Nucleic Acids Research, 2026, **54**, gkag621
https://doi.org/10.1093/nar/gkag621
Methods

Optimal gene panel selection for targeted spatial transcriptomics experiments

Haoran Lu [1], Luyang Fang [1], Orlando Zeng [1], Wenxuan Zhong [1,*], Guo-Cheng Yuan [2,*], Ping Ma [1,*]

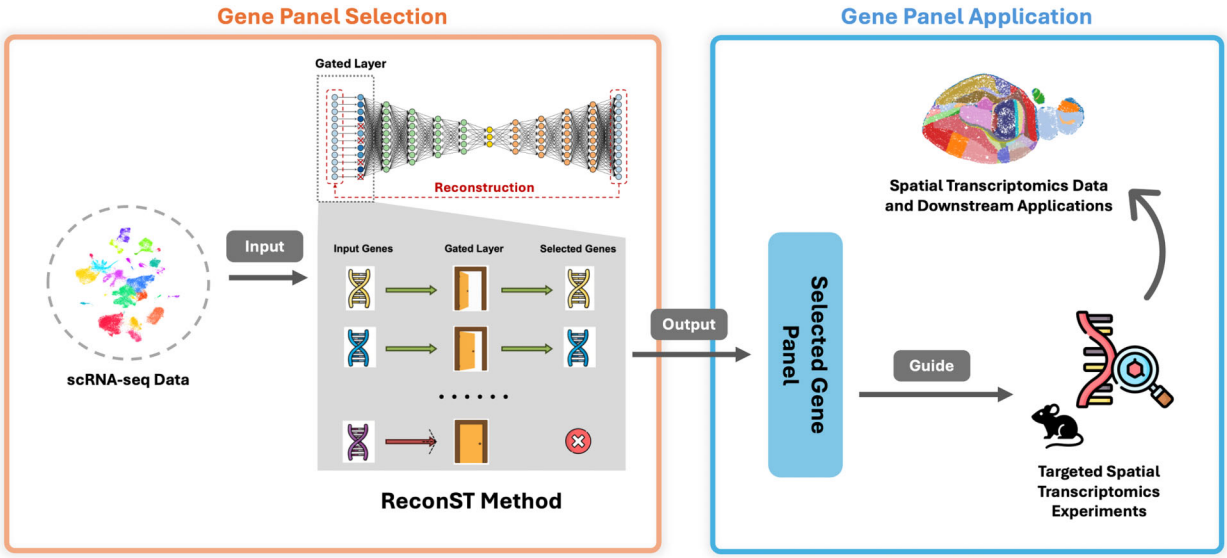
[1]Department of Statistics, University of Georgia, Athens, GA 30602, United States
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States

*To whom correspondence should be addressed. Email: wenxuan@uga.edu
Correspondence may also be addressed to Guo-Cheng Yuan. Email: guo-cheng.yuan@mssm.edu
Correspondence may also be addressed to Ping Ma. Email: pingma@uga.edu

Abstract

Spatial transcriptomics analysis is a powerful approach for dissecting the structure of tissue microenvironment and uncovering the mechanism of cell–cell communications. However, existing technologies are limited by either spatial resolution or gene coverage. Most single-cell-resolution technologies target only a few hundred preselected genes, whose choice plays an important role in the overall analysis. It remains a challenge to optimally design a gene panel to maximize the utility of spatial transcriptomics profiling. To fill this gap, we introduce a novel method, named ReconST, to automatically design optimal gene panels for spatial transcriptomics profiling. ReconST leverages information from existing scRNA-seq data and identifies the optimal subset of genes by using a gated autoencoder. By using a high-coverage mouse brain MERFISH dataset and a fetal lung dataset as the reference benchmarks, we showed that ReconST outperforms existing methods in terms of reconstruction accuracy, spatial pattern preservation, and computing efficiency. As such, ReconST provides a useful and generally applicable tool for optimal gene panel design, which in turn can significantly enhance the utility of spatial transcriptomics profiling in a wide range of biomedical investigations.

