Unified integration of spatial transcriptomics across platforms with LLOKI

Ellie Haber, Ajinkya Deshpande, Jian Ma, et al.

Genome Res. published online November 13, 2025
Access the most recent version at doi:10.1101/gr.280803.125

Method

Unified integration of spatial transcriptomics across platforms with LLOKI

Ellie Haber,[1] Ajinkya Deshpande,[1] Jian Ma,[2] and Spencer Krieger[2]

[1]*Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA;* [2]*Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*

Spatial transcriptomics (ST) has transformed our understanding of tissue architecture and cellular interactions, but integrating ST data across platforms remains challenging due to differences in gene panels, data sparsity, and technical variability. Here, we introduce LLOKI, a novel framework for integrating imaging-based ST data from diverse platforms without requiring shared gene panels. LLOKI addresses ST integration through two key alignment tasks: feature alignment across technologies and batch alignment across data sets. Optimal transport-guided feature propagation adjusts data sparsity to match scRNA-seq references through graph-based imputation, enabling single-cell foundation models such as scGPT to generate unified features. Batch alignment then refines scGPT-transformed embeddings, mitigating batch effects while preserving biological variability. Evaluations on mouse brain samples from five different technologies demonstrate that LLOKI outperforms existing methods and is effective for cross-technology spatial gene program identification, and tissue slice alignment. Applying LLOKI to five ovarian cancer data sets, we identify an integrated gene program indicative of tumor-infiltrating T cells across gene panels. Together, LLOKI provides a robust foundation for cross-platform ST studies, with the potential to scale to large atlas data sets, enabling deeper insights into cellular organization and tissue environments.

[Supplemental material is available for this article.]

In single-cell RNA sequencing (scRNA-seq) analysis, data set integration is crucial for enabling robust comparisons across studies and conditions, with its importance growing alongside large public data repositories such as the Single Cell Portal (Tarhan et al. 2023), CELLxGENE (CZI Cell Science Program et al. 2025), and the Human Cell Atlas (Regev et al. 2017). Similarly, integrating spatial transcriptomics (ST) data sets, which contain both spatial coordinates and gene expression, enables comparative analysis across samples, technologies, and conditions, revealing cellular spatial organization and dynamics across diverse contexts in health and disease (Hu et al. 2024). Although batch integration has been extensively studied in scRNA-seq (Argelaguet et al. 2021), ST presents unique challenges. A key difficulty is variation in gene panels across samples, even within the same technology. Moreover, ST technologies differ significantly in sensitivity to specific genes (Hartman and Satija 2024), and varying sparsity levels of the data further complicate integration. The goal of ST batch integration is to learn a spatially aware embedding function that maps gene expression profiles into a shared feature space, ensuring biologically similar cells remain close, regardless of technology or gene panel.

Existing ST integration methods, such as STAligner (Zhou et al. 2023), SPIRAL (Guo et al. 2023), DeepST (Xu et al. 2022), and PRECAST (Liu et al. 2023a), predominantly align tissue slices based on shared genes, limiting their ability to integrate data sets from platforms with differing gene panels. This challenge is further compounded when integrating multiple data sets, as the intersection of gene sets across technologies becomes increasingly small. For example, our results show that data sets from five different technologies share only 19 genes, despite the smallest gene panel containing around 250 genes.

Recently, single-cell foundation models (scFMs) such as UCE (Rosen et al. 2023), Geneformer (Theodoris et al. 2023), scGPT (Cui et al. 2024), and others (Yang et al. 2022; Bian et al. 2024; Hao et al. 2024) have emerged, leveraging large-scale scRNA-seq data to learn robust, generalizable cell representations. Although promising for scRNA-seq batch integration, attempts to train foundation models specifically on ST data remain limited (Tejada-Lapuerta et al. 2025; Wang et al. 2025) due to challenges such as limited data, non-overlapping gene panels, and pronounced batch effects. Despite this, scFMs remain attractive for ST integration due to their extensive pretraining on full transcriptomes, offering a scalable solution to variability in gene panels across ST technologies. However, a key limitation is that scFMs' cell representations are learned almost exclusively from scRNA-seq data, and the differences in sparsity and gene capture efficiency between scRNA-seq and ST data further complicate their direct application to ST integration.

Here, we introduce LLOKI (pronounced "low-key"), a novel framework for scalable ST integration across diverse technologies without requiring shared gene panels. The framework consists of two key components: (1) LLOKI-FP, which leverages optimal transport and feature propagation to transform ST gene expression profiles, aligning their sparsity with scRNA-seq to optimize scGPT embeddings; and (2) LLOKI-CAE, a conditional autoencoder that integrates embeddings across ST technologies using a novel loss function balancing batch integration with the preservation of biological information from LLOKI-FP embeddings. This two-stage design enables joint analyses of spatially resolved cells across technologies without requiring shared gene panels. Additionally, we show that LLOKI's embeddings facilitate important downstream

Corresponding authors: jianma@cs.cmu.edu, skrieger@andrew.cmu.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.280803.125. Freely available online through the *Genome Research* Open Access option.

tasks such as physical slice alignment and cross-gene-panel spatially variable gene program identification, enabling cross-technology analysis of tissue structure and organization across diverse contexts.

Results

Overview of LLOKI

In LLOKI, we introduce a novel framework for ST integration by decomposing the task into two alignment problems: (1) feature alignment across gene panels; and (2) batch alignment across technologies (see Fig. 1).

For feature alignment with LLOKI-FP, we leverage scGPT (Cui et al. 2024) to embed data into a shared feature space, independent of gene panel differences. However, scGPT—and most existing scFMs—only process nonzero expression genes, making high sparsity problematic, as ST data falls outside the model's scRNA-seq pretraining distribution. LLOKI-FP adjusts ST data sparsity by incorporating spatial information to align the data with an scRNA-seq reference, ensuring compatibility with pretrained scFMs. To achieve this, we first calculate a target sparsity for each cell using optimal transport, aligning ST data sparsity with the scRNA-seq reference. We then employ a novel feature propagation method that integrates gene similarity and spatial proximity to impute missing features while preserving biological variation. The sparsity-aligned ST gene expression is then processed by scGPT, generating feature-aligned embeddings for each cell.

For batch alignment with LLOKI-CAE, we use a conditional autoencoder to integrate data across ST technologies, using LLOKI-FP embeddings as input features for each cell. Our integration strategy combines three loss functions: (1) reconstruction loss to preserve accurate data representations; (2) triplet loss to enhance clustering and mitigate batch effects; and (3) a new biological conservation loss that maintains the local neighborhood structure from the preintegration embedding space, ensuring that cell-type relationships established by LLOKI-FP are preserved during batch correction. Together, these loss functions enable robust integration while preserving cell-type clusters across diverse ST platforms.

LLOKI introduces several key innovations that distinguish it from existing ST integration methods. Unlike previous approaches that rely on shared gene panels, LLOKI-FP accommodates any gene panel, ensuring all available data contributes to biologically meaningful embeddings. LLOKI also uniquely incorporates spatial information to address differences in technical dropout across technologies without oversmoothing, preserving cell-type specificity rather than collapsing cells into region-specific clusters.

Additionally, the three-part loss function in LLOKI-CAE maintains biologically meaningful heterogeneity while enabling robust batch integration. Finally, LLOKI is highly scalable and parallelizable—after initial training, each ST slice is processed individually, and LLOKI embeddings can be computed in under 5 min using <1 GB of GPU memory.

LLOKI enables cross-technology batch integration

We evaluated LLOKI's performance on cross-technology batch integration using one slice from each of five imaging-based ST data sets of coronal mouse brain sections: MERFISH (Zhang et al. 2023) (1122 genes); MERSCOPE (Yao et al. 2023) (550 genes); STARmap (Shi et al. 2023) (1022 genes); CosMx (He et al. 2022; Mallach et al. 2024) (960 genes); and Xenium (10x Genomics 2023) (248 genes). Strain, sex, and age metadata for each sample are summarized in Supplemental Table S4. Despite the smallest panel containing 248 genes, these platforms share only 19 genes, making integration particularly challenging for existing alignment methods. To benchmark LLOKI, we compared it against six baseline methods: (1) PCA; (2) Harmony (Korsunsky et al. 2019) (for scRNA-seq integration); (3) Seurat (Satija et al. 2015) (for scRNA-seq integration); (4) scVI (Lopez et al. 2018) (for scRNA-seq integration); (5) STAligner (Zhou et al. 2023) (for ST integration); and (6) the raw output of LLOKI-FP (prior to batch alignment via LLOKI-CAE). We visualized the embedding space from each method using UMAP (Fig. 2A), coloring cells either by batch or cell type. Because the data sets differ substantially in annotation granularity and their fine-grained labels are neither harmonized nor mutually exhaustive, we used harmonized broad labels for consistency across data sets (Supplemental Table S3). For the Xenium data set, which lacked annotations, cell types were inferred by marker gene analysis (Supplemental Information), yielding only coarse classes (e.g., oligodendrocytes, excitatory neurons). In contrast, MERFISH and MERSCOPE distinguish finer subtypes such as oligodendrocyte maturation states and cortical-layer excitatory populations.

