to capture spatial distribution patterns (Fig. 2b, Supplementary Fig. 2a-b). Using slice-151676 from the DLPFC dataset as an example, we randomly selected 10 spatial spots from cortical layers 3 and 6 and tracked their long-range attention sampling during training. The distribution of sampled spots closely corresponded to the tissue region patterns (Fig. 2c), further validating the effectiveness and biological relevance of long-range information propagation. We also tested the robustness of SpatialFusor by comparing clustering accuracy across different hyperparameters (see Methods), demonstrating that the model is insensitive to network structure and random seeds (Supplementary Fig. 3a-b).

Spatial omics data generated from different experimental platforms and modalities exhibit significant differences in feature structures, distributions, and biological semantics. We further evaluated SpatialFusor's ability to identify spatial domains across multiple platforms and modalities using the osmFISH somatosensory cortex dataset[2] and the CODEX[4] human muscle-invasive bladder cancer tumor dataset[34]. On the image-based osmFISH dataset, SpatialFusor accurately identified spatial domains in the somatosensory cortex with an ARI of 0.61 (Fig. 2d). For the CODEX spatial proteomics dataset, we re-labeled the three main cell types, epithelial cells, stromal cells, and immune cells, mentioned in the original paper based on the annotations provided

by the authors for downstream spatial inference. SpatialFusor successfully captured the spatial distribution patterns of all the three key cell types (Fig. 2e, ARI = 0.45) The relatively lower ARI score may be attributed to the higher spatial heterogeneity of tumor tissues[35]. These results demonstrate the robustness and versatility of SpatialFusor in cross-platform, multi-omics spatial data analysis.

To evaluate SpatialFusor's utility in downstream analysis, we assessed its low-dimensional embeddings for trajectory inference and reconstructed feature for spatial variable gene (SVG) identification. Partition-based graph abstraction[36] (PAGA) was employed to infer trajectories on DLPFC slice-151675 using embeddings learned by SpatialFusor, SEDR, and STAligner. In the UMAP plot (Fig. 2f), SpatialFusor's embedding closely recapitulated the developmental progression from white matter to layer 6 and sequentially to layer 1. Comparatively, embeddings from SEDR and STAligner exhibited less plausible developmental trajectories. We further compared the expression patterns of three region-enriched marker genes between the raw data and SpatialFusor-denoised data in DLPFC slice-151675. After reconstruction, all selected genes showed clear differential expression in their enriched regions, while the raw spatial expression was noisy and inconsistent (Fig. 1g). Collectively, these results demonstrate that SpatialFusor can accurately capture feature expression patterns and reduce noise, thereby better supporting single-sample analysis and spatial heterogeneity studies.

Spatiotemporal alignment for revealing tissue heterogeneity and developmental dynamics

Integrative and comparative analysis of spatial omics datasets is critical for unravelling spatial complexity and temporal change in tissue systems[37]. To assess the performance of SpatialFuser in pairwise slice alignment and integration, we first applied it to two adjacent slices from the BaristaSeq mouse visual cortex dataset[9], comprising 1,525 and 2,042 spots, respectively (Fig. 3a). We evaluated spot-to-spot alignment accuracy

across the entire tissue section (see Methods), comparing SpatialFuser with three state-of-the-art methods reported in previous benchmarks[25], namely SPACEL[38], SLAT[23], and STAligner[22] (Fig. 3b). Owing to its use of ground-truth labels during training (see Methods), SPACEL achieved the highest alignment accuracy (0.987, 1,519 matches), followed closely by SpatialFuser (0.972, 1,516 matches) and STAligner (0.949, 1508 matches), while SALT showed noticeably lower performance (0.857, 1,525 matches).

An effective integration method should not only achieve accurate alignment but also preserve underlying biological patterns in the latent space. To this end, we further evaluated the spatial fidelity of the integrated embeddings. UMAP visualization showed that all four methods effectively removed batch effects (Fig. 3c). With original annotations used as supervision, SPACEL maintained strong region-level separation (Supplementary Fig. 4), but its embedding space appeared fragmented (Supplementary Fig. 5), suggesting limited intra-domain coherence despite minimal cross-domain mixing. STAligner and SLAT preserved the hierarchical structure of the six cortical layers to some extent, but layer boundaries were blurred and showed considerable mixing between adjacent layers. Compared to these methods, SpatialFuser produced embeddings with clearer expression patterns showed in UMAP (Fig. 3c & Supplementary Fig. 5) and more hierarchically organized, anatomically consistent domain structures that closely matched the ground truth (Supplementary Fig. 4).

To quantitatively compare the effectiveness of these methods in supporting integrative spatial heterogeneity analysis, we performed clustering on the joint embeddings from each pair of BaristaSeq slices using the Mclust algorithm, focusing primarily on the ARI before and after alignment and integration. Since SLAT and SPACEL operates only in multi-slice mode, ARI scores before integration were left unpopulated. The bar plot (Fig. 3d) shows that SpatialFuser demonstrated the best performance for post-integration, and the integration process enabled more accurate capture of spatial patterns, as clustering on the integrated embeddings yielded spatial domains more consistent with the ground truth (average ARI=0.86).

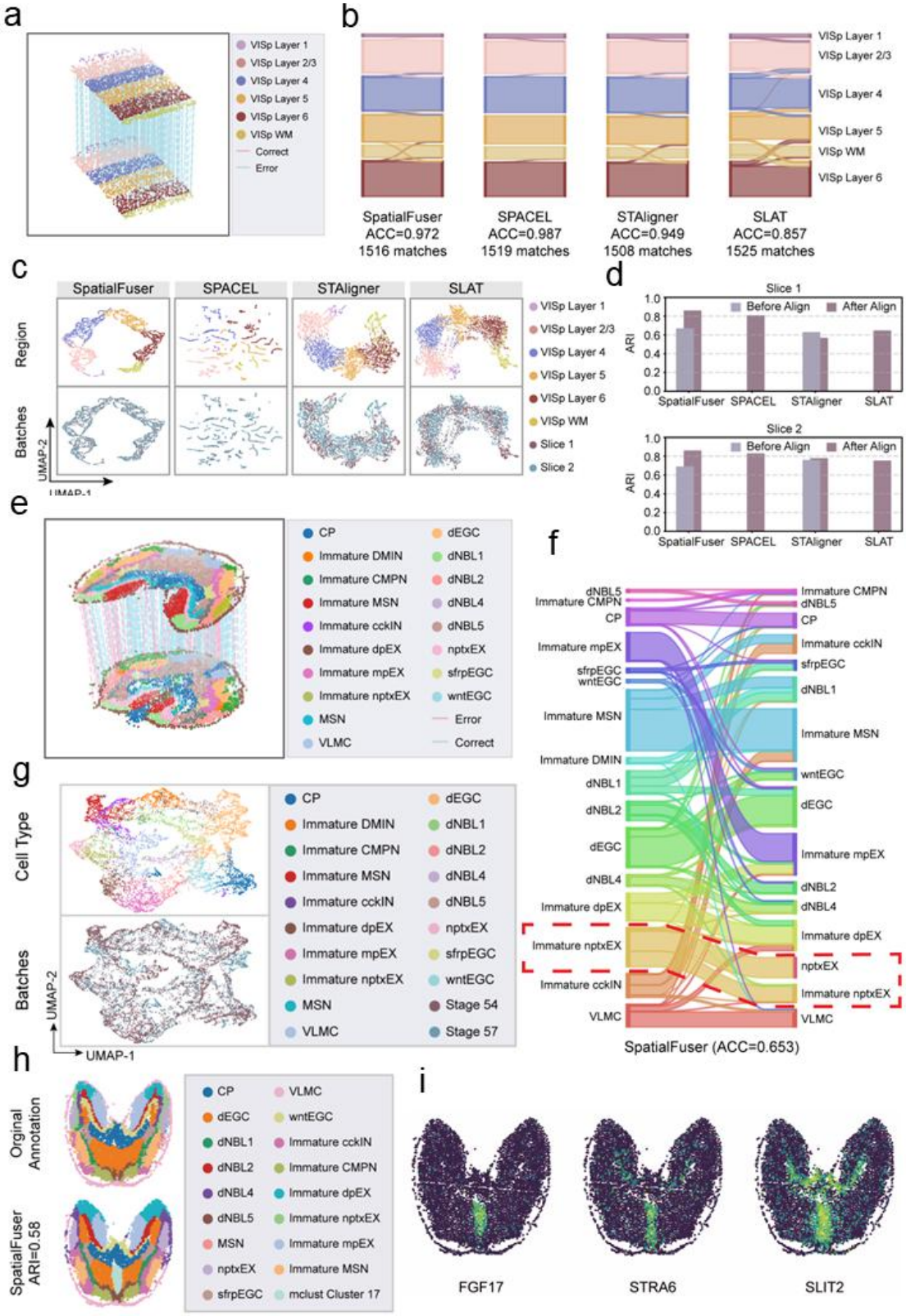


Fig. 3. Spatial alignment across consecutive tissue slices and non-consecutive slices from different developmental stages via SpatialFuser. a, Alignment of slices 1 and 2 from the BaristaSeq mouse visual cortex dataset, coloured by region labels (300 alignment pairs shown for clarity). **b,** Sankey plot showing region type correspondence based on alignments from SpatialFuser, SPACEL, STAligner, and SLAT for slices 1 and 2 of the BaristaSeq dataset. **c,**

UMAP visualizations of integration results for SpatialFuser, SPACEL, STAligner, and SLAT on slices 1 and 2 of the BaristaSeq dataset, coloured by region labels (top) and batch labels (bottom). **d,** ARI scores before and after integration for domain identification in slices 1 and 2 of the BaristaSeq dataset. **e,** Alignment of Stage 54 and Stage 57 slices from the Stereo-seq axolotl regenerative telencephalon dataset, coloured by region labels (300 alignment pairs shown for clarity). **f,** Sankey plot showing region type correspondence based on alignments from SpatialFuser for Stage 54 and Stage 57 slices of the Stereo-seq dataset. **g,** UMAP visualizations of integration results for SpatialFuser on Stage 54 and Stage 57 slices of the Stereo-seq dataset, coloured by region labels (top) and batch labels (bottom). **h,** Spatial domains in the Stage 57 slice of the Stereo-seq dataset identified by Mclust clustering on SpatialFuser-integrated embeddings. **i,** Expression distributions of FGF17, STRA6, and SLIT2, specifically enriched in a potential subregional structure within the dEGCs region of the Stage 57 slice from the Stereo-seq dataset.

We further evaluated SpatialFuser's efficacy in integrating non-consecutive slices from different developmental stages, which exhibit significant differences in cell-type composition and distribution. We employed a Stereo-seq[12] dataset representing axolotl regenerative telencephalon[10], which has manual annotations for each tissue region as reference, and applied SpatialFuser to align and integrate two slices from stage 54 and stage 57 (Fig. 3e). Through alignment, most spots were accurately matched despite substantial variation in spot numbers and types across slices (Fig. 3f), and dynamic changes in cell identities during development were successfully captured (e.g., mature nptxEX at stage 57 could be traced back to immature nptxEX at stage 54). In parallel, SpatialFuser effectively corrected batch effects while preserving expression pattern differences among distinct cell-types (Fig. 3g).

Interestingly, a large region of development-related ependymoglial cells (dEGCs) was enriched with less-aligned cells during the iterative training process of SpatialFuser (Supplementary Fig. 9). We hypothesize that this may be attributed to intra-regional heterogeneity within the seemingly homogeneous dEGCs domain. For instance, previously unannotated substructures may have emerged during telencephalon development from stage 54 to stage 57, reducing cross-slice similarity within this region and thereby hindering the formation of high-quality matches. To support this hypothesis, we performed Mclust clustering on the integrated embeddings and identified a distinct

subpopulation within the dEGCs domain (Cluster 17) (Fig. 3h), likely reflecting emerging subregional or functional differentiation. Further spatial differential analysis confirmed that this subregion exhibited a unique spatial expression pattern, with FGF17, STRA6, and SLIT2 significantly enriched (Fig. 3i). These genes are known to play key roles in embryonic development and neurodevelopmental processes[39-42], aligning well with the spatial distribution and potential functions of the subregion. Collectively, these findings demonstrate SpatialFuser's capability to reveal dynamic cellular changes through spatiotemporal integrative analysis.

Spatial multi-omics integration enables deep tissue profiling across modalities and resolutions

Incorporating multiple modalities into the analysis of biological samples holds great potential for deepening our understanding of the mechanisms underlying cellular and tissue organization[43]. A previous study have demonstrated that co-profiling and joint analysis of the epigenome and transcriptome from the stage E13 mouse embryo section enabled the identification of a neuronal cluster that could not be readily resolved by either modality alone[14]. Here, we firstly applied SpatialFuser to integrate spatial ATAC-RNA-seq derived from this section. Remarkably, Leiden clustering on the reconstructed features from each individual modality after integration successfully recovered the neuronal cluster (cluster 10 in both RNA and ATAC modality Leiden clustering) (Fig. 4a), which was previously identifiable only through joint analysis, suggesting that the integrated representation captures complementary epigenetic and transcriptional signals to enhance the separability of subtle cell populations within each modality. This result highlights the value of SpatialFuser in jointly analysing multi-omics profiles to uncover subtle cellular heterogeneity that may remain undetected in unimodal analyses.