Graphical abstract



Introduction

Spatial transcriptomics is an emerging and transformative technique that enables researchers to study gene expression while preserving spatial context, providing unprecedented insights into tissue organization and function [1–5]. The increasing availability of commercial platforms highlights the growing significance of spatial transcriptomics in biological studies [6, 7]. In contrast to traditional single-cell RNA sequencing (scRNA-seq) techniques, spatial transcriptomics stands out by preserving the spatial information, which is crucial for dissecting the structure of tissue microenvironment and uncovering cell-cell communication mechanisms [3, 8].

Received: November 6, 2025. **Revised:** April 28, 2026. **Accepted:** May 25, 2026

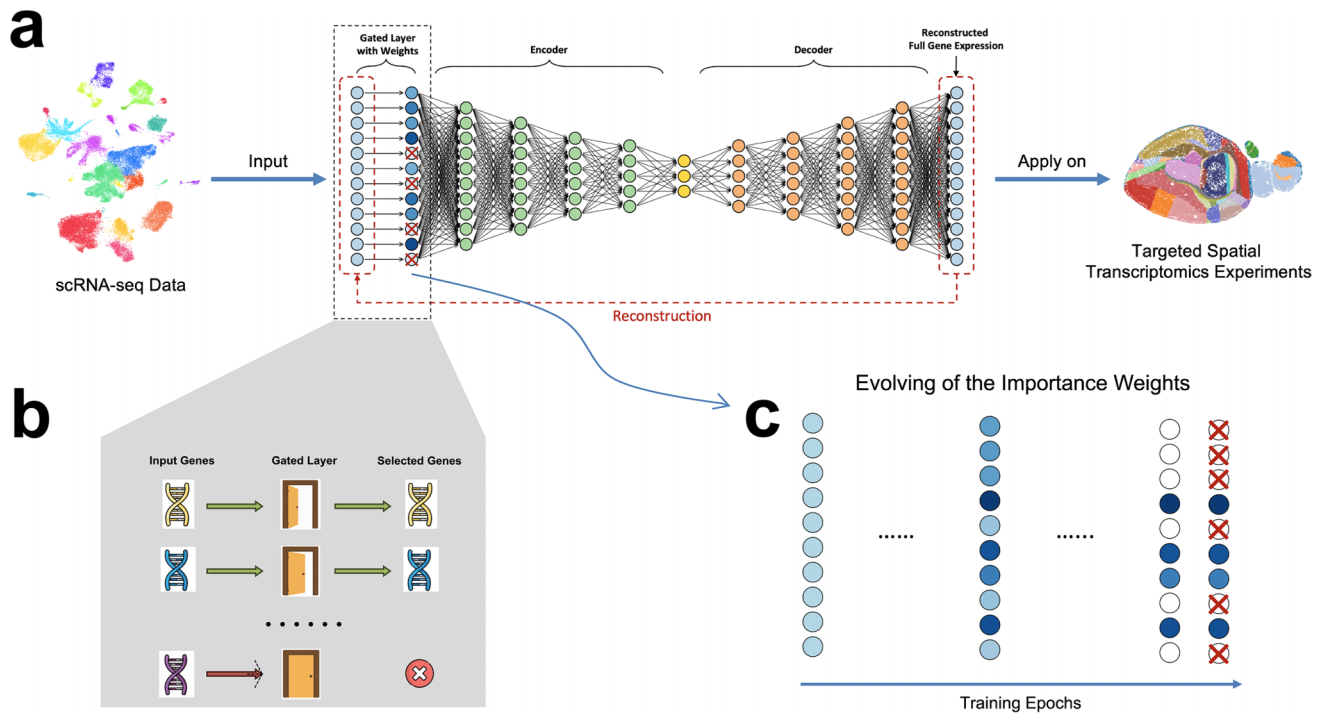


Figure 1. Using ReconST to design an optimal gene panel for spatial transcriptomics analysis. (**a**) A schematic overview of the ReconST workflow using scRNA-seq data as input. (**b**, **c**) A zoomed-in view of the gated layer. (b) A schematic illustration of the gated layer. (c) An illustration of the effect of the gated layer on parameter training. Unselected genes are marked by the "x" mark.

Current spatial transcriptomics technologies can broadly be categorized into two types: whole-transcriptome methods and targeted methods [9]. Whole-transcriptome spatial transcriptomics aims to measure genome-wide gene expression across tissue sections, often using techniques such as Slide-seq [10], 10x Genomics Visium [11, 12], Nanostring GeoMx Digital Spatial Profiler [13], and DBiT-seq [14], though these methods may sacrifice resolution or sensitivity [15].