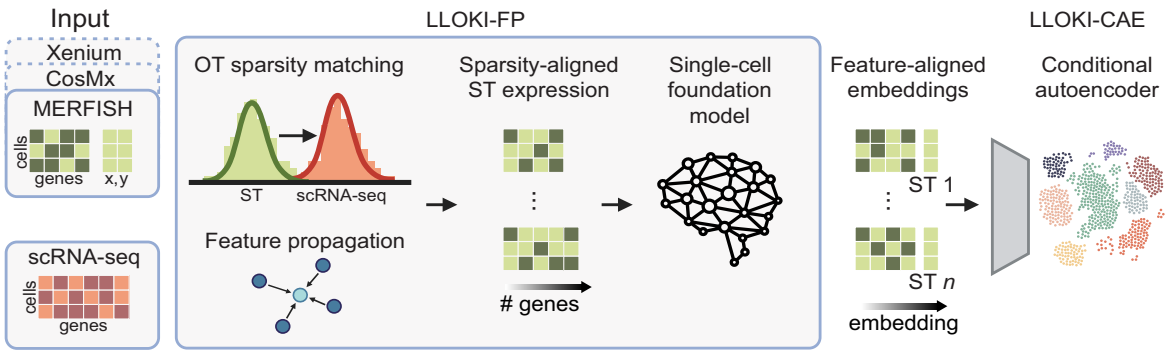
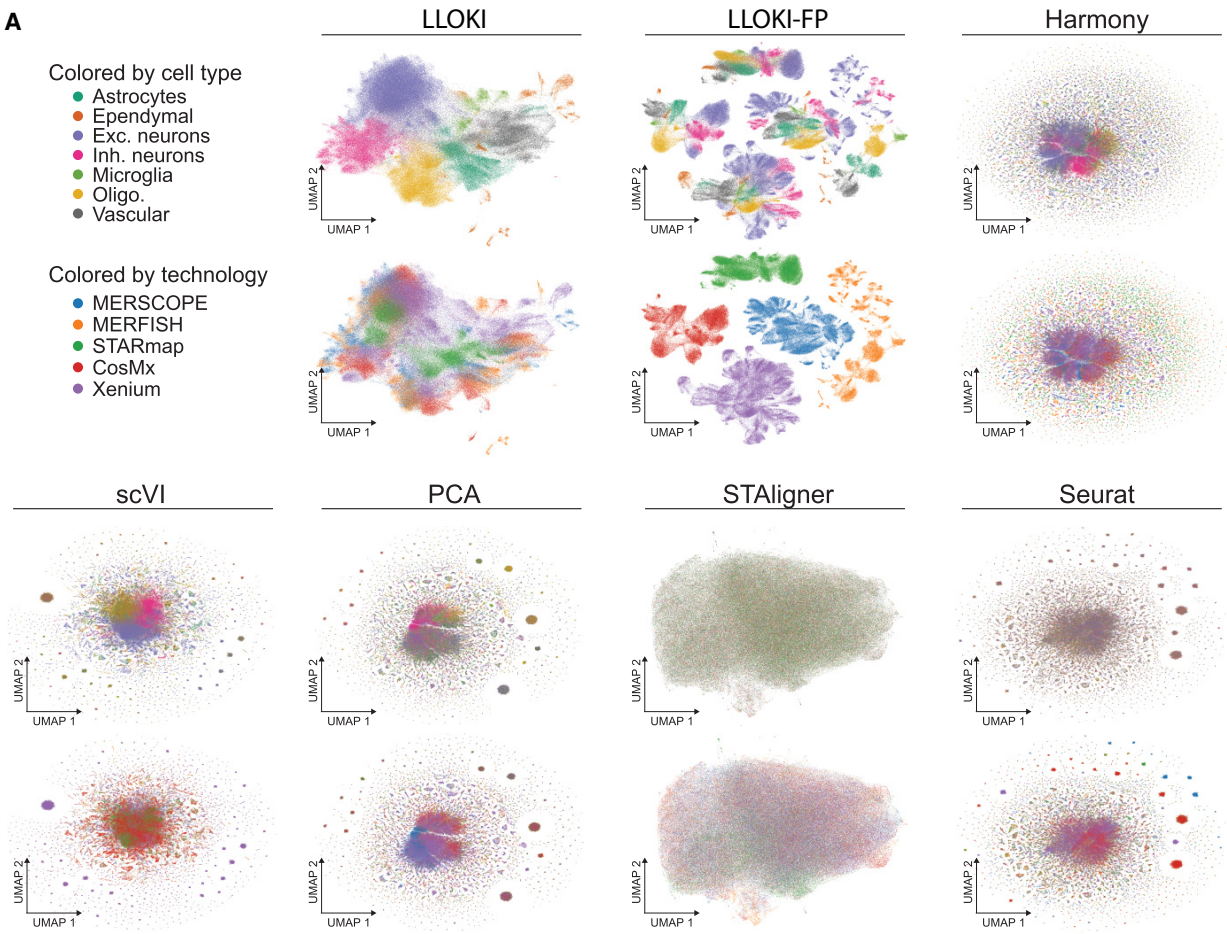


Figure 1. Overview of LLOKI. LLOKI performs a two-stage alignment for integrating ST data: First, it unifies the feature space across ST samples, regardless of gene panel; then, it removes technology-specific batch effects. LLOKI-FP applies optimal transport-guided feature propagation to impute missing gene expression values using spatially informed cell graphs, mitigating data sparsity and gene sensitivity differences. A single-cell foundation model then embeds the data into a unified feature space. LLOKI-CAE further integrates these embeddings across batches using a conditional autoencoder, removing technology-specific effects while preserving biological structure. The resulting embeddings minimize batch effects and support downstream tasks such as cross-technology slice alignment and spatial gene program identification.



B

Method	Bio conservation					Batch correction			Aggregate score	
	Isolated labels	cLISI	KMeans NMI	KMeans ARI	Silhouette label	Graph connectivity	Silhouette batch	PCR comparison	Bio Conservation	Batch Correction
LLOKI	0.56	1.00	0.64	0.57	0.58	0.91	0.88	0.76	0.67	0.85
scVI	0.63	0.89	0.23	0.19	0.49	0.37	0.92	0.80	0.49	0.70
STAligner	0.49	0.59	0.00	0.00	0.49	0.26	0.93	0.86	0.31	0.69
Seurat	0.59	0.68	0.01	0.00	0.47	0.05	0.95	0.88	0.35	0.62
Harmony	0.49	0.58	0.00	-0.00	0.49	0.07	0.92	0.88	0.31	0.62
LLOKI-FP	0.51	1.00	0.13	0.07	0.51	0.29	0.64	0.00	0.44	0.31
PCA	0.40	0.93	0.14	-0.05	0.37	0.53	0.67	0.00	0.36	0.40

Figure 2. Comparative analysis of batch integration using LLOKI versus six baseline methods across slices from five spatial transcriptomics technologies. (*A*) UMAP visualizations of integrated data, colored by cell type (*top*) and by technology (*bottom*), showing the degree of batch separation and preservation of biological variation. (*B*) Quantitative performance evaluation of all methods across eight metrics, measuring both biological variation preservation and batch correction across technologies.

Attempting subtype-level comparisons across all data sets would therefore require collapsing detailed taxonomies to the coarsest resolution or merging nonequivalent labels, both of which risk obscuring genuine biological differences.

For quantitative evaluation, we employed eight metrics from scib-metrics (Luecken et al. 2022; Methods). Five metrics assessed the preservation of biological variation based on cell-type separation using the unified high-level labels for each data set, and three metrics measured batch effect removal, reflecting the degree of integration across technologies into a shared embedding space. For each category, we computed aggregate scores as the mean of their corresponding metrics.

LLOKI achieved the best balance between preserving biological signals and removing batch effects (Fig. 2B). It preserved biological variation between cell types, attaining the highest biological conservation score (0.67) and batch correction score (0.85). In

contrast, LLOKI-FP alone, although yielding a moderate biological conservation score (0.44), performed poorly in batch correction (0.31), highlighting the importance of LLOKI-CAE in aligning LLOKI-FP embeddings across batches.

To further evaluate the importance of the OT alignment step in LLOKI-FP, we compared clustering performance using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) across five ST data sets and related these results to each slice's average per-cell sparsity gap with its matched scRNA-seq reference (Supplemental Table S2). As shown in Supplemental Table S1, LLOKI-FP with OT consistently outperforms unmodified scGPT embeddings on all five data sets and surpasses LLOKI-FP without OT on all three. The size of the performance gain reflects the data set's pre-imputation sparsity gap and gene panel size. STARmap, which is highly sparse and measures over 1000 genes, shows the largest improvement. MERSCOPE, denser than its reference but with a mid-sized panel, still gains noticeably. In contrast, Xenium, which is denser and limited to a narrow panel, sees a modest decline. Thus, OT primarily benefits slices that are highly sparse and profiled with broader panel, adapting to platform sensitivity while improving cell-type separability prior to scGPT embedding.