During the embryonic development stage E13, chromatin accessibility may not fully resolve all the cell types defined by transcriptional profiles, potentially due to lineage priming[26, 44], as some genes exhibit differing spatial signal patterns between ATAC and RNA modalities. For instance, Vax2 (ventral retina marker[45]), Kcnj10

(ventricular zone and astrocyte progenitor marker[46]), Pax6 (cortical progenitor and optic vesicle marker[47]), and Myt1l (hindbrain neuron marker[48]) displayed relatively low RNA expression in their respective regions at E13 in the mouse embryo, despite marked chromatin accessibility. By integrating chromatin accessibility with transcriptomic signals, SpatialFuser enhances the detection of weakly expressed yet epigenetically informative genes (Fig. 4b). However, it is important to note that while SpatialFuser functions as an integrative enhancement tool that leverages complementary information from different modalities to reconstruct a more comprehensive and biologically meaningful gene expression landscape, it does not explicitly differentiate whether low RNA expression arises from true lineage priming or technical limitations in transcript detection. Disentangling these causes requires further experimental validation. Nonetheless, SpatialFuser provides a valuable framework for generating refined hypotheses on transcriptional readiness within complex tissue architectures, facilitating the identification of novel cell subtypes or states, which is crucial for developmental research and early disease detection.

Compared to the one-to-one correspondence between the genome and transcriptome, the linkage between the proteome or metabolome and the transcriptome or epigenome is generally weak[49], as protein abundance and metabolite levels are influenced by multiple regulatory processes beyond gene expression, such as translation, post-translational modifications, and cellular environment[50, 51]. To explore cross-modality integration under such weakly correlated conditions, we applied SpatialFuser to integrate and align transcriptomic (MAGIC-seq[11]), proteomic (PLATO[15]), and metabolomic (MALDI-MSI[7]) data obtained from consecutive mouse cerebellum tissue slices. Despite the significant integration challenges posed by differences in spatial resolution, molecular characteristics, and signal sparsity across weakly linked modalities, SpatialFuser effectively captured shared spatial molecular patterns across the different omics layers and successfully corrected modality biases while preserving true biological variation, as evidenced by the spatial domains identified through Louvain clustering of the latent embeddings, which aligned well with the original anatomical annotations (Fig. 4c). Most spots across the tissue were well

aligned, with high alignment similarity scores (Supplementary Fig. 10) and matching accuracy (Supplementary Fig. 11, Transcriptome–Proteome: 0.84, Transcriptome–Metabolome: 0.75, Proteome–Metabolome: 0.86). It is noteworthy that manual annotation is affected by the granularity and distribution discrepancies across modalities. For example, although MAGIC-seq transcriptomic data and PLATO proteomic data share the same spatial coordinates, the matching rate of their original labels is only 83%, which undermines the reliability of using label-matching accuracy as a metric to evaluate spot-to-region mapping performance. By visualizing the embeddings with UMAP, we further demonstrate SpatialFuser's ability to accurately align the same cell types across different omics modalities, as spots of the same type are consistently co-localized in the embedding space, even under weak cross-modality cellular correspondence (Fig. 4d).

Due to differences in spatial resolution and barcode density, the MALDI-MSI slice captured 3,908 spots, which is significantly more than the 1,677 spots in the transcriptomic and proteomic datasets, thereby enabling finer-grained tissue domain annotations. Through cross-modality and cross-resolution integration and alignment, SpatialFuser was able to leverage the detailed histological distribution patterns in the metabolomic data to enhance the delineation of fiber tracts in the spatial transcriptomic data, which were previously difficult to resolve (Fig. 4e).

To assess whether the reconstructed transcriptomic profiles preserved biologically relevant spatial patterns, we evaluated the ability of known canonical oligodendrocyte marker genes (Cldn11[52], Mog[53], and Cnp[54]) to discriminate the originally annotated fiber tracts before and after integration. ROC curve analysis showed that the reconstructed features achieved higher AUC scores in distinguishing the fiber tracts defined by the original annotation (Fig. 4f), indicating improved concordance with expected spatial expression patterns. Notably, when using Louvain-identified Fiber tracts cluster as the reference, we observed even higher AUC scores (Fig. 4g). The strong spatial concordance between the reconstructed marker gene expression patterns and the refined anatomical delineation supports the biological validity of the finer-grained structure revealed by the SpatialFuser integration.

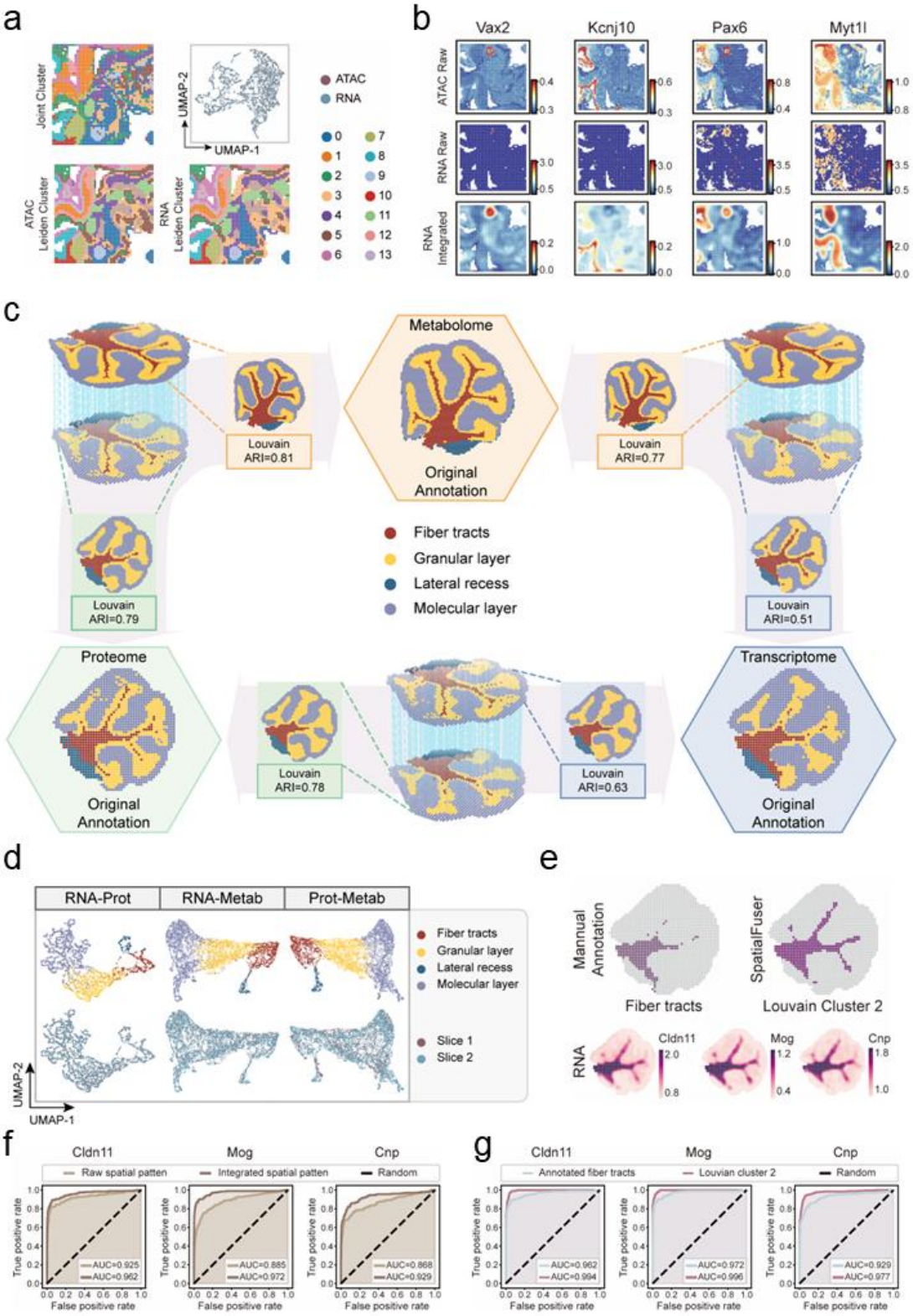


Fig. 4. SpatialFuser effectively integrates and aligns multimodal data from mouse embryo and consecutive mouse cerebellum tissue slices across epigenomic, transcriptomic, proteomic, and metabolomic. a, Spatial distributions identified by Leiden clustering on SpatialFuser-integrated spatial ATAC–RNA-seq data from an E13 mouse embryo, alongside UMAP visualizations coloured by modality. **b,** Spatial expression patterns

of Vax2, Kcnj10, Pax6, and Myt1l in the original ATAC, original RNA, and SpatialFuser-integrated RNA data. **c,** Pair-wise integration and alignment (300 alignment pairs shown for clarity) of transcriptomic (MAGIC-seq), proteomic (PLATO), and metabolomic (MALDI-MSI) data from consecutive mouse cerebellum tissue slices, with Louvain clustering for tissue domain detection. **d,** UMAP visualizations of modality bias correction results from SpatialFuser on pairwise integrations of the mouse cerebellum datasets, coloured by region labels (top) and batch labels (bottom). **e,** Distribution of the original Fiber tract annotation in the MAGIC-seq slice and the corresponding RNA–metabolome integration result (Louvain cluster 2), alongside spatial expression patterns of the Fiber tract marker genes Cldn11, Mog, and Cnp. **f,** ROC curves of marker gene expression before and after integration for distinguishing the original Fiber tract annotation. **g,** ROC curves of reconstructed marker gene expression distinguishing Louvain-detected Fiber tracts (cluster 2) and the original Fiber tract annotation.

Together, these results demonstrate the versatility of SpatialFuser in decoding complex biological systems through spatial multi-omics integration. By bridging modalities with different biological meanings and technical characteristics, SpatialFuser enables deeper and more comprehensive tissue profiling across resolutions, offering a powerful tool for spatial systems biology.

Discussion

Spatial technologies are expanding into multi-omics, enabling researchers to measure transcriptomics, epigenomics, proteomics, metabolomics, and other molecular layers at defined tissue locations. Underlying this is an unprecedented opportunity to integrate and interpret molecular information from diverse complementary perspectives for a deeper understanding of biological mechanisms within the microenvironment[55-57]. In this study, we introduce SpatialFuser, a deep learning-based unified framework designed for detailed molecular profiling within individual tissue sections and cross-modality, multi-sample integrative analysis of spatial multi-omics data. Benchmarking against real spatial transcriptomics datasets generated using diverse technologies, we demonstrate that SpatialFuser outperforms state-of-the-art methods in both spatial domain detection and consecutive slice alignment tasks. While each component of the framework can function independently, their synergistic combination within

SpatialFuser provides a holistic and streamlined solution for accurate spatial interpretation, robust cross-modality integration, and effective cross-slice alignment of spatial multi-omics data.

To accommodate multiple input modalities, SpatialFuser introduces MCGATE, the first unified representation module for multi-platform and multimodal spatial omics data. MCGATE effectively learns interpretable molecular representations under diverse single-omics scenarios. Unlike previous approaches that assume neighbouring cells share similar molecular profiles, MCGATE enhances local information propagation while enabling long-range feature dependency modelling in the embedding space and more accurate identification of spatially coherent patterns. Through ablation studies and interpretability analysis, we validate the effectiveness of the proposed multi-head collaborative attention mechanism in capturing precise spatial expression patterns. This design not only improves model transparency and reliability but also offers new perspectives for spatial data interpretation. However, the computational efficiency of this mechanism is currently limited by the lack of support for multi-channel sparse tensor operations in existing deep learning frameworks, particularly in memory-efficient parallel computation of high-dimensional sparse matrices, which poses challenges when applied to large-scale datasets[58, 59]. We anticipate that these computational bottlenecks will be alleviated with continued advances in hardware and software[60]. Moreover, we plan to upgrade SpatialFuser to incorporate more efficient sparse computation strategies that address current limitations in deep learning frameworks, thereby substantially supporting future challenges in whole-slice analysis at higher spatial resolutions. In fact, SpatialFuser remains generally fast. With a design specifically optimized for sparse computation in the single-head mode, the embedding generation process using MCGATE requires only ~30 seconds per DLPFC slice on an NVIDIA GeForce RTX 4090 GPU (24 GB).

To address diverse spatial multi-omics analysis scenarios, SpatialFuser incorporates multiple coordinate registration algorithms for cross-slice alignment, including the Iterative Closest Point (ICP) and Normal Distributions Transform (NDT). While ICP is widely used in spatial transcriptomics alignment[22, 23], NDT models spatial

characteristics via probabilistic grid fitting, enabling robust and effective registration of slices with densely packed, irregularly distributed spots and indistinct boundaries (e.g., in CODEX data; see Methods).