In contrast, targeted spatial transcriptomics, such as MERFISH [16], seqFISH [17], seqFISH+ [18], Xenium [19], and STARMAP [20], focus on profiling a predefined subset of genes, which is needed to achieve single-cell or subcellular resolution in spatial gene expression profiling [16, 18]. The advantage of single-cell resolution mapping is that it can significantly enhance the ability to study complex tissue architecture and cell-cell interactions [18, 21, 22]. However, its utility strongly depends on the biological information contained in the gene panel, which has to be pre-selected before spatial transcriptomics mapping. As such, an important but difficult task is to design the optimal gene panel for a biological problem at hand.

Although it is possible to manually select genes based on prior biological knowledge of cell-type-specific marker genes or biological pathways, this approach has important practical limitations, including subjectivity, irreproducibility, and strong dependence on prior knowledge. Recently, several methods have been developed to systematically design gene panels by leveraging information from existing single-cell RNA-seq (scRNA-seq) data. Existing gene panel design methods vary in their selection criteria, including highly variable genes [16], marker genes associated with either annotated cell types [23] or original cell clusters [24], and the genes representative of the underlying manifold structure [25]. Despite these developments, it remains challenging to systematically design a gene panel that can optimally summarize the whole transcriptome information.

To overcome this challenge, we develop a novel method, named ReconST, for optimal gene panel design. Instead of targeting known marker genes, ReconST aims to unbiasedly select a subset of genes that optimally reconstruct the whole transcriptome information. By using a novel gated autoencoder model design, ReconST learns to compress and reconstruct gene expression profiles, ensuring that the selected genes retain as much meaningful variation as possible. ReconST offers a principled and generally applicable solution for constructing gene panels from scRNA-seq data that are well-suited for spatial transcriptomic applications. By leveraging the information from a high-coverage spatial transcriptomics dataset [26], we systematically compared the performance of ReconST and existing methods, and showed that it outperforms existing methods in cell-type reconstruction, recovery of underlying spatial patterns, and computational efficiency. We developed the ReconST software as a Python package, which includes a tutorial example to train models, select and export gene panels ready for targeted spatial transcriptomics studies.

Material and methods

Overview of ReconST

ReconST is a principled data-driven method that learns to select genes directly optimized for full-transcriptome reconstruction. ReconST aims to recover as much information from the complete expression landscape as possible. At its core is a gated autoencoder architecture, a neural network designed to compress and reconstruct gene expression profiles

while automatically learning which genes are most informative (Fig. 1a–c). The gating mechanism assigns learnable importance weights to each gene, and effectively selects a subset that best supports accurate reconstruction. By training on scRNA-seq data, ReconST produces a compact panel that is later validated on spatial transcriptomics datasets, ensuring both transcriptomic fidelity and spatial relevance. While neural networks have previously been used in single-cell analysis, our key innovation is the gated selection layer, which dynamically assigns importance weights to each gene. Unlike combinatorial search-based methods that require exhaustive evaluation of different gene subsets, ReconST optimizes gene selection in an end-to-end fashion using stochastic gradient descent [27]. This enables the model to efficiently learn which genes contribute most to transcriptome reconstruction, making the process highly scalable and free from heuristic selection biases.

The gated layer directly controls gene selection during training (Fig. 1b). Unlike standard autoencoders that process all input features equally, the gated layer introduces a learnable weight for every gene, allowing the model to decide which genes are essential and which can be suppressed (Fig. 1c). To further enhance efficiency and precision, we impose an L1 penalty on these weights, automatically shrinking those of non-essential genes toward zero. This sparsity-inducing regularization technique is well-established [28, 29], effectively eliminating uninformative parameters by driving their weights to exactly zero. The L1 penalty thus acts as a built-in filter, discouraging the model from assigning high importance to too many genes and ensuring that only the most informative ones are retained. After training ReconST on the scRNA-seq data, gene selection is guided by the distribution of learned weights: many are driven to zero, while non-zero weights identify genes informative for transcriptome reconstruction. These non-zero weights define the selected gene panel in a fully data-driven manner, removing the need for manual curation while preserving essential biological signals. In cases where a fixed panel size is required, the non-zero weights can be treated as importance scores, and the top-ranked genes are selected accordingly. For example, we constructed a 200-gene panel and evaluated its performance on the MERFISH dataset using various metrics. The strength of the L1 penalty is tuned by cross-validation to balance sparsity and reconstruction accuracy [30] (see the "Materials and methods" section for details).