Among baseline methods, scVI and STAligner achieved the next highest batch correction scores (0.70 and 0.69, respectively), but their biological conservation scores were substantially lower (0.49 and 0.31). Whereas scVI was able to separate major cell types such as excitatory neurons, inhibitory neurons, and oligodendrocytes, it failed to distinguish the remaining cell types. STAligner showed similar limitations, with UMAPs revealing a collapse of the embeddings into a unified structure with minimal cell-type separation. This suboptimal integration is likely due to limited gene panel overlap—only 19 shared genes—restricting the ability to fully capture the spectrum of biological variation. PCA embeddings yielded low biological conservation (0.36) and batch correction (0.40), showing weak cell-type specificity in the UMAP visualization. Seurat and Harmony improved batch correction relative to PCA (0.62) but further reduced biological conservation (0.35 and 0.31), indicating a trade-off in improved batch alignment at the expense of cell-type resolution.

To dissect the contributions of LLOKI-CAE's three-part loss function and to evaluate the importance of the LLOKI-FP feature transformation, we conducted two ablation experiments (Supplemental Fig. S1). In the first experiment, we removed each component of the loss function. Removing the biological conservation loss had the most significant impact, reducing the biological conservation score by 0.37, underscoring its critical role in maintaining cell-type separation in the embedding space. In contrast, removing the triplet loss compromised batch correction, highlighting the delicate balance required between these objectives. In the second experiment, we compared LLOKI-CAE's performance when using scGPT embeddings versus the transformed LLOKI-FP embeddings as input. Whereas using scGPT embeddings alone improved batch correction by 0.04, its biological conservation score dropped by 0.18. Together, these results suggest that, whereas LLOKI-FP is highly effective at maintaining biological variation and cell-type separation, LLOKI-CAE is crucial for integrating slices across technologies and successfully removing batch effects.

LLOKI achieves robust cross-technology spatial alignment

We evaluated LLOKI's ability to align multiple ST slices within a common spatial coordinate framework while ensuring key tissue features are properly aligned. To assess its robustness, we tested two key aspects: (1) its ability to align slices across different technologies; and (2) its performance when slices have minimal overlap in their gene panels. For slice alignment, LLOKI embeddings were used to generate landmark pairs between ST slices by identifying mutual nearest neighbors within the LLOKI embedding space. These landmark pairs were then aligned using the Kabsch algorithm (Kabsch 1976), which finds the optimal rotation and translation to minimize root-mean-square deviation (RMSD) between paired landmarks.

First, we evaluated LLOKI's ability to align slices from different technologies by selecting one slice each from the MERSCOPE, MERFISH, and STARmap data sets (Fig. 3A, left) and performing pairwise alignments using overlapping genes. As baselines, we compared against: (1) the Kabsch algorithm using mutual nearest neighbors from PCA embeddings of shared genes; (2) PASTE (Zeira et al. 2022), a state-of-the-art spatial alignment method that uses fused Gromov–Wasserstein optimal transport to compute alignment based on both transcriptional and spatial similarity; (3) STAligner (Zhou et al. 2023), which provides a slice alignment module in its integration framework; and (4) GPSA (Jones et al. 2023), which uses deep Gaussian processes to map spatial coordinates between slices into a common coordinate system based on gene expression.

All four baselines failed to correctly align the slices. However, LLOKI successfully aligned them such that corresponding regions overlapped (Fig. 3A, right). PCA, PASTE, and STAligner all rotate at least one of the slices incorrectly. In addition to scaling, rotating, and translating the slices, GPSA additionally allows nonlinear scaling, which can produce artifacts like those seen in their alignment when batch effect is pronounced.

To quantify alignment accuracy, we analyzed each pair of aligned slices by defining interslice cell pairs, where each cell in one slice was paired with its nearest spatial neighbor from the other slice in the aligned coordinate space. For each pair, we computed Pearson's correlation using their shared genes. LLOKI outperformed all baselines across all three slice pairs (Fig. 3B). In particular, for the MERFISH–MERSCOPE pair, LLOKI achieved a Pearson's correlation of 0.30, compared to 0.17 for the next-best methods, GPSA and STAligner. The modest Pearson's correlations primarily reflect the small set of genes common to each slice pair, because the metric is calculated solely using those shared transcripts. To test LLOKI's performance under low gene panel overlap, we used the MERFISH and MERSCOPE slices and created 10 simulated data sets by progressively downsampling their shared genes in 10% increments, repeating this process five times per increment to create replicates. Because these two data sets share unified cell-type annotations, we also measured how often interslice pairs share the same annotated cell type, in addition to computing Pearson's correlation. Before alignment, we applied a random rotation and translation to one slice to ensure that alignment methods were actively aligning the data, rather than relying on preexisting spatial similarity.

LLOKI consistently outperformed GPSA, PASTE, and PCA, and tracked closely behind STAligner, which held a small but consistent edge (Fig. 3C,D). Notably, LLOKI and STAligner maintained stable performance with few overlapping genes in this pairwise cross-technology alignment, whereas the other methods exhibited markedly decreased performance.

Overall, these results demonstrate that LLOKI embeddings are highly robust to differences in gene panels, enabling accurate alignment even with minimal overlap between slices. Additionally, LLOKI effectively mitigates batch effects across

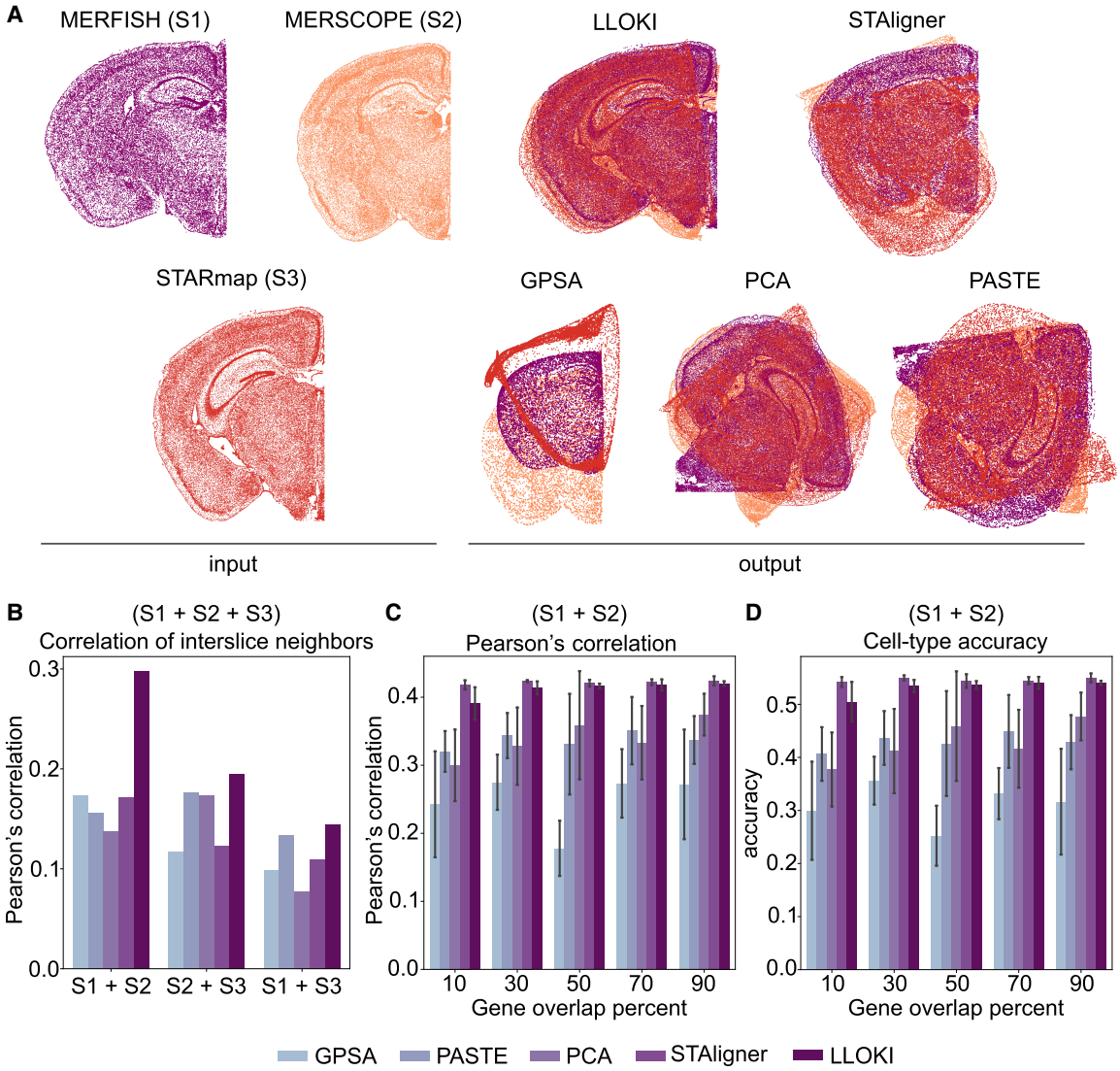


Figure 3. Evaluation of LLOKI's performance in spatial slice alignment. (*A*) In situ visualizations of three slices from the MERFISH, MERSCOPE, and STARmap data sets, along with alignment results using LLOKI and four baseline methods. (*B*) Alignment accuracy as measured by Pearson's correlation of interslice neighbors between each pair of slices from *A*. (*C*,*D*) Comparison of LLOKI and four baseline methods on alignment accuracy of the MERFISH and MERSCOPE slices as gene overlap decreases, showing results for (*C*) Pearson's correlation, and (*D*) cell-type accuracy.

technologies, facilitating accurate cross-technology slice alignment for data sets from diverse ST platforms.