After coordinate pre-matching, SpatialFuser addresses integration and alignment in an iterative manner. By formulating the alignment task as a graph-based optimal transport problem, SpatialFuser employs a relaxed Sinkhorn algorithm to establish reliable anchors for subsequent contrastive learning–based integration. This strategy circumvents the need for assuming similar cell-type composition across samples, a limitation commonly present in earlier methods. Meanwhile, early works such as STAligner, SPACEL, and SpatialGlue typically require input data to be concatenated or to share equal feature dimensions, necessitating dimensionality reduction pre-processing that can lead to information loss, particularly when integrating datasets with substantial feature dimensionality differences (e.g., high-dimensional sequence-based transcriptomics and low-dimensional proteomics)[24]. Leveraging the scalability of MCGATE, SpatialFuser flexibly adjusts its network architecture to accommodate multi-modality and multi-platform data of varying complexity, efficiently and effectively mapping them into a shared latent space. Rather than relying on one-to-one molecular correspondences across modalities, SpatialFuser captures highly correlated spatial patterns of molecules in the embedding space to effectively correct for batch effects and modality biases.

Our experimental results demonstrated that SpatialFuser effectively captures complementary information across spatial epigenomics, transcriptomics, proteomics, and metabolomics, enabling both single-sample interpretation and joint analysis of multiple slices across different conditions or time points. In summary, SpatialFuser provides a unified and powerful framework for diverse spatial multi-omics analysis tasks. To facilitate broad adoption by the research community, we have released the SpatialFuser package along with detailed tutorials and demo cases at our GitHub repository https://github.com/liwz-lab/SpatialFuser.

Methods

Data Preprocessing

Filtering and normalization. SpatialFuser takes cellular information and spatial coordinate information of spatial multi-omics data as input, and we refer to the cells or the capture voxels in omics experiment as spots. We first filtered spots lacking valid annotations for reliable analysis. Spatial coordinates of each spot were normalized to $[0,1]$ to better incorporate spatial information into the model. The normalization of cellular information matrices was conducted using Scanpy package[24]. For each slice, we normalized the raw cellular matrix by total counts across all features, scaling each spot to 10,000 total counts, and applied a log-transformation to the normalized matrix for better variance stabilization and interpretability of expression differences. For dataset generated from sequencing-based and high-throughput AI-enhanced technology, we select the top 3000 spatial variable features.

Construction of graph inputs. SpatialFuser models given spatial multi-omics data as graph-structured inputs $\mathcal{G} = \{(x_i, s_i), i = 1,2,\cdots,N\}$, where N is the number of spots within slice, $x_i \in \mathbb{R}^d$ is d pre-processed feature of spot i, and S_i is adjacency spots of spot i constructed by K-Nearest Neighbors algorithm (KNN) based on spatial coordinates. We adjust the parameter K adaptively based on spatial domain heterogeneity and connectivity patterns across different multi-omics scenarios and platforms.

MCGATE for multi-omics data

Encoder. Given input $\mathcal{G}$, the k-th encoder layer $(k \in \{1,2,\cdots,L\})$ of spot i first adaptively learns the pairwise relationship between its adjacent spots j, represented by an edge weight in the graph structure. The edge weight e_{ij}^k at layer k is computed as:

$$e_{ij}^k = Sigmoid\left(V_1^{k^T}\left(W_1^k x_i^{k-1}\right) + V_2^{k^T}\left(W_1^k x_j^{k-1}\right)\right), \quad (1)$$

where W_1 is a learnable weight matrix, V_1 and V_2 are trainable weight vectors. To obtain the local attention weights for information aggregation, the edge e_{ij}^k weights are normalized using a sparse $Softmax$ function over S_i of spot i:

$$att_{ij}^k = Softmax_{j \in S_i}(e_{ij}^k). \tag{2}$$

Based on local attention weights, the encoder aggregates information from S_i of spot i to compute the smoothed representation x'^k as:

$$x'^k = BN\left(\sum_{j \in S_i} att_{ij}^k \left(W_1^k x_j^{k-1}\right)\right), \tag{3}$$

where BN denotes batch normalization[61], which stabilizes training and accelerates convergence[62].

Finally, a feedforward transformation with residual connection is applied to introduce nonlinearity and enhance the model's representational capacity:

$$x_i^k = BN\left(\sigma\left(W_2^k x'^k + b^k\right) + x'^k\right), \tag{4}$$

where W_2 is a learnable weight matrix, b is a trainable bias, and σ is a customized activation function. The encoder takes the pre-processed feature vector x_i^0 as input, and the output of the final layer x_i^L serves as the low-dimensional feature embedding learned by MCGATE.

Adaptive Adjacency Expansion Layer. Based on the assumption that spatially adjacent cells exhibit high feature similarity, previous methods typically rely on convolutional or neighborhood-based attention mechanisms for information aggregation. However, our experiments show that in spatial omics data, the spot pairs with the highest feature similarity are often not spatially adjacent. This suggests that relying solely on local features as regularization may limit the model's ability to capture true spatial expression patterns, highlighting the importance of long-range information propagation for accurate embedding learning. However, computing global attention within the encoder becomes increasingly expensive as the number of spots grows, making it impractical for large-scale slices learning. To address this, we introduce a

collaborative attention mechanism that efficiently models spatial expression patterns in the low-dimensional embedding space.

The adaptive adjacency expansion layer first computes cosine similarity between spot embeddings as follow:

$$\begin{aligned} Cos_{ij}^{local} &= Softmax_{S_i}\left(\frac{x_i^L \cdot x_j^L}{\|x_i^L\| \cdot \|x_j^L\|}\right), \\ Cos_{ij}^{long} &= Softmax_{S_i^c}\left(\frac{x_i^L \cdot x_j^L}{\|x_i^L\| \cdot \|x_j^L\|}\right), \end{aligned} \tag{5}$$

where S_i denotes the spatial neighborhood of spot i, and S_i^c denotes the set of non-spatial neighbors.

The local cosine similarity matrix Cos_{ij}^{local} is fed back to the encoder to enhance the local attention mechanism previously defined in Equation (2):

$$att_{local_{ij}}^k = Softmax_{S_i}\left(Cos_{ij}^{local} * e_{ij}^k\right), \tag{6}$$

where e_{ij}^k is the attention coefficient computed at k-th encoder layer.

For each spot i, the adaptive adjacency expansion layer selects the mutual nearest neighbors[63] (MNNs) spots with highest similarity outside its spatial neighborhood from Cos_{ij}^{long}, generating a sparse long-range adjacency matrix that defines reliable pathways for long-range information propagation. during training process. Let G_i denotes the set of long-range neighbors for spot i, the long-range attention weights are computed as:

$$att_{long_{ij}}^k = Softmax_{G_i}\left(e_{ij}^k\right). \tag{7}$$

The collaborative attention weights are then defined as the linear addition of local and long-range attention:

$$att_{co_{ij}}^k = \alpha\, att_{long_{ij}}^k + (1-\alpha) att_{local_{ij}}^k, \tag{8}$$

where $\alpha \in [0,1]$ (default $\alpha = 0$) is a slice-specific hyperparameter controlling the relative influence of long-range attention during information aggregation. By integrating long-range and local collaborative attention, the adaptive adjacency expansion layer enables efficient global information propagation and aggregation,

thereby enhancing MCGATE's ability to learn meaningful spatial expression patterns in spatial omics data.

Decoder. Symmetric to the encoder, the k-th decoder layer ($k \in \{1,2,\cdots,L\}$) reconstructs the representation of spot i as follows:

$$\hat{x}_i^{k-1} = \sigma\left(\widehat{W}_1^k\left(\sum_{j\in S_i} \widehat{att}_{co_{ij}}^k \hat{x}_j^k\right)\right). \tag{9}$$

The decoder takes the low-dimensional feature embedding x^{L} as input and outputs the reconstructed low-noise representation $\hat{x}_i^0$, which approximates the original data.

To reduce the risk of overfitting, MCGATE adopts a parameter sharing strategy between the encoder and decoder. Specifically, for the k-th decoder layer, we set $\widehat{W}_1^k = W_1^{k^T}$ and $\widehat{att}_{co_{ij}}^k = att_{co_{ij}}^k$ to share both transformation matrices and attention weights. However, the decoder does not replicate the feedforward structure of the encoder, in order to prevent overly rigid architectural symmetry, which may constrain the model's expressiveness.

The training of MCGATE is driven by a reconstruction objective, which optimizes the low-dimensional embeddings through minimization of the mean squared error (MSE) loss between input and decoder output:

$$Loss_{rec} = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i^0 - \hat{x}_i^0\right\|_2, \tag{10}$$

where N is the total number of spots.

Multi-head Attention. Inspired by previous works[64], MCGATE incorporates multi-head attention mechanism to enable more fine-grained information aggregation by partitioning input features into multiple sub-space. In the k-th encoder layer ($k \in \{1,2,\cdots,L\}$), the h-th collaborative attention weight ($h \in \{1,2,\cdots,H\}$) between spots i and j, denoted $att_{co_{ij}}^{k,h}$, is computed in parallel using independent parameters $W_1^{k,h}$, $V_1^{k,h}$, and $V_2^{k.h}$. Notably, in the adaptive adjacency expansion layer, cosine similarity is also computed separately in each subspace, ensuring that feature aggregation in each head captures the diversity of underlying feature structures.

Under the multi-head attention paradigm, the overall information aggregation processes in both encoder and decoder is realized by concatenating the outputs of multiple parallel attention heads. The aggregation formula in the encoder (extending Equation 3) is:

$$x'^{k} = BN\left(Concat_{h=1}^{H}\left(\sum_{j\in S_i} att_{co_{ij}}^{k,h}\left(W_1^{k,h}x_j^{k-1,h}\right)\right)\right). \tag{11}$$

Similarly, the decoder propagation process (extending Equation 9) becomes:

$$\hat{x}_i^{k-1} = \sigma\left(Concat_{h=1}^{H}\left(\widehat{W}_1^{k,h}\left(\sum_{j\in S_i} \widehat{att}_{co_{ij}}^{k,h}\,\hat{x}_j^{k,h}\right)\right)\right). \tag{12}$$

Due to the lack of efficient sparse multi-channel tensor operations in deep learning frameworks (e.g., PyTorch[64]), the multi-head attention mechanism often leads to high GPU memory usage. To ensure scalability on large spatial multi-omics datasets, MCGATE adopts sparse tensor operations in the single-head mode to effectively reduce memory usage and computational cost during training.

Matching layer

SpatialFuser models the alignment of spatial omics slices as a graph matching problem, which is essentially a classic optimal transport task. The core objective is to find an optimal coupling matrix by minimizing the Wasserstein distance[65] between embedding feature distributions. Let the low-dimensional embeddings of the two slices be $E_1 \in \mathbb{R}^{n\times d'}$ and $E_2 \in \mathbb{R}^{m\times d'}$, SpatialFuser introduces spatial proximity constraints to define a similarity measure across slices:

$$Simi(i,j) = \begin{cases} \dfrac{\dfrac{E_1[i]\cdot E_2[j]}{\|E_1[i]\|\cdot\|E_2[j]\|}+1}{2\tau+10^{-6}}, & if\ \|Coo_1[i]-Coo_1[j]\|_2 \le \mathrm{r} \\ 0, & otherwise \end{cases} \tag{13}$$

where the indices $i \in \{1,2,\cdots,n\}$ and $j \in \{1,2,\cdots,m\}$ denote arbitrary spots on each slice, τ is a temperature coefficient, $Coo_1[i]$ and $Coo_2[j]$ represent their

corresponding spatial coordinate vectors, and r denotes the spatial neighborhood radius, restricting the cross-sample matching to a local spatial window.

Based on the similarity matrix defined in Equation (13), SpatialFuser formulates the alignment task as the following optimal transport problem:

$$\underset{T\in\mathcal{T}}{\operatorname{argmax}} \sum_{i=0}^{n}\sum_{j=0}^{m} Simi(i,j)T_{ij} + \lambda \cdot Reg(T), \tag{14}$$

where $T \in [0,1]^{n\times m}$ is the soft coupling matrix to be optimized. The set $\mathcal{T} = \{T \in \mathbb{R}^{n\times m} \,|T1_m \leq 1_n, T^T 1_n \leq 1_m\}$ defines the feasible region. $Reg(\cdot)$ denotes the entropy regularizer, and λ is its weighting factor. SpatialFuser employs the Sinkhorn algorithm[66] to solve this linear program via efficient iterative updates.

While Equation (14) implicitly assumes that every spots in the two slices can be reliably matched, heterogeneous spatial omics datasets often differ in cell-type composition and batch size. A robust alignment algorithm should therefore tolerate unmatched spots. To address this, SpatialFuser introduces a relaxed mapping strategy by augmenting the original similarity matrix $Simi \in \mathbb{R}^{n\times m}$ defined in Equation (13) with additional dustbin channels (Supplementary Fig), assigning them predefined weights as follows:

$$\overline{Simi}(i,j) = \begin{cases} Simi(i,j), & for\ \mathrm{i} \leq \mathrm{n}, \mathrm{j} \leq \mathrm{m} \\ S_{bin}, & \text{otherwise} \end{cases} \tag{15}$$

where $i \in \mathbb{Z}_{[0,n+1]}$ and $j \in \mathbb{Z}_{[0,m+1]}$, $\overline{Simi}$ is the augmented similarity matrix with the dustbin channels, and S_{bin} is a constant weight (default $S_{bin} = 1$) assigned to the dustbin.