Statistical formulation of the ReconST method

We present the detailed statistical formulation of ReconST. At its core, ReconST uses an autoencoder architecture [31], a neural network that compresses input data into a low-dimensional latent representation and reconstructs the original data from this representation. The key innovation of ReconST over traditional autoencoders is that we introduce a gate layer, positioned right after the input. The gating layer assigns non-negative learnable importance weights to each gene. Genes with zero weight are excluded from the reconstruction process, while those with higher weights are prioritized, allowing the model to focus on the most informative genes and effectively filter out noise.

Following the gate layer, the autoencoder is composed of two parts: an encoder and a decoder. The input to the encoder is the gated full gene expression profile. The encoder consists of fully connected layers with gradually decreasing dimensions to map the high-dimensional gene input into a compressed representation in a lower-dimensional latent space. Then, the decoder with increasing number of nodes reconstructs the original input from this representation. Together, the encoder and decoder form a bottleneck architecture, in which the latent space serves as a constraint that forces the network to capture the most salient features of the data. This structure prevents the model from learning a trivial identity mapping and encourages the extraction of compact and meaningful representations. Finally, the decoder output is the reconstructed full gene expression profile.

Mathematically, the input gene expression vector for the i-th cell is denoted as $\boldsymbol{X}_i \epsilon R^G$, where G is the total number of genes. The key trainable gating layer is introduced immediately after the input layer, assigning a learnable scalar weight w_j to each gene j. These weights form a gating vector $\boldsymbol{w} = (w_1, w_2, \cdots, w_G)^T$, which modulates the input gene expression vector $\boldsymbol{X}_i$ through element-wise multiplication to produce a gated input $\boldsymbol{X}_i' = \boldsymbol{w} \odot \boldsymbol{X}_i$, where $\odot$ denotes element-wise multiplication. Genes with near-zero weights w_j are suppressed, while those with non-zero weights are prioritized for reconstruction. The encoder then maps the gated input $\boldsymbol{w} \odot \boldsymbol{X}_i$ to a compressed latent representation $\boldsymbol{h}_i = f_{\boldsymbol{\theta}}(\boldsymbol{w} \odot \boldsymbol{X}_i)$ using a multi-layer perceptron structure [32]. This encoder comprises H sequential nonlinear transformations:

$$f_{\boldsymbol{\theta}}(\boldsymbol{w} \odot \boldsymbol{X}_i) = \mathrm{T}_H \circ \sigma \circ \mathrm{T}_{H-1} \circ \sigma \circ \cdots \sigma \circ T_1 (\boldsymbol{w} \odot \boldsymbol{X}_i),$$

where each T_j represents an affine transformation $T_j(\boldsymbol{u}) = \boldsymbol{v}^{(j)}\boldsymbol{u} + \boldsymbol{b}^{(j)}$, $j = 1, \cdots, H$ parameterized by weights $\boldsymbol{v}^{(j)} \epsilon R^{d_j \times d_{j-1}}$ and biases $\boldsymbol{b}^{(j)} \epsilon R^{d_j}$. Here, d_j denotes the dimension of the j-th layer, and $\sigma(\cdot)$ is the LeakyReLU activation function [32]. The encoder parameters $\boldsymbol{\theta}$ are the collection of all $\boldsymbol{v}^{(j)}$ and $\boldsymbol{b}^{(j)}$. Similarly, the decoder $g_{\boldsymbol{\phi}}$, parameterized by $\boldsymbol{\phi}$, reconstructs the input from the latent representation $\boldsymbol{h}_i$:

$$\hat{\boldsymbol{X}}_i = g_{\boldsymbol{\phi}}(\boldsymbol{h}_i) = g_{\boldsymbol{\phi}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{w} \odot \boldsymbol{X}_i)\right),$$

aiming to minimize the reconstruction error between the original input $\boldsymbol{X}_i$ and the reconstructed output $\hat{\boldsymbol{X}}_i$.