LLOKI enhances cross-technology spatial gene program detection

To assess LLOKI's utility for downstream analysis, we integrated data from four ST platforms—MERFISH, MERSCOPE, CosMx, and Xenium—and evaluated whether LLOKI embeddings improve cross-technology analyses using SpiceMix (Chidester et al. 2023). SpiceMix models gene expression as a combination of latent factors ("metagenes") that capture transcriptional programs with spatial organization. Joint analysis enables a more comprehensive understanding of tissue biology by revealing conserved spatial patterns that may be missed when data sets are analyzed in isolation. However, SpiceMix requires a common feature space, which limits its application to data sets with few overlapping genes. We address this limitation using LLOKI embeddings as an alternative input to SpiceMix, providing a unified feature space across data sets.

We computed metagenes using SpiceMix under three conditions: (1) LLOKI-based integration, using 128-dimensional LLOKI embeddings as input to SpiceMix; (2) a shared-gene approach, applying SpiceMix to the 22 genes common across data sets; and (3) independent SpiceMix runs on each data set, with a post hoc optimal metagene mapping.

To evaluate the consistency and biological relevance of metagenes across conditions, we matched metagenes to the major cell types annotated in each data set (Fig. 4A). For the LLOKI-based integration, we found consistent associations between metagenes and major cell types across data sets. In particular, metagene 1 was enriched in astrocytes, metagene 6 in microglia, metagenes 5, 7, and 14 in vascular cells, and metagenes 11 and 13 in oligodendrocytes, with additional metagenes capturing excitatory and inhibitory neurons. In contrast, both the shared-gene and

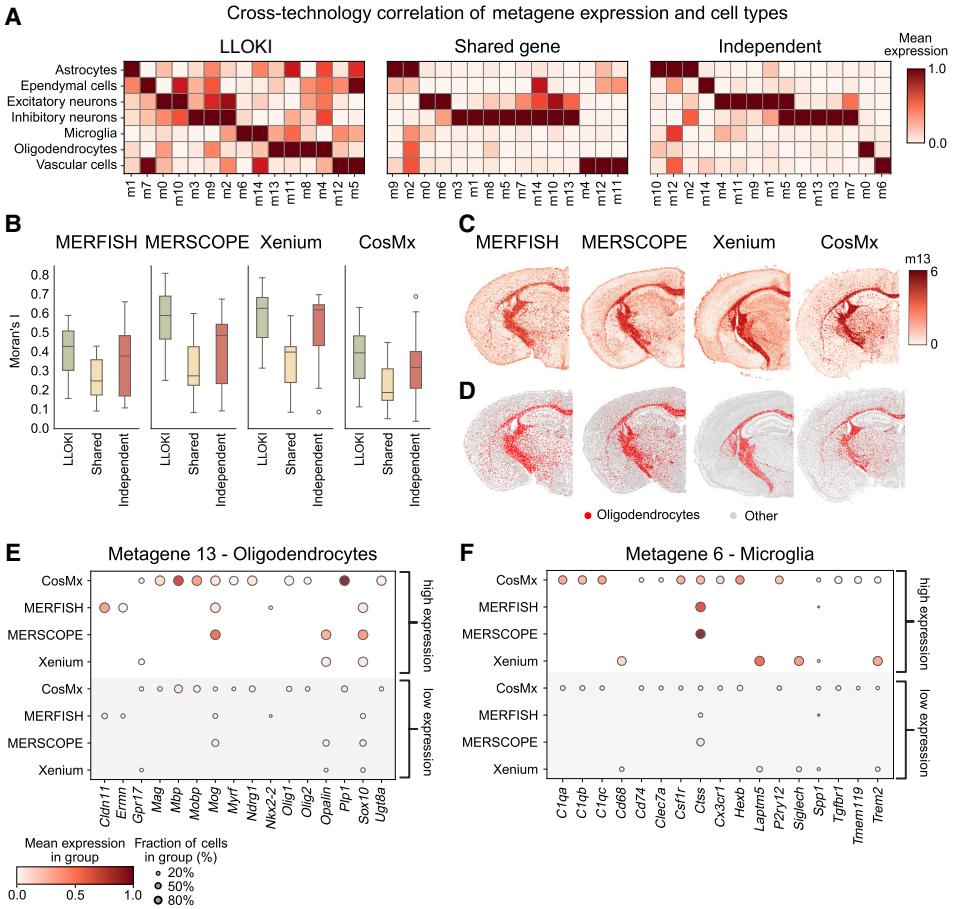


Figure 4. SpiceMix with LLOKI embeddings enhances cross-technology metagene discovery. (*A*) Cell-type enrichment of metagenes across MERFISH, MERSCOPE, CosMx, and Xenium data sets, comparing three approaches: LLOKI-based integration, the shared-gene approach (22 common genes), and independent analyses. (*B*) Quantitative assessment of metagene quality using Moran's I spatial autocorrelation scores, shown as box plots. (*C*) Spatial distribution of metagene 13 across all four technologies, showing consistent patterns using LLOKI embeddings with SpiceMix. (*D*) Spatial distribution of oligodendrocytes across MERFISH, MERSCOPE, Xenium, and CosMx, demonstrating the correspondence between metagene 13 and oligodendrocytes. (*E*) Expression levels of canonical oligodendrocyte marker genes across MERFISH, MERSCOPE, Xenium, and CosMx in cells highly expressing metagene 13 versus cells lowly expressing metagene 13. (*F*) Expression levels of canonical microglia marker genes across MERFISH, MERSCOPE, Xenium, and CosMx in cells highly expressing metagene 6 versus cells lowly expressing metagene 6.

independent approaches captured only partial associations, failing to consistently identify microglia- and oligodendrocyte-specific metagenes.

We further evaluated metagene quality for each run using Moran's I to quantify spatial autocorrelation (Fig. 4B). In brain tissue, functionally similar cells tend to cluster spatially, so a metagene exhibiting spatial autocorrelation indicates that it recapitulates known cell-type patterns. Our results show that the LLOKI-based approach consistently produced metagenes with strong spatial autocorrelation in all data sets, whereas the shared-gene approach exhibited the lowest autocorrelation and the independent analyses showed high variability.

To assess biological relevance, we examined metagenes with consistent spatial and cell-type associations in all data sets. We found that metagene 13 showed consistent spatial expression patterns in all four data sets (Fig. 4C) and was enriched in oligodendrocytes (Fig. 4D), indicating it captures a cell type–specific gene program.

To further investigate the molecular underpinnings of metagene 13, we stratified cells by metagene expression and cross-referenced their original gene panels to compare expression of oligodendrocyte marker genes between groups (Fig. 4E). Nearly all oligodendrocyte markers exhibited strong correlation with metagene 13. Many of these markers were only measured in one data set, including *Mbp*, *Mobp*, and *Plp1* in CosMx, and *Cldn11* and *Ermn* in MERFISH. Other markers, such as *Mag*, *Opalin*, and *Sox10*, were measured in multiple data sets, Collectively, these findings highlight the value of leveraging full gene panels via LLOKI—rather than restricting analysis to a limited set of shared genes—and validate the association of metagene 13 with oligodendrocyte identity.

We used a similar approach to associate metagene 6 with microglia (Fig. 4F). In both MERFISH and MERSCOPE, cells with high metagene 6 expression consistently exhibited elevated expression of the microglia marker *Ctss*. Higher expression of metagene 6 was associated with increased expression of *Laptm5*, *Siglech*, and *Trem2* in Xenium, and *C1qa*, *C1qb*, *C1qc*, and *Hexb* in CosMx. These markers—many of which are uniquely expressed across the different technologies—enabled robust identification of a conserved program in microglia only when using full gene panels

with LLOKI. As with metagene 13, microglia-associated metagenes were not consistently detected using baseline methods.

Together, these results show that LLOKI's integrated embeddings enable SpiceMix to delineate biologically meaningful spatial gene programs, overcoming the limitations imposed by data set-specific gene panels and enhancing cross-technology analyses. For further validation, in situ metagene plots for SpiceMix runs using LLOKI embeddings, shared-genes, and independent gene panels are provided in Supplemental Figures S2–S4, respectively.

LLOKI identifies a shared gene program for tumor infiltration in ovarian cancer

We next applied LLOKI to integrate five ST data sets of human ovarian cancer (Yeh et al. 2024), which differ in both gene panel composition and technology (Fig. 5A). These data sets include three from CosMx (SMI), one from Xenium (ISS), and one from MERFISH. The CosMx data sets include: (1) 960 genes measured across 100 small tissue samples (0.9 mm × 0.6 mm); (2) 1000 genes measured across four large whole-tissue samples (up to 2 cm); and (3) 6175 genes measured across 62 small tissue samples (0.5 mm × 0.5 mm). The Xenium data set contains 240 genes from 32 small tissue samples (1.5 mm × 1.5 mm), whereas the MERFISH data set includes 140 genes from four large whole-tissue samples (up to 2 cm). Integration via LLOKI achieved robust mixing of shared cell types across these technologies (Fig. 5B), yielding a biological conservation score of 0.47 and a batch correction score of 0.23 when evaluated on nonmalignant cells. Malignant cells remained largely distinct between data sets, reflecting tumor-specific transcriptional differences from diverse origins.