The dustbin channel strategy is a form of relaxation, which essentially converts the inequality-constrained problem in Equation (14) into an equality-constrained problem with an expanded feasible region:

$$\bar{\mathcal{T}} = \{\bar{T} \in \mathbb{R}^{(n+1)\times(m+1)} \,|\bar{T}1_{m+1} = [1_n^T, m]^T \ , T^T 1_n = [1_m^T, n]^T\}. \tag{16}$$

By allowing unmatched or low-confidence points to be assigned to the dustbin channels, the model can adaptively distinguish meaningful alignments from noisy ones, thus relaxing the strong assumption of compositional similarity between slice and avoiding forcible alignment.

To ensure sufficient supervision during early training and further improve matching quality, SpatialFuser introduces a Mutual Top-K Neighbors (MKNs) strategy based on the soft assignment matrix generated by the Sinkhorn algorithm. Specifically, the model initially selects MNNs from the soft couplings to construct high-confidence matches. However, strong batch effects in the early stages may result in insufficient MNNs. To address this, the matching criteria are relaxed to include mutual top-K neighbors, enhancing the supervision signal and ensuring stable training. These MKNs serve as anchors to guide the fusion layer.

Fusion layer

In the fusion layer design, a shared multilayer perceptron (MLP) is connected to two independent MCGATE modules to project the low-dimensional embeddings of spatial omics data from different experimental platforms and modalities into a unified latent space. To prevent the model from memorizing batch-specific patterns, the embeddings are first concatenated and then randomly permuted along the row prior to being input into the MLP, encouraging it to learn a generalizable mapping function.

The fusion layer employs contrastive learning for spatial multi-omics data integration. Specifically, high-confidence cross-slice matches derived from the matching layer serve as anchor-positive pairs to guide the correction of batch effects or modality biases. To enhance cell-type discriminability, negative samples are randomly drawn from non-neighboring regions within the same tissue slice. Given that the number of spatial spots typically far exceeds the local neighborhood size, this strategy is both efficient and effective in minimizing the risk of false negatives.

The training process employs an improved triplet loss[29] as the modality fusion objective, simultaneously promoting the clustering of similar samples (anchor-positive) and the separation of dissimilar ones (anchor-negative), while enforcing both absolute and relative distance constraints to improve training stability:

$$Loss_{fuse} = \left\|E_i - E_p^i\right\|_2 + \frac{1}{N_a}\sum_{i=1}^{N_a} \max\left\{\left\|E_i - E_p^i\right\|_2 - \left\|E_a^i - E_n^i\right\|_2 + \tau, 0\right\}, \quad (17)$$

where E_i is the fused embedding of spot i, serving as anchor, E_p^i and E_n^i denote the embeddings of the corresponding positive and negative spots respectively, and τ is a margin parameter. To prevent embedding collapse and preserve spatial representational fidelity, a reconstruction loss is incorporated into the integration process. Meanwhile, to address potential inconsistencies caused by the dual independent encoder design of SpatialFuser, we introduce an embedding direction constraint loss $Loss_{dir}$ to align the global orientation of the representations across slices:

$$Loss_{dir} = 1 - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{E_1[i]\cdot E_2[j]}{\|E_1[i]\|\cdot\|E_2[j]\|}. \tag{17}$$

The total objective function of the integration process combines the fusion objective, the reconstruction loss, and the embedding direction constraint loss, and is formulated as:

$$Loss_{integrate} = Loss_{fuse} + \beta_{rec}\, Loss_{rec} + \beta_{dir} Loss_{dir}, \tag{18}$$

where β_{rec} and β_{dir} are the weight of the reconstruction loss and the as embedding direction constraint loss, respectively (default $\beta_{rec} = 1$, $\beta_{dir} = 0.1$). The embeddings integrated by the fusion layer are used in the next iteration to guide the matching layer in identifying more accurate cross-slice correspondences, enabling progressively improved alignment and integration across training rounds.

Coordinate Pre-matching Modules (Optional)

Spatial registration of spot coordinates is a key preparatory step for aligning and integrating spatial omics slices. SpatialFuser formulates spatial omics data as 2D point clouds and corrects coordinate discrepancies via rigid transformations defined by a rotation matrix M and a translation t .The coordinate registration problem is formulated as the following optimization objective:

$$argmin_{M,t}\, \|Y_1 - Y_2 \cdot M - t\|^2, \tag{19}$$

where $Y_1 \in \mathbb{R}^{n\times 2}$ and $Y_2 \in \mathbb{R}^{m\times 2}$ denote the spatial coordinate matrices of the two datasets to be aligned, with n and m are the numbers of spots in each set.

SpatialFuser provides multiple solution strategies to accommodate different scenarios. The ICP algorithm is a classical method for solving point cloud registration problems[67, 68]. It first establishes spot correspondences by finding the nearest neighbor for each spot in one slice, and then addresses the above optimization problem via singular value decomposition, iteratively updating the rotation matrix M and translation vector t by minimizing the Euclidean distance between matched spot pairs. The ICP algorithm has been widely applied to spatial transcriptomics slice alignment, particularly in cases where the slices display well-defined structural grids or regular anatomical boundaries. However, ICP is not consistently reliable on slices with uneven spot distributions. To address this limitation, SpatialFuser also incorporates the NDT algorithm[69], which models the reference slice by partitioning it into a 2D grid. Each grid cell $V_{i,j}$ contains the spots within its bounds, and a multivariate Gaussian distribution $\mathcal{N}(\mu_{i,j}, \Sigma_{i,j})$ is fitted to the spots in each non-empty grid cell. Instead of relying on pairwise Euclidean distances, NDT evaluates the Mahalanobis distance between each transformed spot $z_t \in \mathbb{R}^2$ and the Gaussian distribution of the corresponding grid cell in the reference slice. The coordinate pre-matching is achieved by minimizing the cumulative deviation:

$$\min_{R,t} \sum_{k=1}^{K} \left(z_k - \mu_{i,j}\right)^T \Sigma_{i,j}^{-1} \left(z_k - \mu_{i,j}\right), \tag{20}$$

where K is the total spot number of the transformed slice. This probabilistic approach enables more robust registration in scenarios involving sparse, noisy, or irregular spot distributions[70].

Whole-slice Alignment

We propose a whole-slice alignment strategy to support downstream analysis and performance comparison against existing alignment methods. By default, the slice with fewer spatial spots is mapped to the one with more spots. For each spot i in the source slice, a candidate set is first identified in the target slice by selecting all spots within a spatial distance radius r, and we further compute the cosine similarity between spot i and its candidate spots based on the fused embeddings. Inspired by previous work[23], we introduce a probabilistic quality control mechanism. Rather than randomly pairing

spots across both slices, we randomly sample 1,000 spots from the source slice and, for each sampled spot, randomly select a nearby spot within radius r in the target slice to compute cosine similarities, in order to form a stricter null distribution. The 95th percentile of this null distribution is used as the threshold for quality control. In scenarios where the source and target slices have the same resolution, we retain the candidate with the highest similarity score as the final alignment result for each source spot.

Implementation details of SpatialFuser

Spatial Domain Detection Task. To accommodate datasets with varying complexity, SpatialFuser adopts different MCAGATE configurations. For high-throughput data such as 10X Visium, Stereo-seq, and MAGIC-seq, a two-layer MCAGATE is used, with hidden dimensions set to 512 and 32 respectively. For low-throughput data such as CODEX and MALDI-MSI, a single-layer MCAGATE with 32 hidden units is employed. The MCAGATE is trained using the Adam optimizer, with the ELU function selected as the activation. The model is trained for 500 steps by default, during which the Adaptive Adjacency Expansion Layer is updated every 100 epochs to refine long-range relationships and their associated attention weights. Both the learning rate and the long-range attention coefficient α are customized for each dataset.

Alignment and Integration Task. The training process is divided into two stages: pre-training and contrastive learning. In the pre-training stage, a fusion layer is connected to two independent MCAGATE encoders and jointly trained with a reconstruction loss for 500 epochs (default), producing high-quality embeddings for downstream alignment and integration. For slices with semantically similar molecular profiles, shared features across slices are used. The matching layer and the fusion layer are then iteratively optimized to align the embeddings into a shared latent space. Specifically, the matching layer first identifies anchor-positive pairs based on the pre-trained embeddings. These identified pairs are then used to guide the fusion layer in integrating the modalities. By default, triplet pairs are updated every 20 epochs, and this iterative process continues until 500 steps. The learning rate is dataset-specific.

Clustering, Differential expression analysis, and Spatial trajectory inference

Clustering was performed by constructing a shared nearest neighbor graph based on the spatially resolved data, followed by cluster identification using the Louvain[33], Leiden[32], or Mclust[31] algorithm. The default resolution parameter used in benchmarking for the Louvain and Leiden clustering algorithms was set to 0.1, while for other downstream analyses it was adjusted to the optimal value for each case. Cluster biomarkers and differentially expressed genes between groups were determined using the Wilcoxon rank-sum test, as implemented in the rank_genes_groups() function of the Scanpy[71] package. To infer spatial trajectories, we applied partition-based graph abstraction via the paga() function in Scanpy.

Spatial distribution detection across diverse experimental platforms and modalities

Data description. The 10X Visium DLPFC dataset[30] comprises 12 tissue sections from 3 individuals, providing whole-transcriptome coverage with 33,438 gene features. Each section contains between 3,498 and 4,789 spatially resolved spots. The mouse somatosensory cortex dataset[2], generated using the osmFISH platform, includes 5,328 spots at single-cell resolution, based on the targeted detection of 33 genes. The CODEX human bladder cancer dataset[4] consists of 75 tumor tissue sections from 31 patients, profiling 35 protein markers across nuclear and membrane compartments. After quality control, ~360,000 epithelial cells, ~140,000 immune cells, and ~90,000 stromal cells were retained. A 1,440-dimensional feature matrix was initially constructed using statistical summaries (e.g., percentages and quantile distributions) of protein expression. For the detection of cell type spatial distributions, the 210308_TMA2_reg6 slice was selected, and the average nuclear and membrane expression levels of each protein were extracted as input features. All datasets include detailed annotations of tissue regions or cell types, which serve as ground truth for evaluation.

Hyperparameter settings. The hyperparameters for SpatialFuser were optimized through grid search. For all baseline methods, we adopted the recommended data pre-

processing steps and hyperparameter settings as specified in their official implementations and documentation for DLPFC dataset, except for STAligner. Since STAligner's embedding model structure is very similar to that of STAGATE, we referred to the documentation of STAGATE for its parameter settings. The documentation links for these methods are as follows:

SpaGCN: https://github.com/jianhuupenn/SpaGCN/blob/master/tutorial/tutorial.md

STAGATE: https://stagate.readthedocs.io/en/latest/index.html

SEDR: https://sedr.readthedocs.io/en/latest/index.html

Hyperparameter robustness. We tested SpatialFuser's robustness to key hyperparameters of MCGATE including: (1) number of encoder and decoder layers, (2) dimension of layers, (3) dimension of MCGATE embedding. We ran MCGATE on all 12 slices of DLPFC dataset and every experiment was run with different random seeds. To identify a suitable attention architecture for the DLPFC dataset, we benchmarked different numbers of attention heads based on a selected network structure from the robustness testing.

Spatiotemporal Alignment and integration

Data description. The BaristaSeq mouse visual cortex dataset[9] comprises three tissue slices with similar domain distributions and morphological structures, containing 4,491, 3,545, and 3,390 spatial spots, respectively. A total of 78 transcript features were retained, along with detailed anatomical region annotations. Slices 1 and 2 were selected for integration and alignment. The Stereo-seq axolotl regenerative telencephalon dataset[10] includes 18 brain tissue slices with single-cell resolution, capturing over 20,000 genes across various stages of regeneration and development. Two slices from developmental stages 54 and 57, containing 2,929 and 4,410 spatial spots respectively, were selected for alignment and integration. The original annotations were used as references for evaluation.

Hyperparameter settings. The hyperparameters for SpatialFuser were optimized through grid search, while the hyperparameter settings for all other methods followed the official guidelines and tutorials provided by their respective authors:

SALT: https://slat.readthedocs.io/en/latest/tutorials.html

SPACEL: https://spacel.readthedocs.io/en/latest/index.html

STAligner: https://staligner.readthedocs.io/en/latest/index.html

To ensure a fair comparison of the integration and alignment performance on consecutive BaristaSeq slices, we conducted a grid search for several key parameters related to both graph construction and network architecture, which are critical yet not explicitly specified in the original documentation of the compared methods. Specifically, SPACEL constructs the spatial graph using 8-nearest neighbors and employs a two-layer Spline, with a hidden layer size of 256 and an output embedding dimension of 16 (Supplementary Fig. 6). STAligner builds the graph using 20-nearest neighbors and adopts a two-layer encoder with node dimensions set to 64 and 30, respectively (Supplementary Fig. 7). For contrastive learning-based integration process, the number of nearest neighbors used when constructing MNNs was set to 10. Notably, while typical GCNs tend to suffer from over-fitting, gradient vanishing, or over-smoothing beyond 4–5 layers, SLAT consistently demonstrated improved performance with increased depth, achieving its best results with over 20 LGCN layers (Supplementary Fig. 8). Therefore, we report SLAT's performance using 20 layers GCN with mlp_hidden = 128 and hidden_size = 64 in the experiments on the BaristaSeq dataset. Before training, SLAT constructs a KNN graph with K=10.