To encourage sparsity in gene selection, we imposed an L_1 penalty on the gating weights $\boldsymbol{w}$. The model is trained by minimizing a composite loss function:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{X}, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{X}_i - g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\boldsymbol{w} \odot \boldsymbol{X}_i)) \right\|_2^2 + \lambda \sum_{j=1}^{G} |w_j|,$$

where N is the total number of samples, λ controls the strength of the L_1 regularization, and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$ denotes the whole training dataset. The first term in the loss function quantifies the reconstruction error using mean squared error, while the second term encourages sparsity by L_1 regularization of the gating weights.

By optimizing this loss, ReconST learns to retain only genes with non-zero w_j values, which are most critical for reconstructing the full transcriptome, while discarding uninformative genes. The L1 regularization leads the weights of non-informative genes to shrink toward zero [28, 33, 34] and make the selected gene subset sparse, reducing redundancy and achieving our goal of gene subset selection. As a result, ReconST automatically adjusts these weights to suppress irrelevant genes and emphasize those critical for accurate reconstruction of the full gene expression profile.

After model training, genes with non-zero gate weights are selected as the informative gene panel. This approach allows the model to automatically determine a data-driven subset of genes that contribute meaningfully to reconstruction. In cases where a pre-specified number of genes is required and this

number is smaller than the total number of non-zero weights, the gate weights are treated as importance scores, and the top-ranking genes are selected accordingly.

Data source and preprocessing

To construct and evaluate gene panels under realistic experimental settings, we used two complementary datasets spanning different tissues and measurement modalities. The first dataset consists of scRNA-seq and spatial transcriptomics data from mouse brain [26], including MERFISH measurements, enabling gene panel selection from scRNA-seq and evaluation in a spatial context. The second dataset is a human fetal lung dataset consisting of scRNA-seq data and corresponding spatial transcriptomics measurements generated using Xenium. Details of each dataset and preprocessing steps are described below.

Mouse brain data

We used matched single-cell RNA-seq and MERFISH resources from the Allen Brain Cell Atlas as described in [26]. The single-cell reference dataset is the cortex subset. The spatial validation dataset is the subset from file Zhuang-ABCA-4-log2.h5ad, along with the associated cell metadata, gene annotations, and CCF section coordinates. In the main analysis, for computational efficiency, we used a randomly selected regional subset consisting of 31 299 cells from the neocortex. To further assess robustness and scalability, we additionally constructed a larger and more representative dataset by sampling 10% of cells from each of the 24 brain region datasets, resulting in a total of 404 298 cells. This stratified multi-region sampling scheme is used in an extra robustness analysis to evaluate the stability of gene selection and performance across broader training inputs.

For scRNA-seq, cells with fewer than 200 detected genes and genes expressed in fewer than 100 cells were removed. The data were then normalized by library size to a target of 10 000 counts, followed by log transformation and scaling with a maximum absolute value of 10. MERFISH measurements were used as provided in log scale. Gene symbols were harmonized between modalities and intersected to obtain a shared gene set. CCF section x–y coordinates were linked to cell identifiers for downstream spatial analyses.

Fetal lung data

The fetal lung dataset is obtained from a human fetal lung scRNA-seq atlas [35]. Raw count matrices, gene annotations, and cell-level metadata are loaded from the original dataset. We focus on gestational week 18 (GW18), as this stage has corresponding spatial transcriptomics data available for downstream spatial visualization and validation.

Cells are first subset to GW18 prior to analysis. Genes with zero total expression across all selected cells are removed to eliminate uninformative features. The data are then normalized by total counts per cell, scaled to 10 000, followed by log transformation. No additional gene filtering or feature restriction is applied, ensuring that gene selection methods operate on the full expressed transcriptome within this subset.