Using LLOKI embeddings, we further investigated T cell heterogeneity with a focus on tumor infiltration. We combined T cells from all data sets and clustered them based solely on their LLOKI embeddings. This unsupervised approach partitioned T cells into two groups that corresponded well with the percentage of malignant cells among their 100 nearest spatial neighbors—a proxy for local tumor infiltration (Fig. 5C). In contrast, performing the same analysis using shared genes resulted in poor integration, making it impossible to partition cells in a way that corresponded to tumor infiltration (Supplemental Fig. S5A). Analyzing each data set individually also failed to recover meaningful T cell partitions, particularly for the ISS 240 data set, where T cells could not be grouped by local tumor infiltration (Supplemental Fig. S5B).

To characterize the genes responsible for this tumor infiltration program, we performed integrated differential expression analysis in two stages: (1) comparing T cells against all other cell types to obtain T cell-specific genes for each data set's gene panel; and (2) identifying differentially expressed genes between the two T cell groups detected via LLOKI embeddings. Genes consistently up- or downregulated across data sets were retained, yielding a shared gene panel that reliably distinguishes tumor-infiltrating T

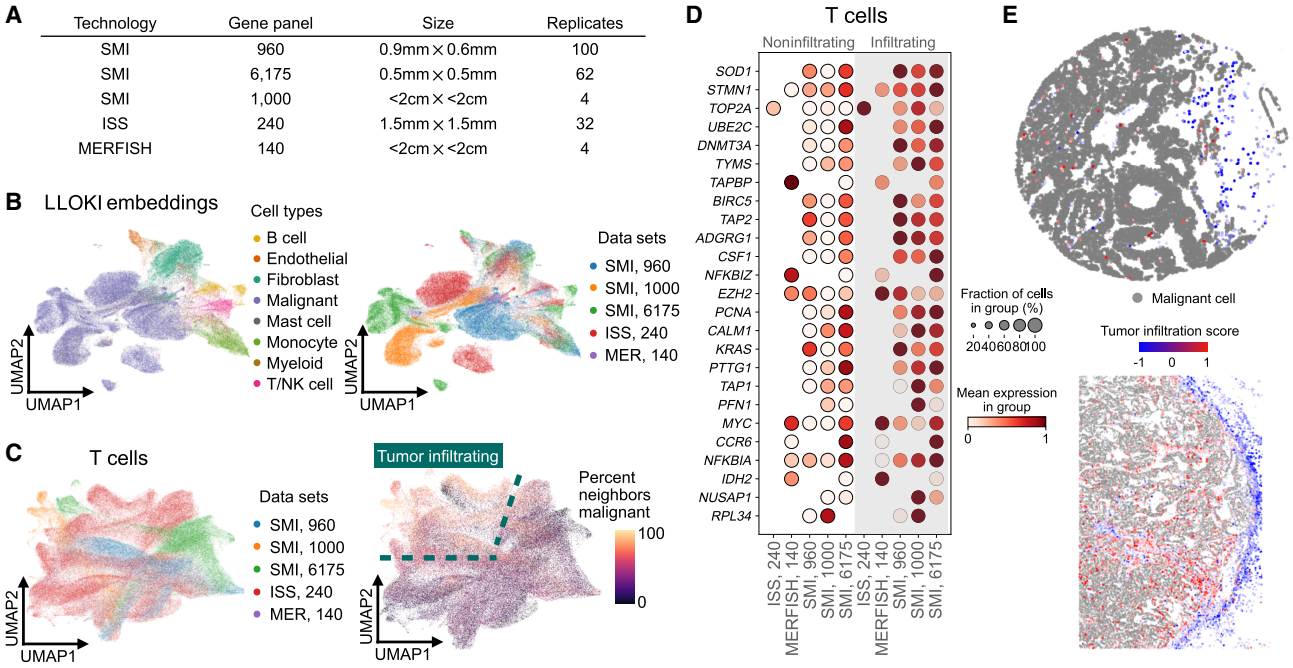


Figure 5. LLOKI identifies cross-technology gene program indicative of tumor infiltration in T cells in ovarian cancer. (*A*) Summary of the five ovarian cancer ST data sets used in this analysis. (*B*) UMAP visualization of LLOKI embeddings shows successful integration of nonmalignant cells across data sets, whereas transcriptionally distinct tumor cells remain separate. (*C*) UMAP visualization of LLOKI embeddings of T cells, colored by data set (*left*) or by the proportion of 100-nearest spatial neighbors that are malignant (*right*). Dashed outlines indicate regions identified as tumor-infiltrating. (*D*) Expression levels of upregulated genes from the cross-technology tumor infiltration gene program in infiltrating versus noninfiltrating T cells. Missing dots indicate genes that were not measured in particular data sets. (*E*) Representative in situ visualization showing T cells colored by the tumor infiltration score and malignant cells in gray.

cells (Fig. 5D). In particular, this gene panel includes markers associated with cell proliferation (e.g., *STMN1*, *TOP2A*, *UBE2C*, *BIRC5*, *NUSAP1*) and antigen processing (e.g., *TAP1*, *TAP2*, *TAPBP*), suggesting that these processes are correlated with tumor infiltration.

To assess the biological relevance of this gene panel, we devised a tumor infiltration score, weighting upregulated genes positively and downregulated genes negatively. When spatially projected onto individual T cells, this score revealed a clear pattern: T cells within tumor regions exhibited consistently higher scores than those outside the tumor (Fig. 5E). We further quantified the correlation between this tumor infiltration score and the percentage of malignant cells among the 100 nearest spatial neighbors. Across 170,000 T cells from all data sets, we observed a significant correlation (Pearson $R = 0.326$, $P < 1 \times 10^{-200}$; Spearman's $\rho = 0.366$, $P < 1 \times 10^{-200}$), confirming the biological relevance of our scoring approach.

These findings further demonstrate that LLOKI not only enables effective integration of heterogeneous spatial transcriptomics data but also facilitates the discovery of biologically meaningful gene programs, such as the one defining tumor-infiltrating T cells in ovarian cancer.

Discussion

In this work, we developed LLOKI, a novel framework for integrating ST data across diverse platforms without requiring shared gene panels. LLOKI combines feature alignment and batch integration through two complementary components: LLOKI-FP, which aligns gene panels into a shared feature space by addressing differences in data sparsity, and LLOKI-CAE, which mitigates batch effects across ST technologies while preserving biological specificity. By leveraging both components, LLOKI enables robust cross-technology integration while retaining meaningful biological variation.

Our results demonstrate that LLOKI outperforms baseline methods across key integration metrics, maintaining biological variation and effectively correcting batch effects—even when gene panel overlap is minimal. LLOKI's integrated embeddings support challenging downstream tasks such as spatial gene program identification and slice alignment, underscoring its versatility in addressing the inherent complexities of ST data integration and paving the way for more comprehensive analysis of tissue architecture across diverse technologies.

Several avenues exist for extending LLOKI's capabilities. Refinements to LLOKI-FP—such as improved mitigation of technology-specific biases via optimal transport—could enhance the alignment between ST and scRNA-seq data. Similarly, the conditional autoencoder design of LLOKI-CAE offers promising generalization capabilities: by updating only the conditional weights while retaining shared parameters, LLOKI can be readily adapted to new ST platforms without full retraining. As the field of scFMs evolves, our approach can be extended to incorporate emerging models, further enhancing integration performance.

Beyond methodological advancements, LLOKI has several practical applications. One notable use case is leveraging samples from one ST technology as controls for disease samples collected using a different technology, reducing data requirements and conserving resources. Moreover, LLOKI could be applied to the integration of atlas-scale data sets, enabling large-scale, cross-technology analyses to identify rare cellular niches without necessitating prohibitively large sample sizes. Overall, LLOKI provides a powerful tool for harmonizing ST data sets across different platforms and gene panels, facilitating a deeper understanding of spatial cellular organization in diverse biological contexts.

By addressing key challenges in ST data integration, LLOKI lays the groundwork for scalable, cross-technology spatial transcriptomics analysis, empowering researchers to uncover biologically meaningful patterns across diverse tissue samples and experimental conditions.

Methods

LLOKI is a framework for integrating spatial transcriptomics data from diverse platforms by addressing two key alignment challenges (Fig. 1): (1) feature alignment across varying gene panels, handled by LLOKI-FP, which imputes gene expression and embeds cells using a single-cell foundation model; and (2) batch alignment across different ST technologies, handled by LLOKI-CAE via a conditional autoencoder with a novel three-part loss function. This approach allows for unified analysis of heterogeneous ST data sets while preserving essential biological signals.