Spatial Multi-Omics Alignment and integration

Data description. The spatial ATAC-RNA-seq dataset[14] of the E13 mouse embryo comprises 5,328 spots with genome-wide co-profiling of epigenome and transcriptome, enabling simultaneous measurement of approximately 24,017 accessible chromatin regions and 15,748 messenger RNA expression. Manual annotation of anatomical regions was not provided by the authors. Instead, the dataset includes eight major ATAC-defined clusters, 14 RNA-defined clusters, and 14 integrative clusters based on joint analysis of spatial ATAC and RNA data. The multi-omics mouse cerebellum dataset[15] comprises three consecutive tissue slices, each representing a distinct spatial omics modality: MAGIC-seq for spatial transcriptomics, PLATO for spatial proteomics,

and MALDI-MSI for spatial metabolomics. The MAGIC-seq slice captured 1,677 spatial spots at a resolution of 32 μm, profiling 16,116 genes. The PLATO slice provides high-throughput proteomic profiling aligned with the transcriptomic slice, identifying 5,722 protein groups after AI enhancement and quality control. The MALDI-MSI slice achieves higher spatial resolution, capturing 3,908 spots with 491 metabolite peaks. Spatial domain annotations are available for all three slices, though slight differences exist in domain distributions across modalities. Notably, the slices with different resolutions exhibit significant discrepancies in spatial coordinates.

Hyperparameter settings. For the high-throughput spatial ATAC-RNA-seq data, we constructed a 4-nearest neighbor graph using the top 3,000 spatially variable features, which were normalized and log-transformed before being used as input. The number of layers in MCGATE was set to two, with hidden dimensions of 512 and 32, respectively. We first pre-trained two 4-head MCGATE model separately on the ATAC and RNA modalities for 500 epochs using a learning rate of 1e-3 and 1e-4. Further, we set the reconstruction loss weight $\beta_{rec} = 50$ and the directional loss weight $\beta_{dir} = 0.1$ and trained the fusion and matching layers for 500 steps with a learning rate of 1e-4 to achieve integration and alignment. In the experiments on the multi-omics mouse cerebellum dataset, we set K=4 for KNN graph construction, and employed the NDT algorithm for coordinate registration prior to training. During the pretraining stage, we extracted the top 3,000 spatially variable features from both the MAGIC-seq transcriptomic slice and the PLATO proteomic slice. These features were normalized and log-transformed before being input into separate two-layer MCGATE models with hidden dimensions of 512 and 32. The number of attention heads was set to 4, and the models were trained for 500 steps using the Adam optimizer with learning rates of 3e-3 and 1e-3, respectively. Considering the relatively low throughput of the MALDI-MSI metabolomic data, we used all available features as input and employed a single-layer MCGATE model with a 32-dimensional embedding and four attention heads, trained for 500 steps with a learning rate of 5e-4. In the integration and alignment stage, we set $\beta_{rec} = 50$ and $\beta_{dir} = 0.1$, and trained the fusion and matching layers for 200 steps using the Adam optimizer with a learning rate of 3e-3.

Data Availability

All experimental data of this study have already been published and are accessible within the corresponding articles and public repositories. Specifically, the 10X Visium DLPFC dataset is available in the spatialLIBD package[72] (http://spatial.libd.org/spatialLIBD); the osmFISH mouse somatosensory cortex dataset is available at http://linnarssonlab.org/osmFISH/; the CODEX human bladder cancer dataset is available from the Aquila database[73] (https://aquila.cheunglab.org/); the BaristaSeq mouse visual cortex dataset is available from the SODB database[74] (https://gene.ai.tencent.com/SpatialOmics/); the Stereo-seq axolotl regenerative telencephalon dataset is available from the STOmicsDB database[75] (https://db.cngb.org/stomics/artista/); the spatial ATAC–RNA-seq E13 mouse embryo dataset is available at https://ki.se/en/mbb/oligointernode; and the spatial multi-omics mouse cerebellum dataset is available from the original article.

Code Availability

The SpatialFuser framework was implemented in the "spatialFuser" Python package, which is open-source and accessible at https://github.com/liwz-lab/SpatialFuser.

Competing interests

To maximize the impact of this study, Sun Yat-sen University has submitted a patent application to the State Intellectual Property Office of Chia (SIPO).

Acknowledgements

This work was supported by the grants of National Natural Science Foundation of China (92474107), National Key R&D Program of China (2021YFF1200903), Guangdong Basic and Applied Basic Research Foundation of China (2022B1515120077), and Major Project of Guangzhou National Laboratory of China (GZNL2024A01003).

References

[1] Qiu X, Zhu D Y, Lu Y, et al. Spatiotemporal modeling of molecular holograms. Cell, 2024, 187(26): 7351-7373. e61

[2] Codeluppi S, Borm L E, Zeisel A, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nature Methods, 2018, 15(11): 932-935

[3] Zhang M, Eichhorn S W, Zingg B, et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. Nature, 2021, 598(7879): 137-143

[4] Black S, Phillips D, Hickey J W, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nature Protocols, 2021, 16(8): 3802-3835

[5] Ptacek J, Locke D, Finck R, et al. Multiplexed ion beam imaging (MIBI) for characterization of the tumor microenvironment across tumor types. Laboratory Investigation, 2020, 100(8): 1111-1123

[6] Williams C, Reeves J W, Danaher P, et al. Spatial insights into tumor immune evasion illuminated with 1000-plex RNA profiling with CosMx Spatial Molecular Imager. Cancer Research, 2023, 83(6765): 10

[7] Rohner T C, Staab D, Stoeckli M. MALDI mass spectrometric imaging of biological tissue sections. Mechanisms of Ageing and Development, 2005, 126(1): 177-185

[8] Ståhl P L, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 2016, 353(6294): 78-82

[9] Chen X, Sun Y-C, Church G M, et al. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. Nucleic Acids Research, 2018, 46(4): e22-e22

[10] Wei X, Fu S, Li H, et al. Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. Science, 2022, 377(6610): eabp9444

[11] Zhu J, Pang K, Hu B, et al. Custom microfluidic chip design enables cost-effective three-dimensional spatiotemporal transcriptomics with a wide field of view. Nature Genetics, 2024, 56(10): 2259-2270

[12] Liao S, Heng Y, Liu W, et al. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. bioRxiv, 2023: 2023.04. 28.538364

[13] Deng Y, Bartosovic M, Ma S, et al. Spatial profiling of chromatin accessibility in mouse and human tissues. Nature, 2022, 609(7926): 375-383

[14] Zhang D, Deng Y, Kukanja P, et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. Nature, 2023, 616(7955): 113-122

[15] Hu B, He R, Pang K, et al. High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning. Cell, 2025, 188(3): 734-748. e22

[16] Sun C, Wang A, Zhou Y, et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. Nature Communications, 2023, 14(1): 2692

[17] Chen A, Liao S, Cheng M, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell, 2022, 185(10): 1777-1792. e21

[18] Li H, Bao S, Farzad N, et al. Spatially resolved genome-wide joint profiling of epigenome and transcriptome with spatial-ATAC-RNA-seq and spatial-CUT&Tag-RNA-seq. Nature protocols, 2025: 1-35

[19] Hu J, Li X, Coleman K, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature Methods, 2021, 18(11): 1342-1351

[20] Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nature Communications, 2022, 13(1): 1739

[21] Xu H, Fu H, Long Y, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Medicine, 2024, 16(1): 12

[22] Zhou X, Dong K, Zhang S. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. Nature Computational Science, 2023, 3(10): 894-906

[23] Xia C-R, Cao Z-J, Tu X-M, et al. Spatial-linked alignment tool (SLAT) for aligning heterogenous slices. Nature Communications, 2023, 14(1): 7236

[24] Long Y, Ang K S, Sethi R, et al. Deciphering spatial domains from spatial multi-omics with SpatialGlue. Nature Methods, 2024, 21(9): 1658-1667

[25] Hu Y, Xie M, Li Y, et al. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. Genome Biology, 2024, 25(1): 212

[26] Ma S, Zhang B, LaFave L M, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell, 2020, 183(4): 1103-1116. e20

[27] Lais P, Mishra S, Xiong K, et al. Image guided construction of a common coordinate framework for spatial transcriptome data. Scientific Reports, 2025, 15(1): 18074

[28] Sarlin P-E, DeTone D, Malisiewicz T, et al. Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020

[29] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. Proceedings of the iEEE conference on computer vision and pattern recognition. 2016

[30] Maynard K R, Collado-Torres L, Weber L M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience, 2021, 24(3): 425-436

[31] Fraley C, Raftery A E, Murphy T B, et al., mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. 2012, Technical report.

[32] Traag V A, Waltman L, Van Eck N J. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, 2019, 9(1): 1-12

[33] Blondel V D, Guillaume J-L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008

[34] Gouin III K H, Ing N, Plummer J T, et al. An N-Cadherin 2 expressing epithelial cell subpopulation predicts response to surgery, chemotherapy and immunotherapy in bladder cancer. Nature Communications, 2021, 12(1): 4906

[35] Yuan Y. Spatial heterogeneity in the tumor microenvironment. Cold Spring Harbor Perspectives in Medicine, 2016, 6(8): a026583

[36] Wolf F A, Hamey F K, Plass M, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology, 2019, 20(1): 59

[37] Long Y, Ang K S, Li M, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nature Communications, 2023, 14(1): 1155

[38] Xu H, Wang S, Fang M, et al. SPACEL: deep learning-based characterization of spatial transcriptome architectures. Nature Communications, 2023, 14(1): 7603

[39] Hoshikawa M, Ohbayashi N, Yonamine A, et al. Structure and expression of a novel fibroblast growth factor, FGF-17, preferentially expressed in the embryonic brain. Biochemical and Biophysical Research Communications, 1998, 244(1): 187-191

[40] Dhokia V, Macip S. A master of all trades–linking retinoids to different signalling pathways through the multi-purpose receptor STRA6. Cell Death Discovery, 2021, 7(1): 358

[41] Xavier-Neto J, Costa Â M S, Figueira A C M, et al. Signaling through retinoic acid receptors in cardiac development: Doing the right things at the right times. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2015, 1849(2): 94-111

[42] Hu H. Chemorepulsion of neuronal migration by Slit2 in the developing mammalian forebrain. Neuron, 1999, 23(4): 703-711

[43] Li Z, Cui X, Chen X, et al. Cross-modality representation and multi-sample integration of spatially resolved omics data. bioRxiv, 2024: 2024.06. 10.598155

[44] Meijer M, Agirre E, Kabbe M, et al. Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. Neuron, 2022, 110(7): 1193-1210. e13

[45] Barbieri A M, Lupo G, Bulfone A, et al. A homeobox gene, vax2, controls the patterning of the eye dorsoventral axis. Proceedings of the National Academy of Sciences, 1999, 96(19): 10729-10734

[46] Tang X, Schmidt T M, Perez-Leighton C E, et al. Inwardly rectifying potassium channel Kir4. 1 is responsible for the native inward potassium conductance of satellite glial cells in sensory ganglia. Neuroscience, 2010, 166(2): 397-407

[47] Manuel M N, Mi D, Mason J O, et al. Regulation of cerebral cortical neurogenesis by the Pax6 transcription factor. Frontiers in cellular neuroscience, 2015, 9: 70

[48] Chen J, Fuhler N A, Noguchi K K, et al. MYT1L is required for suppressing earlier neuronal development programs in the adult mouse brain. Genome Research, 2023, 33(4): 541-556

[49] Chen S, Zhu B, Huang S, et al. Integration of spatial and single-cell data across modalities with weakly linked features. Nature Biotechnology, 2024, 42(7): 1096-1106

[50] Lundberg E, Borner G H. Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology, 2019, 20(5): 285-302

[51] Alexandrov T. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. Annual Review of Biomedical Data Science, 2020, 3(1): 61-87

[52] Maheras K J, Peppi M, Ghoddoussi F, et al. Absence of claudin 11 in CNS myelin perturbs behavior and neurotransmitter levels in mice. Scientific Reports, 2018, 8(1): 3798

[53] Scolding N, Frith S, Linington C, et al. Myelin-oligodendrocyte glycoprotein (MOG) is a surface marker of oligodendrocyte maturation. Journal of Neuroimmunology, 1989, 22(3): 169-176

[54] Lappe-Siefke C, Goebbels S, Gravel M, et al. Disruption of Cnp1 uncouples oligodendroglial functions in axonal support and myelination. Nature Genetics, 2003, 33(3): 366-374

[55] Vandereyken K, Sifrim A, Thienpont B, et al. Methods and applications for single-cell and spatial multi-omics. Nature Reviews Genetics, 2023, 24(8): 494-515

[56] Liu X, Peng T, Xu M, et al. Spatial multi-omics: deciphering technological landscape of integration of multi-omics and its applications. Journal of Hematology & Oncology, 2024, 17(1): 72

[57] Kiessling P, Kuppe C. Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases. Genome Medicine, 2024, 16(1): 14

[58] Smith S, Karypis G. Tensor-matrix products with a compressed sparse tensor. Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms. 2015

[59] Auddy A, Xia D, Yuan M. Tensor methods in high dimensional data analysis: Opportunities and challenges. arXiv preprint arXiv:2405.18412, 2024,

[60] Dave S, Baghdadi R, Nowatzki T, et al. Hardware acceleration of sparse and irregular tensor computations of ml models: A survey and insights. Proceedings of the IEEE, 2021, 109(10): 1706-1752

[61] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. 2015

[62] Bjorck N, Gomes C P, Selman B, et al. Understanding batch normalization. Advances in Neural Information Processing Systems, 2018, 31

[63] Haghverdi L, Lun A T, Morgan M D, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology, 2018, 36(5): 421-427

[64] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30

[65] Panaretos V M, Zemel Y. Statistical aspects of Wasserstein distances. Annual Review of Statistics and Its Application, 2019, 6(1): 405-431

[66] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems, 2013, 26

[67] Dong Z, Liang F, Yang B, et al. Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 163: 327-342

[68] Umeyama S. Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 13(4): 376-380

[69] Biber P, Straßer W. The normal distributions transform: A new approach to laser scan matching. Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453). 2003. 3

[70] Magnusson M. The three-dimensional normal-distributions transform: an efficient representation for registration, surface analysis, and loop detection. Örebro Universitet. 2009

[71] Wolf F A, Angerer P, Theis F J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 2018, 19(1): 15

[72] Pardo B, Spangler A, Weber L M, et al. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. BMC genomics, 2022, 23(1): 434

[73] Zheng Y, Chen Y, Ding X, et al. Aquila: a spatial omics database and analysis platform. Nucleic Acids Research, 2023, 51(D1): D827-D834

[74] Yuan Z, Pan W, Zhao X, et al. SODB facilitates comprehensive exploration of spatial omics data. Nature Methods, 2023, 20(3): 387-399

[75] Xu Z, Wang W, Yang T, et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. Nucleic Acids Research, 2024, 52(D1): D1053-D1061

Supplementary Figures

Supplementary Fig. 1. Comparison of spatial domains by clustering assignments.