Benchmark methods for gene panel selection

We benchmarked the performance of ReconST against the following methods: (i) Default Panel: For the mouse brain study, we used Vizgen 1000 Plex Gene Panel Lists for Mouse Brain [36], and for the fetal lung study, we used a panel from fetal lung dataset study itself [35]; (ii) Differentially Expressed Genes (DE) [37]: genes selected based on differential expression analysis performed on the scRNA-seq data; (iii) Highly Variable Genes (HVG): top variable genes identified from scRNA-seq data, a widely used heuristic for feature selection in single-cell studies; (iv) Spapros [15]: a method that selects gene set specificity for cell type identification while capturing within-cell type expression variation; (v) PERSIST [38]: a panel design framework that integrates scRNA-seq information to preserve global transcriptomic structure, ensuring selected genes approximate the full transcriptome; (vi) Triku [39]: an unsupervised method that prioritizes genes with localized expression patterns, thereby enhancing recovery of spatially informative features; (vii) COSG [40]: a correlation-based marker selection algorithm that efficiently identifies cell-type-specific genes with low redundancy; (viii) scPNMF [23]: a sparse projective non-negative matrix factorization approach that selects informative genes through low-dimensional parts-based representation learning; (ix) GeneBasis [25]: a method that identifies compact gene sets that preserve transcriptional diversity and local neighborhood structure across cell types; and (x) NS-Forest [24]: a random-forest-based approach that identifies a minimal set of necessary and sufficient marker genes for cell-type discrimination.

Evaluation

Evaluation was performed using MERFISH data for the mouse brain study and scRNA-seq data for the fetal lung study. To ensure a comprehensive assessment of gene selection quality, we employed a series of complementary evaluation metrics. First, we computed the reconstruction loss using the mean squared error between the full and reconstructed gene expression profiles in the MERFISH dataset. Second, we quantified the global similarity between the reconstructed and full gene expression distributions using the Kullback–Leibler (KL) divergence [41]. Third, we used the Pearson and Spearman correlations to quantify the similarity of mean expression of the full and selected gene panels in the MERFISH dataset. Further, to evaluate the cell-type resolution of each selected gene panel, we computed the adjusted Rand index (ARI) [42] at multiple annotation levels, including division, class, subclass, supertype, and cluster. This was done by performing Leiden clustering [43] based on each method's selected gene set and comparing the resulting clusters with the annotated cell-type level from this dataset. Finally, we assessed spatial pattern preservation by calculating Moran's I index [44], a standard measure of spatial autocorrelation, which reflects the degree to which the spatial arrangement of selected genes recapitulates that of the full transcriptome. Finally, runtime was measured as the wall-clock time required to obtain each panel from scRNA-seq. Together, these metrics provide a multifaceted evaluation of each method's ability to recover transcriptomic structure, preserve spatial organization, and distinguish biologically meaningful cell types.

Implementation details

For the proposed ReconST method, we use a symmetric autoencoder structure where the sizes of input and gated layers both equal the dimension of genes. This is followed by layers with 512, 512, 256, 256, and 128 nodes in the encoder.

The decoder mirrors this structure with layers of 256, 256, 512, 512, and finally the same number of nodes as the gene dimension in the output layer. In our example, the gene dimension is 1147. Each hidden layer (other than the gated layer and the output layer) is followed by a LeakyReLU [45] activation function with a negative slope of 0.3 and a dropout layer [46] with a dropout probability of 0.05. The model was trained using the Adam optimizer [47] with a learning rate of 0.001 and weight decay of 0.00001, and training proceeded for 1000 epochs. All hyperparameters, including the learning rate, weight decay, dropout rate, and LeakyReLU slope, were selected via grid search based on a validation set separated from the training set. The L_1 penalty was applied only to the training procedure, while test loss was computed using only the reconstruction term to assess the reconstruction performance.

All analyses were performed mainly in Python, with some benchmark methods implemented in R. Scanpy handled preprocessing, highly variable genes, differential expression, neighbors, and Leiden. Squidpy was used for Moran's I index computation. Scikit-learn provided evaluation metrics including the ARI. PyTorch was used for the autoencoder and reconstruction loss, and NumPy, SciPy, and pandas supported numerical operations and data handling; Matplotlib and Seaborn were used for plotting. Spapros, COSG, NSForest, and PERSIST were run from their public Python implementations; scPNMF and GeneBasis were run via their public R packages.