Feature alignment and denoising with LLOKI-FP

scFMs offer a promising solution for aligning features across ST data sets without requiring shared gene panels (Rosen et al. 2023; Theodoris et al. 2023; Cui et al. 2024). However, ST data often appears out-of-distribution for scFMs pretrained on scRNA-seq due to differences in sparsity and gene detection sensitivity. Additionally, many scFMs process only genes with nonzero expression, making higher sparsity a major challenge, as it reduces the effective feature dimensionality and introduces variability in embedding quality across ST platforms.

To address this, LLOKI-FP employs a sparsity-matching and denoising approach, combining Wasserstein optimal transport (Peyré and Cuturi 2019) with graph-based feature propagation (Rossi et al. 2022). Rather than performing conventional gene imputation, LLOKI-FP transforms the sparsity profile of ST data to better match the scRNA-seq distribution, optimizing compatibility with scFMs and improving embedding quality (see Supplemental Table S2 for per-data set sparsity values). Unlike prior feature propagation methods for single-cell data (Lee et al. 2024), LLOKI-FP integrates spatial information and computes a sparsity prior via optimal transport, shifting the focus from mere imputation to distribution alignment.

At a high level, LLOKI-FP first determines a target sparsity for each ST cell using optimal transport. It then constructs a cellular graph incorporating both gene similarity and spatial proximity, which is used to iteratively impute missing gene expression until each cell's target sparsity is reached.

Graph construction

LLOKI-FP begins by constructing an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where nodes $v_i \in \mathbf{V}$ represent cells, and edges $e_{ij} \in \mathbf{E}$ encode similarity between cells. Because feature propagation assumes high graph homophily—where connected cells share similar gene expression—it is critical that the graph reflects true biological similarity.

To achieve this, LLOKI-FP integrates gene expression and spatial proximity into graph construction. Each cell's k-nearest neighbors (k-NNs) are identified in the gene expression space to establish graph connectivity, whereas edge weights (w: $E \rightarrow \boldsymbol{R}^+$) are set using a Radial Basis Function kernel based on spatial distance

$$w(e_{ij}) = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \quad (1)$$

where d_{ij} is the spatial distance between cells i and j and σ is a bandwidth parameter computed as the mean distance from a random sample of 100 cell pairs. This weighting reinforces graph homophily by leveraging the biological insight that physically adjacent cells tend to share microenvironments and gene expression patterns. The graph is then symmetrically normalized to ensure balanced feature propagation across cells. This graph-guided feature propagation aligns ST sparsity distributions with scRNA-seq while ensuring imputation is informed by both transcriptional and spatial information.

Feature propagation for data imputation with optimal transport

To impute sparse gene expression values, LLOKI-FP employs feature propagation (Rossi et al. 2022), enhanced with optimal transport (Peyré and Cuturi 2019) to align ST sparsity with scRNA-seq references. Given that ST data is typically more sparse than scRNA-seq data, our approach selectively imputes missing values while preserving a degree of inherent sparsity—completely eliminating sparsity would reduce embedding diversity and obscure biologically meaningful variation. Each ST slice is processed independently, with feature propagation performed on the graph constructed from that slice's assayed genes and spatial coordinates.

This process starts with the constructed similarity graph **G** with spatially informed, normalized adjacency matrix **A,** and the original gene expression matrix **X**. Gene expression is iteratively updated as follows:

1. *Feature propagation with controlled sparsity*. Imputed gene expression values are updated iteratively as

$$\mathbf{X}^{(t+1)} = (1 - \mathbf{M}) \circ \mathbf{A}\mathbf{X}^{(t)} + \mathbf{M} \circ \mathbf{X}^{(0)}, \tag{2}$$

   where $\mathbf{X}^{(0)}$ is the original gene expression matrix, **M** is a binary mask indicating observed (nonzero) entries in $\mathbf{X}^{(0)}$, and $\circ$ denotes the element-wise Hadamard product. Thus, imputation is applied only to genes included in the assayed panel but observed as zero due to dropout or sparsity. Genes with nonzero expression values are preserved, whereas unmeasured genes are excluded entirely from imputation. This update minimizes the graph Dirichlet energy as shown in Rossi et al. (2022), ensuring that imputation is both spatially coherent and biologically relevant.
2. *Sparsity matching with optimal transport*. To harmonize sparsity profiles between ST and scRNA-seq data, we compute per-cell sparsity as $s_i = 1 - \frac{\text{nonzero}(x_i)}{N}$, where nonzero($x_i$) is the number of nonzero gene expression values in cell i and N is the total number of genes. We then compute the empirical cumulative distribution function (CDF) for the sparsity values. For a sorted set $\{s_{(1)}, s_{(2)}, \ldots, s_{(n)}\}$, the CDF is defined as

$$F(s_{(i)}) = \frac{i}{n}. \tag{3}$$

   Denoting the CDFs for the ST data and the scRNA-seq reference as $F_{\text{ST}}(s)$ and $F_{\text{scRNA}}(s)$, respectively, we align the distributions by mapping each ST sparsity value s to a target value s^* via

$$s^* = F^{-1}_{scRNA}(F_{\text{ST}}(s)). \tag{4}$$

   This quantile alignment minimizes the Wasserstein distance between the distributions, yielding target sparsity levels that mirror the scRNA-seq profile (Peyré and Cuturi 2019). Subsequently, the feature propagation update is applied iteratively until each cell reaches its target sparsity, defined as having no more than $N_{\max} = \lceil (1 - s_{\text{ref}})G \rceil$ nonzero genes, where s_{ref} is the reference sparsity from scRNA-seq data and G is the number of genes. To prevent overimputation—where the imputed data becomes denser than intended—we monitor the number of nonzero entries per cell and reintroduce zeros as needed to recapitulate the natural dropout observed in scRNA-seq data. We additionally apply early stopping when the Frobenius norm of the change in imputed expression falls below a threshold ε.

This iterative process—alternating between controlled feature propagation and sparsity adjustment via 1D optimal transport mapping —ensures that LLOKI-FP embeddings preserve critical biological signals while aligning with scRNA-seq sparsity, thereby enhancing compatibility with scFMs.

Denoising with feature diffusion

Following optimal transport-enhanced feature propagation, we refine gene expression data using feature diffusion. We construct a new cell similarity graph based on the imputed gene expression matrix $\mathbf{X}_{\text{imputed}}$, with $\mathbf{A}_{\text{imputed}}$ as the corresponding adjacency matrix. During this phase, the original gene expression matrix is diffused across the new graph without modifying nonzero values, using

$$\mathbf{X}_{\text{denoised}} = \alpha\, \mathbf{A}_{\text{imputed}}\, \mathbf{X}^{(0)} + (1 - \alpha)\, \mathbf{X}^{(0)}. \tag{5}$$

This diffusion step propagates gene expression features across the spatially weighted k-NN graph, updating missing or noisy values while preserving the original gene expression measurements. The hyperparameter $\alpha \in [0, 1]$ controls the extent to which information from neighboring cells (via $\mathbf{A}_{\text{imputed}}\, \mathbf{X}^{(0)}$) influences the denoised output. A value of α close to 1 emphasizes the diffused signal from neighboring cells, whereas a value closer to 0 favors retaining the original gene expression values. This formulation leverages the assumption that cells that are both spatially and transcriptionally similar share comparable true gene expression patterns.

Feature alignment using a single-cell foundation model

Once ST sparsity is aligned to match the scRNA-seq reference and the data is denoised, we proceed with feature alignment via a scFM pretrained on scRNA-seq data. This step ensures that all data sets regardless of their original gene panels share a common feature space, allowing the full gene panel to inform cell embeddings.

scGPT (Cui et al. 2024) demonstrated strong performance across cell-type clustering metrics (Supplemental Information). It is not only computationally efficient but also particularly adept at handling smaller gene panels due to its autoregressive pretraining strategy. The final output of LLOKI-FP is a 512-dimensional embedding space, consistent across all input data sets, enabling robust feature alignment across diverse ST technologies.

We compared the performance of LLOKI-FP to scGPT alone (without feature propagation) across five ST technologies (Supplemental Information). LLOKI-FP consistently outperforms scGPT, with the most significant improvement observed on the STARmap data set (ARI of 0.423 compared to 0.197). We also conducted an ablation study evaluating the impact of the optimal transport sparsity alignment step in LLOKI-FP (Supplemental Information). This analysis revealed that incorporating optimal transport alignment significantly enhanced performance for the MERSCOPE and STARmap data sets, confirming the importance of this step in our workflow.

Batch integration with LLOKI-CAE

Cell embeddings from LLOKI-FP effectively capture biological variation within each ST data set but still exhibit large batch effects between technologies. To address this, we develop a conditional autoencoder (CAE) for batch integration, conditioned on both ST technology and specific gene panel, which allows the model to differentiate between data sets by technology or gene panel. Integration is achieved through a novel three-part loss function that aligns embeddings across technologies while preserving the structure of cell-type clusters within each data set.

Conditional autoencoder architecture

Our CAE architecture follows a standard design with some key new developments. Both the encoder and decoder comprise three layers, reducing the dimensionality of the 512-dimensional input embeddings from LLOKI-FP down to 128. Many downstream tasks require positive embeddings, so we apply a softplus activation function in the final layer of the encoder to ensure all embeddings are positive.

The conditional component of the autoencoder is implemented as a learnable embedding that maps a batch token to a 10-dimensional embedding, appended to the input of both the encoder and decoder.