Supplementary Fig. 2. Benchmarking results of tissue domain identification in the DLPFC dataset using commonly applied clustering methods.

Supplementary Fig. 3. Robustness of SpatialFuser clustering accuracy across hyperparameter settings and random seeds.

Supplementary Fig. 4. Comparison of tissue domain identification ARI before and after alignment for BaristaSeq dataset.

Supplementary Fig. 5. UMAP visualizations for BaristaSeq dataset before and after alignment.

Supplementary Fig. 6. Hyperparameter search for SPACEL.

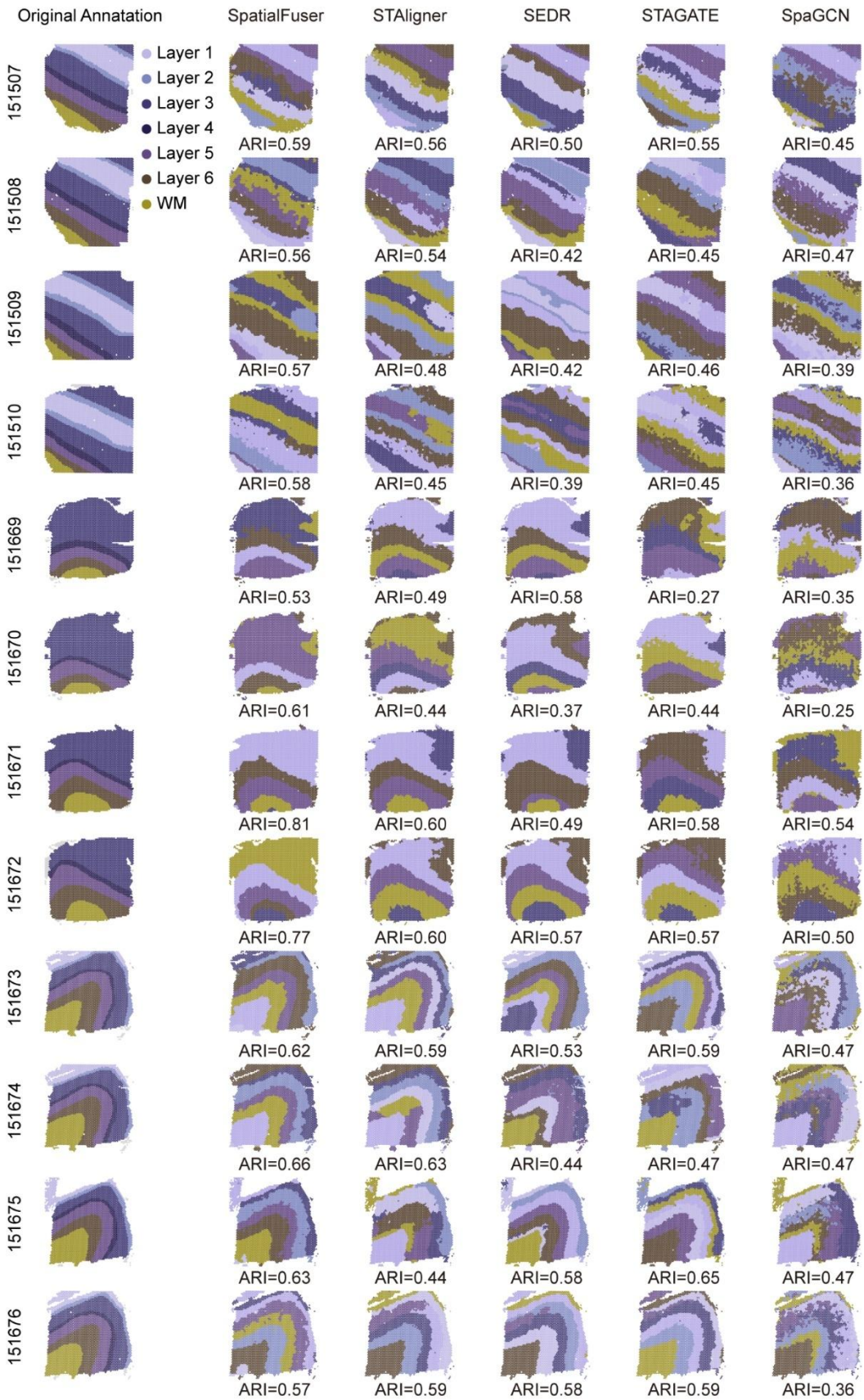
Supplementary Fig. 7. Hyperparameter search for STAligner.

Supplementary Fig. 8. Hyperparameter search for SLAT.

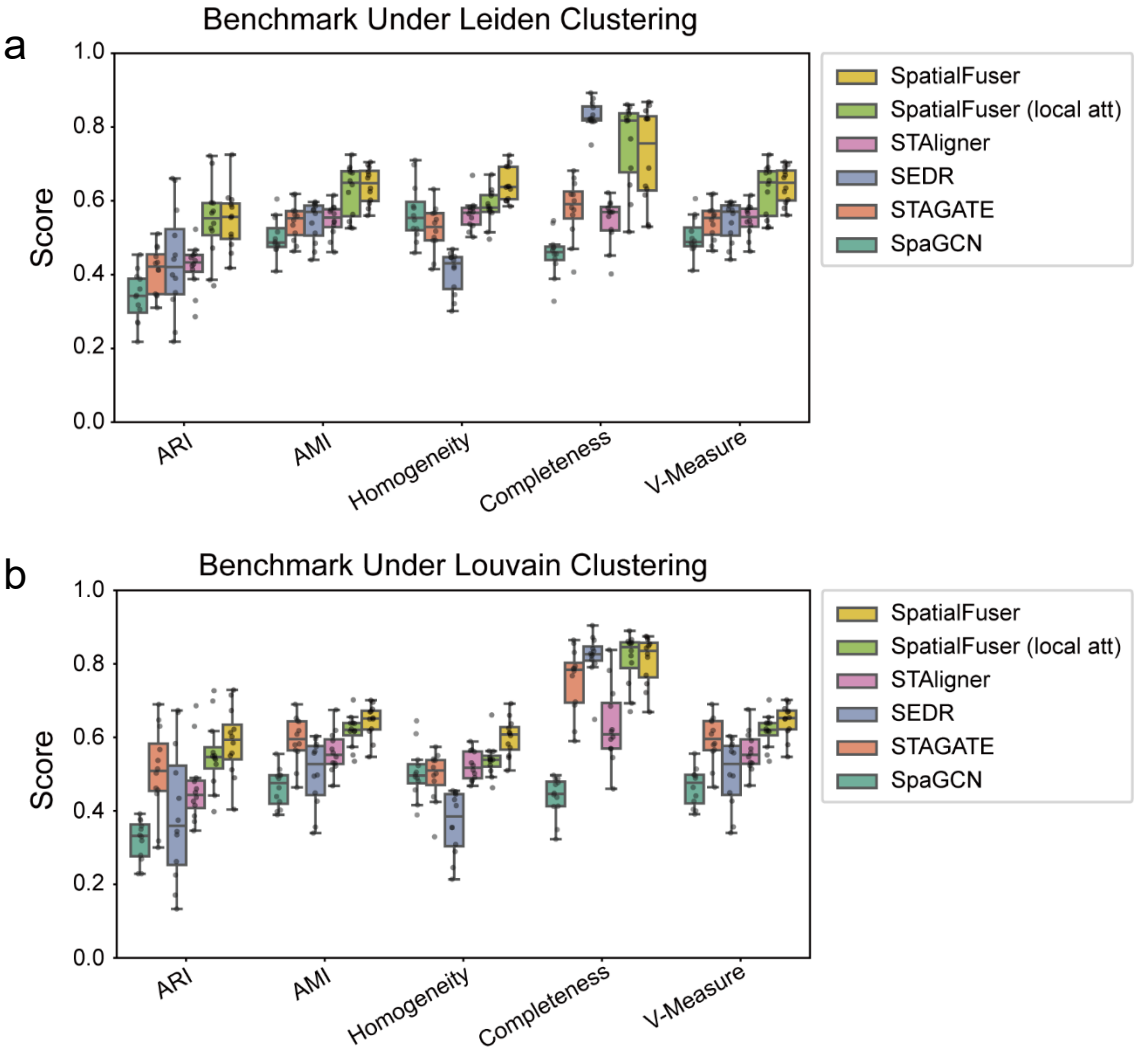
Supplementary Fig. 9. Original annotated dEGCs field and mutual top-K neighbours distributions in the Stage 57 slice of the Stereo-seq axolotl regenerative telencephalon dataset.

Supplementary Fig. 10. Similarity score distributions from random matching and SpatialFuser matching in pairwise alignments across spatial transcriptomics (MAGIC-seq), proteomics (PLATO), and metabolomics (MALDI-MSI) datasets.

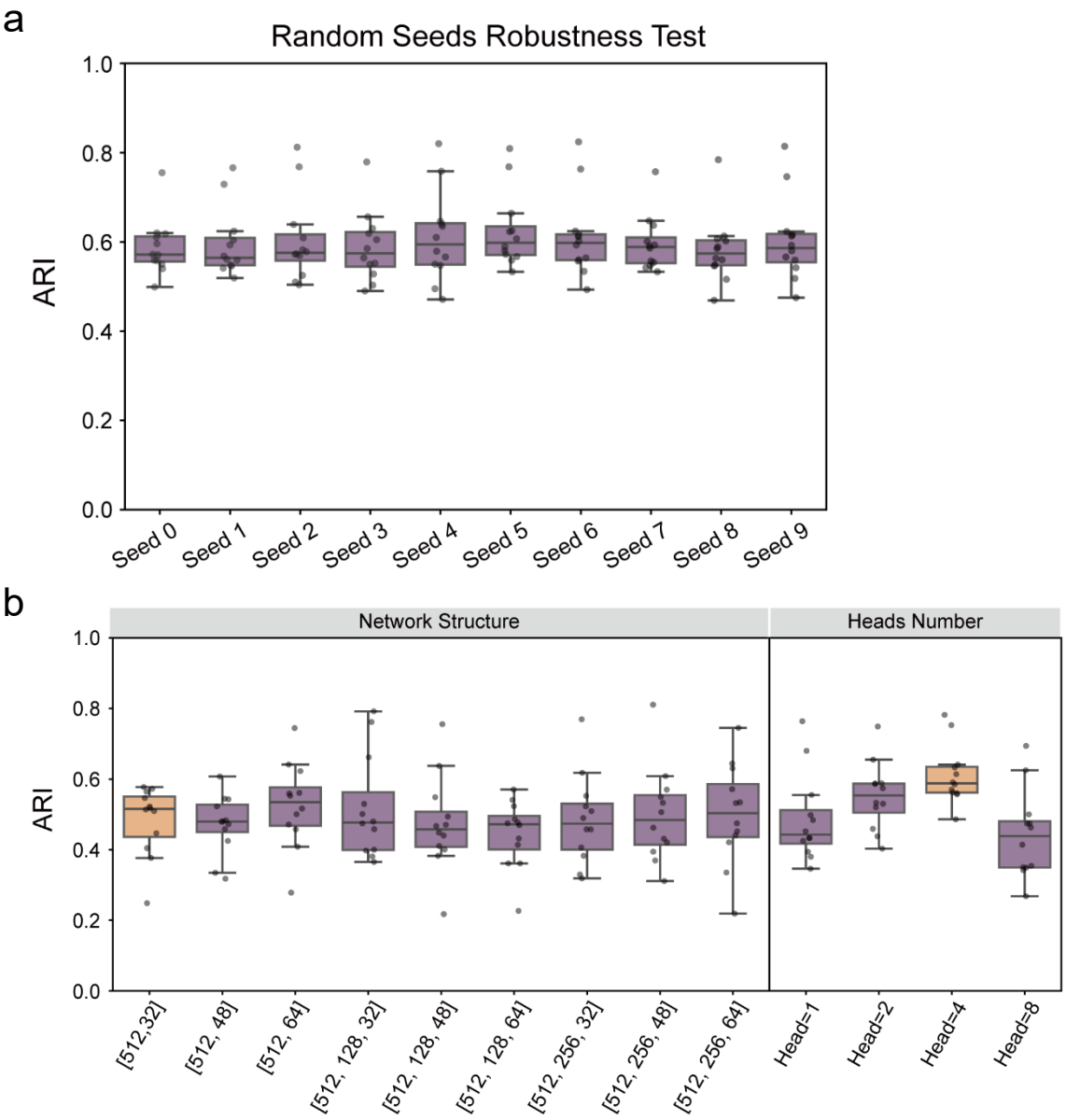
Supplementary Fig. 11. Sankey plots showing region type correspondence based on pairwise alignments from SpatialFuser across spatial transcriptomics (MAGIC-seq), proteomics (PLATO), and metabolomics (MALDI-MSI) datasets.



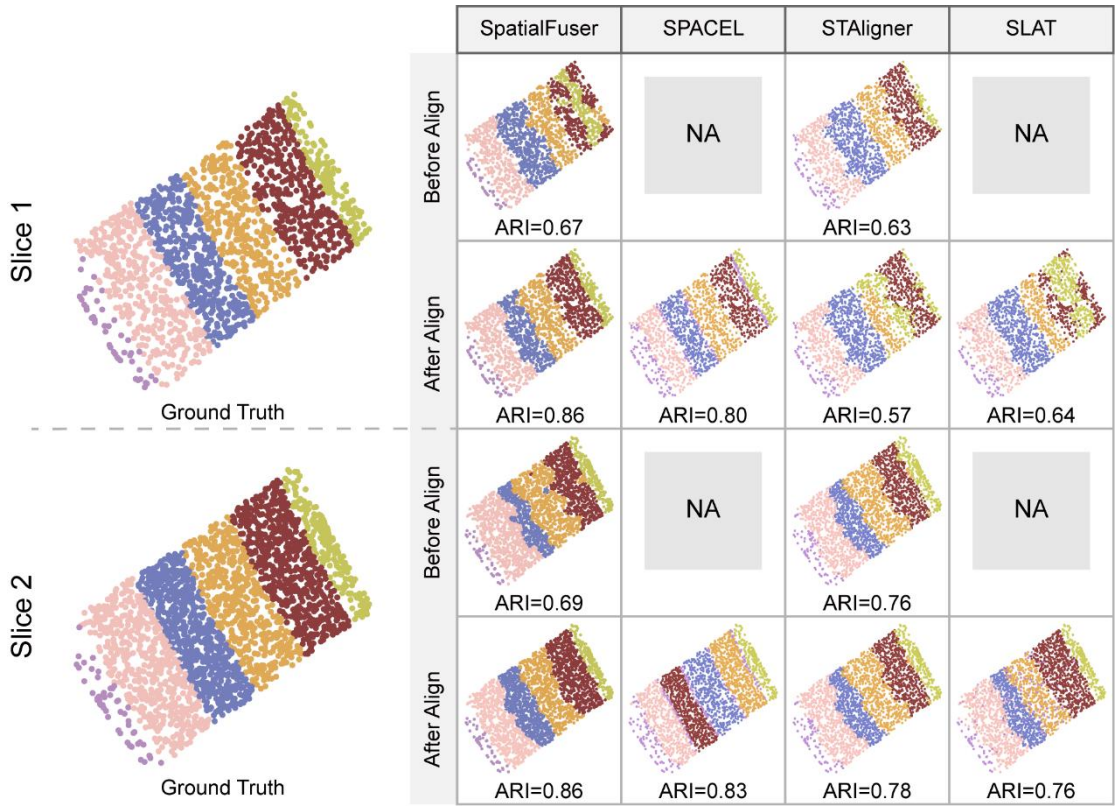
Supplementary Fig. 1. Comparison of spatial domains by clustering assignments via SpatialFuser, STAligner, SEDR, STAGATE, SpaGCN and manual annotation in all 12 sections of the DLPFC dataset.



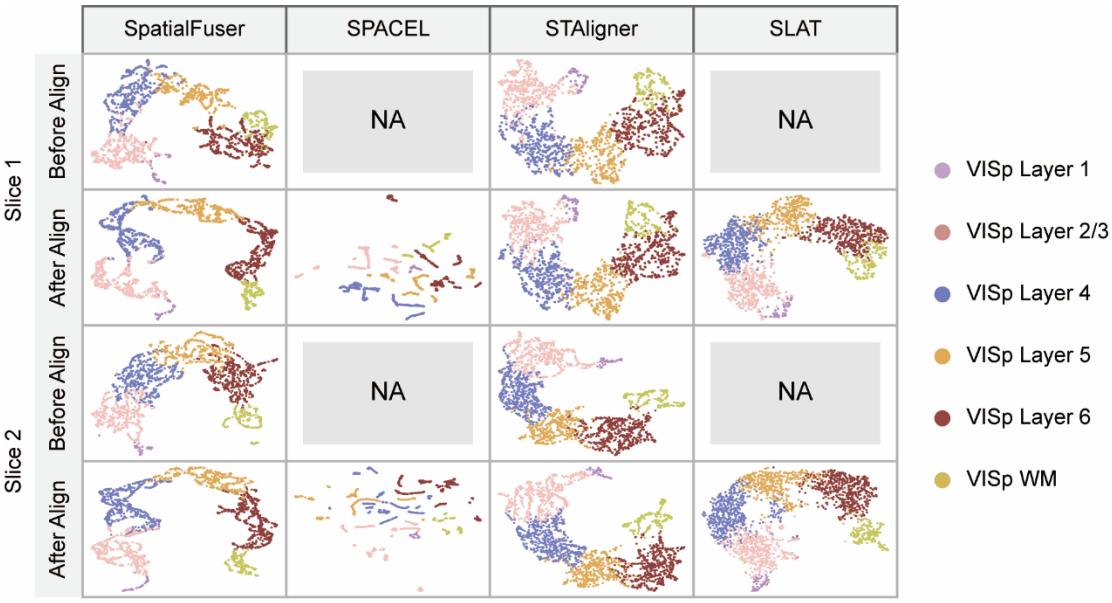
Supplementary Fig. 2. Benchmarking results of tissue domain identification in the DLPFC dataset using commonly applied clustering methods. a, Boxplots of five evaluation metrics (ARI, AMI, Homogeneity, Completeness, V-measure) for Leiden clustering results across 12 DLPFC sections. **b,** Boxplots of five evaluation metrics (ARI, AMI, Homogeneity, Completeness, V-measure) for Louvain clustering results across 12 DLPFC sections.



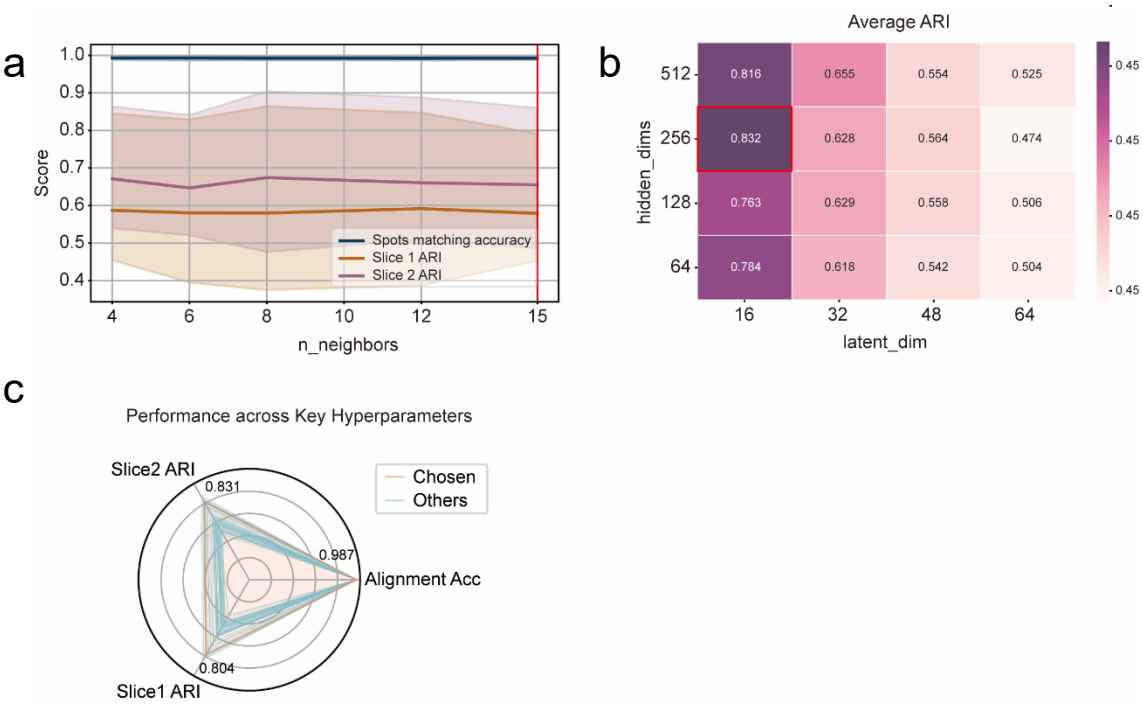
Supplementary Fig. 3. Robustness of SpatialFuser clustering accuracy across hyperparameter settings and random seeds. a, Clustering ARI of SpatialFuser across all 12 DLPFC sections under different hyperparameter settings. Hyperparameters were tuned via grid search over encoder/decoder layer numbers (2 or 3), hidden layer dimensions (512, 256, or 128), and latent dimensions (32, 48, or 64). **b,** Clustering ARI of SpatialFuser across all 12 DLPFC sections using the default hyperparameters under different random seeds.



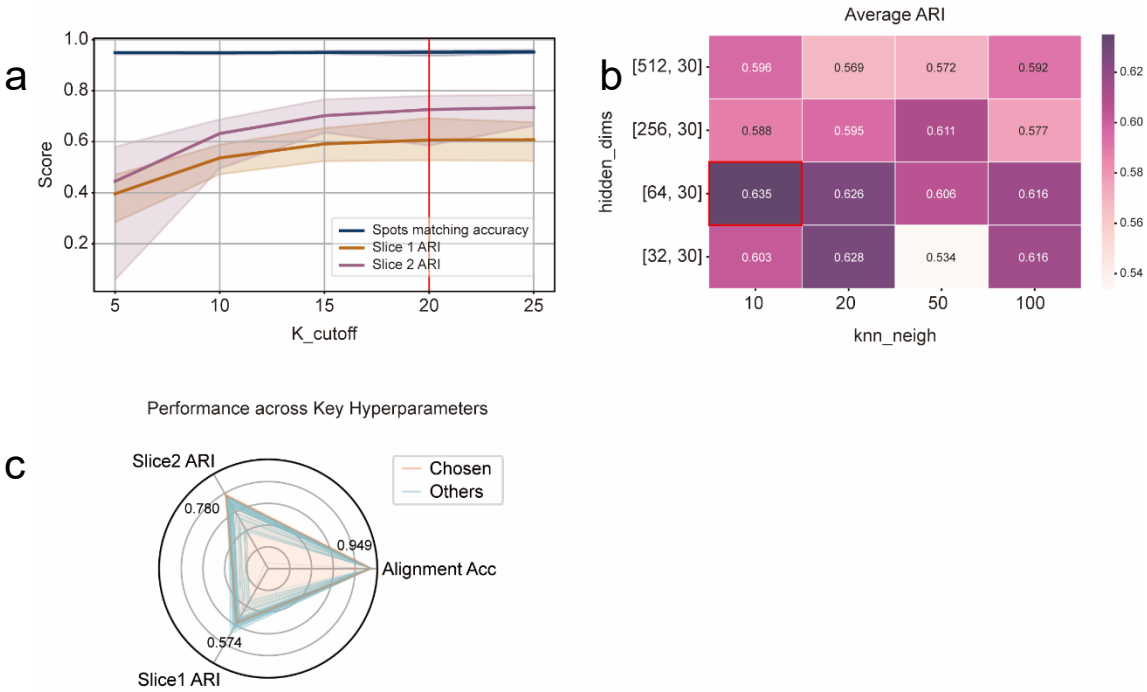
Supplementary Fig. 4. Comparison of tissue domain identification ARI before and after alignment in slice 1 and slice 2 of BaristaSeq dataset using SpatialFuser, SPACEL, STAligner, and SALT.