Three-part loss function

Conventional methods for integrating ST data sets—such as STAligner (Zhou et al. 2023)—typically balance cell-type heterogeneity and batch integration using a combination of reconstruction loss and triplet loss. Although triplet loss effectively promotes batch integration, we found that reconstruction loss alone does not sufficiently preserve critical biological information. We therefore introduce a third loss function, the biological conservation loss, designed to maintain the local structure of the original LLOKI-FP embeddings, which inherently capture robust biological signal across cell types. By jointly optimizing reconstruction loss, triplet loss, and our new biological conservation loss, our approach aligns embeddings across batches while preserving essential biological information.

1. **Reconstruction loss.** The reconstruction loss seeks to recreate the original LLOKI-FP embeddings for each cell X. Given the decoder output Y, we minimize the mean squared error (MSE) between X and Y, ensuring that the CAE preserves original information through the encoding-decoding process

$$\mathcal{L}_{\text{rec}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2. \tag{6}$$

2. **Triplet loss.** Triplet loss encourages embeddings of biologically similar cells from different technologies to be closer, while pushing apart embeddings of dissimilar cells within the same technology. Anchor-positive pairs consist of a cell from one technology (anchor) and its mutual nearest neighbor from another. Negative pairs are randomly selected cells from the same slice as the anchor, which are likely to be dissimilar. When reliable cell-type annotations are available, we optionally refine this selection by: (1) restricting positive pairs to cells sharing the same cell type across batches; and (2) selecting negative pairs using hard negative sampling, where the negative pair is chosen as the closest cell from the same batch that belongs to a different cell type than the anchor. The triplet loss is defined as

$$\mathcal{L}_{\text{trip}} = \frac{1}{n}\sum_{i=1}^{n}\max\left(0, \left\|Z_i^{\text{anchor}} - Z_i^{\text{positive}}\right\|^2 - \left\|Z_i^{\text{anchor}} - Z_i^{\text{negative}}\right\|^2 + \alpha\right), \tag{7}$$

where Z_i is the CAE embeddings, and α is the margin separating positive and negative pairs.

3. **Biological conservation loss.** Our new biological conservation loss preserves cell-to-cell similarity by ensuring that the distances between each cell and its k-nearest neighbors in the original LLOKI-FP embedding space are maintained in the LLOKI-CAE embedding space. The biological conservation loss is defined as

$$\mathcal{L}_{\text{bc}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i}(\|Z_i - Z_j\| - \|X_i - X_j\|)^2, \tag{8}$$

where $\mathcal{N}_i$ is the set of nearest neighbors of cell i in the original embedding space.

To balance these components, we introduce weighting parameters λ_{rec}, λ_{trip}, and λ_{bc}, which are tuned to determine the optimal balance among the three competing objectives. The total loss is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{trip}}\mathcal{L}_{\text{trip}} + \lambda_{\text{bc}}\mathcal{L}_{\text{bc}}. \tag{9}$$

These weights are optimized to ensure that each aspect of the loss appropriately influences training, preserving cell-type clustering while effectively reducing batch effects across technologies.

Further details about hyperparameter choices for LLOKI-CAE are in Supplemental Information.

Evaluation metrics for batch integration

To quantitatively evaluate LLOKI and compare it to baselines for batch integration, we used eight metrics from scib-metrics (Luecken et al. 2022). Five of these metrics assessed the biological variation preserved by the embeddings, based on cell-type separation using cell-type annotations provided with each data set: isolated labels silhouette score, k-means clustering NMI and ARI (KMeans NMI and KMeans ARI), silhouette width (Silhouette label), and cell-type local inverse Simpson index (cLISI). The remaining three metrics evaluated batch effect removal, measuring how well technologies were integrated into a shared embedding space: silhouette batch score, batch connectivity in a cell type–specific subgraph (graph connectivity), and principal component regression comparison (PCR comparison). Aggregate scores were computed by averaging biological conservation and batch correction metrics.

Scalability of LLOKI

LLOKI demonstrates robust scalability for processing large ST data sets. The computational framework consists of two main components with complementary performance characteristics. LLOKI-FP processes each ST slice independently, resulting in computational requirements that scale linearly with the number of slices. For a typical slice containing around 50,000 cells, running LLOKI-FP takes ~5 min and uses 20 GB of memory. Importantly, this component can be fully parallelized across compute nodes to process multiple slices simultaneously, offering substantial acceleration for large-scale studies. LLOKI-CAE employs an efficient batching approach during training and maintains a lightweight memory footprint of <2 GB regardless of data set size. For the five-slice integration task described in our results, model training

was completed in <1 h on a single GPU. Posttraining embedding generation is computationally efficient, requiring only a single forward pass operation through the trained encoder. This process takes <1 sec to generate embeddings for cells from all five slices (250,000 cells), enabling rapid integration of new data sets without retraining. Notably, only one slice of each technology is needed to train the model; processing all slices from large data sets after training takes only 5 min each and is fully parallelizable.

Slice alignment

For slice alignment with LLOKI, we implemented the Kabsch algorithm (Kabsch 1976) for image alignment and used our LLOKI embeddings to determine landmark pairs between two slices. The algorithm then seeks to minimize the root-mean-squared distance for all landmark pairs. For baselines, we ran PASTE, STAligner, GPSA, and a version of the Kabsch algorithm that uses PCA embeddings of shared gene expression instead of LLOKI embeddings. We ran PASTE with default parameters and five different values for alpha (0.01, 0.05, 0.001, 0.005, 0.0001), using the best-performing parameter set for our comparison. We note that we run PASTE instead of PASTE2 (Liu et al. 2023b), because the advantage of PASTE2 is in partial alignment, and all of our analysis was on full-slice alignments. We ran STAligner and GPSA using their default parameters.

SpiceMix analysis

For the SpiceMix analysis on the four mouse brain coronal slices from MERFISH, MERSCOPE, Xenium, and CosMx, we ran SpiceMix with the following parameter settings. We set the number of metagenes $K = 15$ and the spatial affinity regularization to 10^{-4}. We performed pretraining without optimizing spatial affinity parameters for 10 iterations, followed by 200 iterations with full optimization.

For baseline comparisons where SpiceMix was run independently on each data set using full gene panels, we constructed a cross-data set metagene mapping to enable direct comparison of metagene identities across technologies. To do this, we first computed metagene signature matrices for each data set, defined as the Pearson's correlation between each metagene and the unified cell-type annotations (excluding "Other/Unannotated" cells). These signatures capture the degree to which each metagene is associated with distinct cell types. For each pair of data sets, we computed metagene similarity based on their signature vectors and applied the Hungarian algorithm (Kuhn 1955) to derive a one-to-one metagene mapping that maximizes correspondence. Using MERSCOPE as the reference, we identified the best-matched metagene in each of the other data sets (MERFISH, Xenium, and CosMx) for each reference metagene and retained only those with consistent matches in at least two additional data sets. This mapping was then used to reorder the metagene dimensions in the independently trained data sets' SpiceMix embedding matrices, so that mapped metagenes align across data sets.

To assess the correspondence of metagenes identified by SpiceMix with cell-type annotations, we computed an enrichment matrix where each row represents a cell type and each column a metagene, using only high-confidence cell-type annotations. For each data set, we excluded "Other/Unannotated" cells, retaining seven cell-type labels. For each data set, we computed the mean metagene expression for each cell type using only nonzero values, then aggregated the results into a unified matrix. Metagenes were then reordered based on their maximum enrichment across cell types to highlight cell type–specific patterns and enable cross-technology comparisons.

We used Moran's I to evaluate spatial autocorrelation of metagene expression for each data set. Spatial neighborhood graphs were constructed using Squidpy (Palla et al. 2022). For each metagene, Moran's I was computed to quantify how strongly its expression was spatially clustered.

To further evaluate the biological relevance of individual metagenes, we conducted a marker gene analysis stratified by metagene expression. Specifically, for metagenes of interest (e.g., metagene 13 for oligodendrocytes, metagene 6 for microglia), we divided cells into high and low expression groups based on the 95th percentile of metagene expression within each data set. We then examined the expression of known cell type–specific marker genes across these groups using the original, unprocessed gene expression matrices. For each gene and group, we computed the mean expression and the fraction of expressing cells (nonzero expression), enabling comparison across technologies with nonoverlapping gene panels.

Ovarian cancer analysis

We integrated the ovarian cancer data sets using LLOKI with the following hyperparameters: $\lambda_{bc} = 500.0$, $\lambda_{trip} = 0.2$, $\lambda_{recon} = 1.0$, a learning rate of 0.0005, a chunk size of 16,000, and a batch dimension of 10. Hyperparameter tuning was performed to optimize for both a high biological conservation score and effective batch correction which was measured using scib-metrics.

To identify tumor-infiltrating T cells, we first performed Leiden clustering on T cells from all data sets, using LLOKI embeddings. We then selected clusters corresponding to regions in the UMAP with a high percentage of malignant neighbors, thereby defining a cross-technology group of tumor-infiltrating T cells.

To derive a gene list associated with tumor infiltration, we used a two-stage process. In the first stage, for each data set, we performed differential expression analysis comparing T cell populations to identify genes that were enriched in T cells. We applied filters to retain only genes expressed in at least 5% of T cells and exhibiting a minimum fold change of 0.3. This produced 816 T cell-enriched genes in all data sets. In the second stage, we split T cells by infiltration status and repeated differential expression analysis on the subset of genes identified in the first stage. These results were aggregated using a weighted rank-based aggregation method: for each data set, each gene received a normalized ranking and the final ranking was averaged across data sets.

Finally, using the shared gene list, we scored T cells for their expression of the tumor infiltration program with Scanpy, using positive weights for upregulated genes and negative weights for downregulated genes. This provided a robust metric to quantify T cell infiltration in the integrated ovarian cancer data sets.

Software availability

The source code for LLOKI is available on GitHub (https://github.com/ma-compbio/Lloki/) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported, in part, by National Institutes of Health Common Fund 4D Nucleome Program grant UM1HG011593 (J.M.); National Institutes of Health Common Fund Cellular Senescence Network Program grant UH3CA268202 (J.M.); and National Institutes of Health grants R01HG007352 (J.M.),

R01HG012303 (J.M.), R21DA061481 (J.M.), R03OD039980 (J.M.), and U24HG012070 (J.M.). J.M. was additionally supported by the Ray and Stephanie Lane Professorship, a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation, a Google Research Award, and a Single-Cell Biology Data Insights award from the Chan Zuckerberg Initiative. S.K. is a Lane Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: Conceptualization: E.H., S.K., and J.M.; methodology: E.H., S.K., and J.M.; software: E.H., A.D., and S.K.; investigation: E.H., S.K., and J.M.; writing: E.H., S.K., and J.M.; funding acquisition: J.M.

References

10x Genomics. 2023. Mouse brain coronal section using a pre-designed 248-gene Xenium Mouse Brain Gene Expression panel. https://www.10xgenomics.com/products/xenium-in-situ/mouse-brain-dataset-explorer. Accessed: Sep 3, 2025.

Argelaguet R, Cuomo AS, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* **39:** 1202–1215. doi:10.1038/s41587-021-00895-7

Bian H, Chen Y, Dong X, Li C, Hao M, Chen S, Hu J, Sun M, Wei L, Zhang X. 2024. ScMulan: a multitask generative pre-trained language model for single-cell analysis. In *Proceedings of the 28th Annual International Conference, Research in Computational Molecular Biology*, Vol. 14758, pp. 479–482, Cambridge, MA. https://dl.acm.org/doi/10.1007/978-1-0716-3989-4_57

Chidester B, Zhou T, Alam S, Ma J. 2023. SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nat Genet* **55:** 78–88. doi:10.1038/s41588-022-01256-z

Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21:** 1470–1480. doi:10.1038/s41592-024-02201-0

CZI Cell Science Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Cakir B, Chaffer J, Chambers S, et al. 2025. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* **53:** D886–D900. doi:10.1093/nar/gkae1142

Guo T, Yuan Z, Pan Y, Wang J, Chen F, Zhang MQ, Li X. 2023. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biol* **24:** 241. doi:10.1186/s13059-023-03078-6

Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, Wang T, Ma J, Zhang X, Song L. 2024. Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21:** 1481–1491. doi:10.1038/s41592-024-02305-7

Hartman A, Satija R. 2024. Comparative analysis of multiplexed in situ gene expression profiling technologies. *eLife* **13:** RP96949. doi:10.7554/eLife.96949

He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, Danaher P, Dunaway D, Garrison RG, Geiss G, et al. 2022. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* **40:** 1794–1806. doi:10.1038/s41587-022-01483-z

Hu Y, Xie M, Li Y, Rao M, Shen W, Luo C, Qin H, Baek J, Zhou XM. 2024. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biol* **25:** 212. doi:10.1186/s13059-024-03361-0

Jones A, Townes FW, Li D, Engelhardt BE. 2023. Alignment of spatial genomics data using deep Gaussian processes. *Nat Methods* **20:** 1379–1387. doi:10.1038/s41592-023-01972-2

Kabsch W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Section A* **32:** 922–923. doi:10.1107/S0567739476001873

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* **16:** 1289–1296. doi:10.1038/s41592-019-0619-0

Kuhn HW. 1955. The Hungarian method for the assignment problem. *Nav Res Logist Q* **2:** 83–97. doi:10.1002/nav.3800020109

Lee J, Yun S, Kim Y, Chen T, Kellis M, Park C. 2024. Single-cell RNA sequencing data imputation using bi-level feature propagation. *Brief Bioinformatics* **25:** bbae209. doi:10.1093/bib/bbae209

Liu W, Liao X, Luo Z, Yang Y, Lau MC, Jiao Y, Shi X, Zhai W, Ji H, Yeong J, et al. 2023a. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with precast. *Nat Commun* **14:** 296. doi:10.1038/s41467-023-35947-w

Liu X, Zeira R, Raphael BJ. 2023b. Partial alignment of multislice spatially resolved transcriptomics data. *Genome Res* **33:** 1124–1132. doi:10.1101/gr.277670.123

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15:** 1053–1058. doi:10.1038/s41592-018-0229-2

Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19:** 41–50. doi:10.1038/s41592-021-01336-8

Mallach A, Zielonka M, van Lieshout V, An Y, Khoo JH, Vanheusden M, Chen W-T, Moechars D, Arancibia-Carcamo IL, Fiers M, et al. 2024. Microglia-astrocyte crosstalk in the amyloid plaque niche of an Alzheimer's disease mouse model, as revealed by spatial transcriptomics. *Cell Rep* **43:** 114216. doi:10.1016/j.celrep.2024.114216

Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al. 2022. Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* **19:** 171–178. doi:10.1038/s41592-021-01358-2

Peyré G, Cuturi M. 2019. Computational optimal transport. *Found Trends Mach Learn* **11:** 355–607. doi:10.1561/2200000073

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. 2017. The human cell atlas. *eLife* **6:** e27041. doi:10.7554/eLife.27041

Rosen Y, Roohani Y, Agrawal A, Samotorcan L, Consortium TS, Quake SR, Leskovec J. 2023. Universal cell embeddings: a foundation model for cell biology. bioRxiv doi:10.1101/2023.11.28.568918

Rossi E, Kenlay H, Gorinova MI, Chamberlain BP, Dong X, Bronstein MM. 2022. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Proceedings of the First Learning on Graphs Conference*, PMLR **198:** 11:1–11:16. PMLR, Cambridge, MA.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33:** 495–502. doi:10.1038/nbt.3192

Shi H, He Y, Zhou Y, Huang J, Maher K, Wang B, Tang Z, Luo S, Tan P, Wu M, et al. 2023. Spatial atlas of the mouse central nervous system at molecular resolution. *Nature* **622:** 552–561. doi:10.1038/s41586-023-06569-5

Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. 2023. Single Cell Portal: an interactive home for single-cell genomics data. bioRxiv doi:10.1101/2023.07.13.548886

Tejada-Lapuerta A, Schaar AC, Gutgesell R, Palla G, Halle L, Minaeva M, Vornholz L, Dony L, Drummer F, Richter T, et al. 2025. Nicheformer: a foundation model for single-cell and spatial omics. *Nat Methods* doi:10.1038/s41592-025-02814-z

Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. 2023. Transfer learning enables predictions in network biology. *Nature* **618:** 616–624. doi:10.1038/s41586-023-06139-9

Wang C, Cui H, Zhang A, Xie R, Goodarzi H, Wang B. 2025. scGPT-spatial: continual pretraining of single-cell foundation model for spatial transcriptomics. bioRxiv doi:10.1101/2025.02.05.636714

Xu C, Jin X, Wei S, Wang P, Luo M, Xu Z, Yang W, Cai Y, Xiao L, Lin X, et al. 2022. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res* **50:** e131. doi:10.1093/nar/gkac901

Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, Lu H, Yao J. 2022. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* **4:** 852–866. doi:10.1038/s42256-022-00534-z

Yao Z, van Velthoven C, Kunst M, Zhang M, McMillen D, Lee C, Jung W, Goldy J, Abdelhak A, Aitken M, et al. 2023. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624:** 317–332. doi:10.1038/s41586-023-06812-z

Yeh CY, Aguirre K, Laveroni O, Kim S, Wang A, Liang B, Zhang X, Han LM, Valbuena R, Bassik MC, et al. 2024. Mapping spatial organization and genetic cell-state regulators to target immune evasion in ovarian cancer. *Nat Immunol* **25:** 1943–1958. doi:10.1038/s41590-024-01943-5

Zeira R, Land M, Strzalkowski A, Raphael BJ. 2022. Alignment and integration of spatial transcriptomics data. *Nat Methods* **19:** 567–575. doi:10.1038/s41592-022-01459-6

Zhang M, Pan X, Jung W, Halpern A, Eichhorn S, Lei Z, Cohen L, Smith K, Tasic B, Yao Z, et al. 2023. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624:** 343–354. doi:10.1038/s41586-023-06808-9

Zhou X, Dong K, Zhang S. 2023. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nat Comput Sci* **3:** 894–906. doi:10.1038/s43588-023-00528-w

Received April 19, 2025; accepted in revised form September 8, 2025.