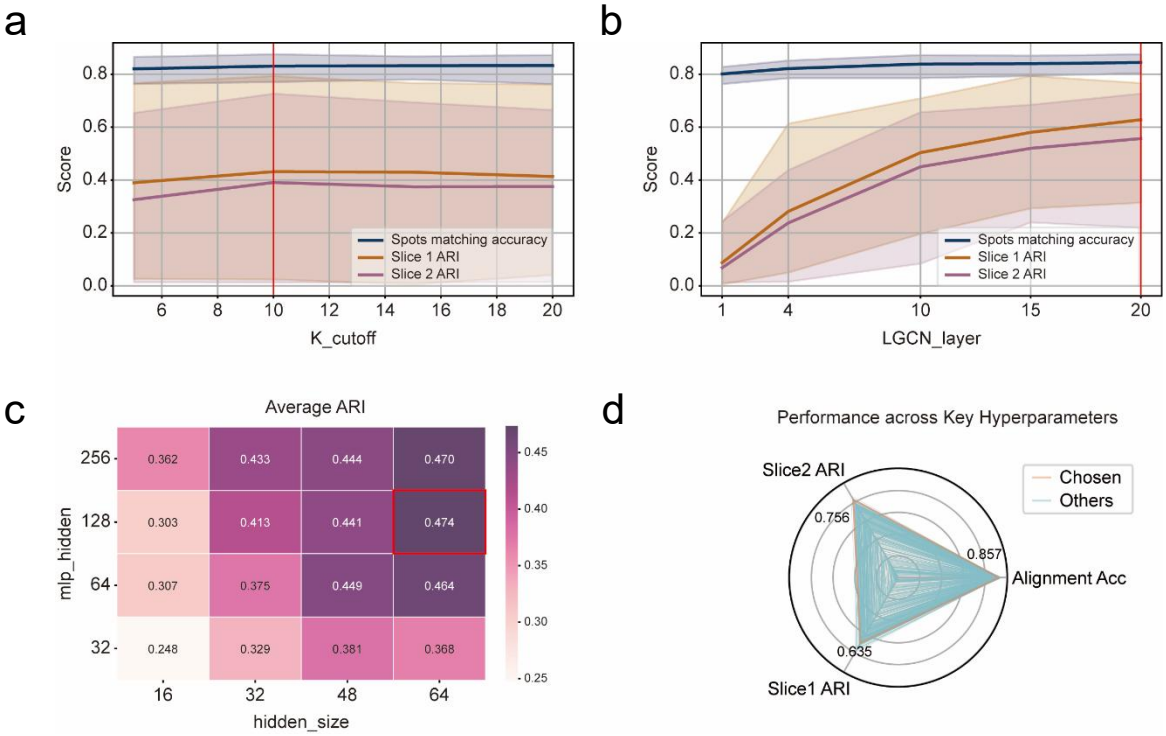
Supplementary Fig. 5. UMAP visualizations of slice 1 and slice 2 of BaristaSeq dataset before and after alignment using SpatialFuser, SPACEL, STAligner, and SALT. Spots are coloured according to the manual annotations.



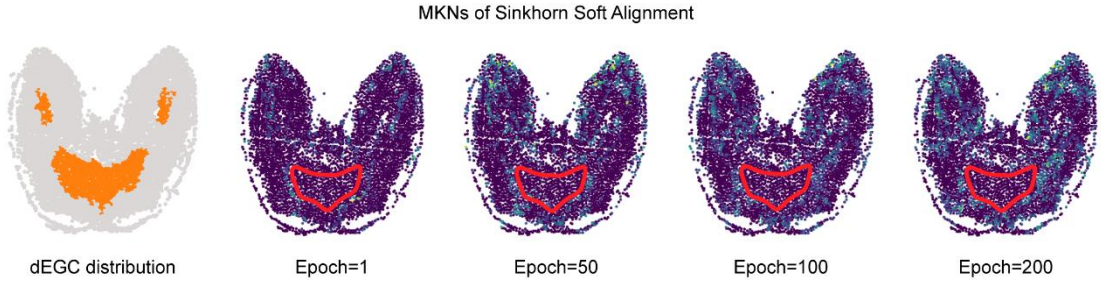
Supplementary Fig. 6. Hyperparameter search for SPACEL. a, Tissue domain detection performance (ARI) and spot type matching accuracy under different values of n_neighbors. **b,** Average tissue domain detection performance (ARI) across different network structures (hidden_dims and latent_dim). **c,** Radar plots of performance (single-slice domain detection ARI and cross-slice spot matching accuracy) across all grid search parameter combinations. The red line indicates the optimal parameter setting selected.



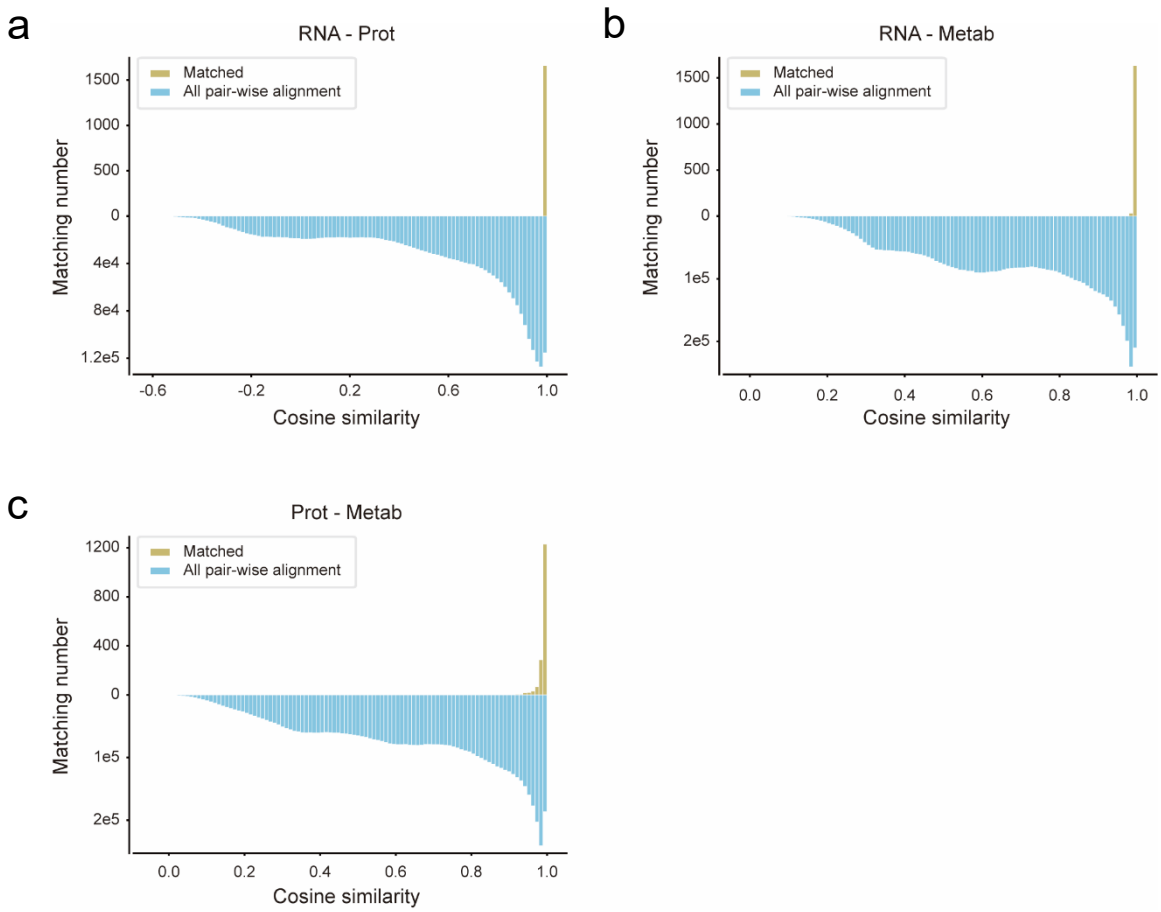
Supplementary Fig. 7. Hyperparameter search for STAligner. a, Tissue domain detection performance (ARI) and spot type matching accuracy under different values of K_cutoff. **b,** Average tissue domain detection performance (ARI) across different network setting (hidden_dims and knn_neigh). **c,** Radar plots of performance (single-slice domain detection ARI and cross-slice spot matching accuracy) across all grid search parameter combinations. The red line indicates the optimal parameter setting selected.



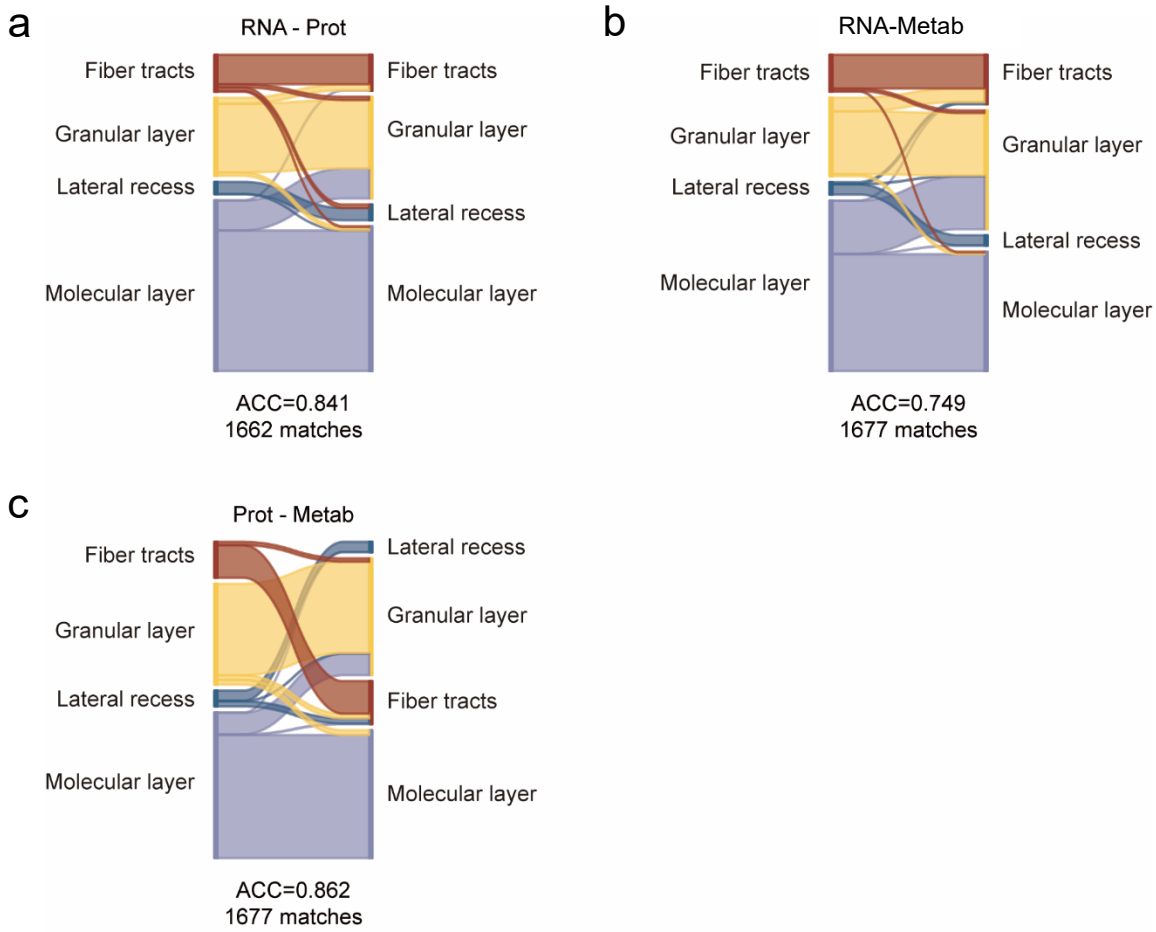
Supplementary Fig. 8. Hyperparameter search for SLAT. a, Tissue domain detection performance (ARI) and spot type matching accuracy under different values of K_cutoff. **b,** Tissue domain detection performance (ARI) and spot type matching accuracy under different numbers of LGCN_layers (1 and 4, as suggested by the authors). **c,** Average tissue domain detection performance (ARI) across different network structures (mlp_hidden and hidden_size). **d,** Radar plots of performance (single-slice domain detection ARI and cross-slice spot matching accuracy) across all grid search parameter combinations. The red line indicates the optimal parameter setting selected.



Supplementary Fig. 9. Original annotated dEGCs field and mutual top-K neighbours distributions in the Stage 57 slice of the Stereo-seq axolotl regenerative telencephalon dataset generated by the Sinkhorn-based matching layer during training. The red box highlights the dEGCs region, which consistently lacks matched points across multiple training steps (steps 1, 50, 100, and 200).



Supplementary Fig. 10. Similarity score distributions from random matching and SpatialFuser matching in pairwise alignments across spatial transcriptomics (MAGIC-seq), proteomics (PLATO), and metabolomics (MALDI-MSI) datasets. a, Similarity score distribution in MAGIC-seq and PLATO alignment. **b,** Similarity score distribution in MAGIC-seq and MALDI-MSI alignment. **c,** Similarity score distribution in PLATO and MALDI-MSI alignment.



Supplementary Fig. 11. Sankey plots showing region type correspondence based on pairwise alignments from SpatialFuser across spatial transcriptomics (MAGIC-seq), proteomics (PLATO), and metabolomics (MALDI-MSI) datasets. a, Spot type matching in MAGIC-seq and PLATO alignment. **b,** Spot type matching in MAGIC-seq and MALDI-MSI alignment. **c,** Spot type matching in PLATO and MALDI-MSI alignment.