Results

Performance evaluation and benchmark using a mouse brain MERFISH dataset

To benchmark the performance of ReconST and existing methods, we utilized a publicly available dataset with a high-resolution transcriptomic and spatial atlas of the mouse brain [26]. The dataset provides both scRNA-seq and multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) data, allowing gene selection using scRNA-seq and validation in spatial contexts. The original scRNA-seq dataset comprises ~7 million cells and 19 000 genes, with 4 million high-quality cells passing stringent quality control. The data were collected from 267 donor mice, ensuring robust representation across biological replicates. While most commercially available spatial transcriptomics platforms only allow the profiling of a few hundred genes, this MERFISH dataset contains spatial profiles of 4.3 million cells measured on a panel of 1147 genes, thereby allowing the opportunity to evaluate and compare the performance of different smaller gene panel designs in reconstruction of both transcriptomic and spatial patterns.

We compare ReconST against the ten benchmark methods, including the default panel by Vizgen [36] for mouse brain, as detailed in the "Materials and methods" section. For fair comparison, we apply each method to select 200 genes out of the 1147 genes used in the MERFISH dataset. To mimic real application setting, the gene panel selection was based on scRNA-seq expression data, while reserving the MERFISH data for performance evaluation. Of note, the scRNA-seq and MERFISH data were obtained from the same donor mouse, thereby providing an excellent example for performance evaluation. After preprocessing the scRNA-seq data with the SCANPY pipeline [48], a total of 31 299 cells and 3295 genes were retained. Details of preprocessing are provided in the "Materials and methods" section. Further analysis identified 2019 highly variable genes, among which 1014 genes were included in the MERFISH dataset. For model evaluation, we tested each gene panel regarding the full transcriptomic, cluster, and spatial information from the entire MERFISH dataset, which contains 215 278 cells and 1014 genes after preprocessing. This carefully curated subset provides a representative foundation for evaluating our gene selection method while ensuring computational feasibility and alignment with spatial transcriptomics experiments. We included a comprehensive set of quantitative metrics, including the reconstruction loss, KL divergence, Pearson and Spearman, clustering ARI, Moran's I index, and running time. Details of the evaluation metrics are introduced in the "Materials and methods" section.

We first illustrate the ReconST results by an exploratory analysis shown in Fig. 2a with an example that the UMAP visualization [49] is based on the ReconST-selected panel versus compared to the UMAP derived from the full reference panel. We see that the selected subset successfully preserves the overall cell-type clustering structure and maintains resolution across diverse populations. Further, we quantitatively assess the similarity with two metrics: the pairwise distances correlation [50] to assess global geometry preservation and the Jaccard index of the k-nearest neighbor graph [51] to assess similarity of local neighborhood structure. We observe a pairwise distance correlation of 0.792 and a k-nearest-neighbor Jaccard index of 0.726, indicating excellent global and local structure preservation of the ReconST selected gene panel.

The comparative performance of the evaluated methods on the MERFISH dataset is presented in Fig. 2b, with the methods ordered by their average rank across all performance metrics, excluding the running time. ReconST achieves the strongest overall performance across nine out of the ten metrics. In terms of reconstructing the whole transcriptome, ReconST attains the lowest reconstruction loss and KL divergence, together with the highest Pearson and Spearman correlations, demonstrating the closest agreement with the full expression distribution across both linear and monotonic relationships. Its clustering performance is the first or second best in various annotation levels, with the highest ARI at the class, subclass, supertype, and cluster levels, confirming that the selected panel retains sufficient information for accurate unsupervised cell-type discrimination at both coarse and fine resolutions. In terms of spatial patterns, ReconST attains the highest Moran's I index, indicating superior preservation of spatial autocorrelation patterns.

Among the comparison methods, Spapros also performs strongly and is the closest competitor to ReconST across most metrics, which is consistent with its design to maximize cell-type specificity while preserving within-type variation. In particular, it achieves the second-best performance in reconstruction, clustering, and spatial preservation. PERSIST, which also employs an autoencoder-based approach similar to ReconST, achieves reasonable performance but remains clearly behind ReconST in both reconstruction and clustering accuracy while requiring 8 h of runtime compared to ReconST's 12 min. This dramatic difference stems from ReconST's gated layer with L1 penalty that enables efficient gradient-based training, whereas PERSIST's optimization scheme is computationally intensive. The heuristic methods HVG and DE are computationally lightweight and achieve moderate performance, but their gene selection criteria do not explicitly optimize for

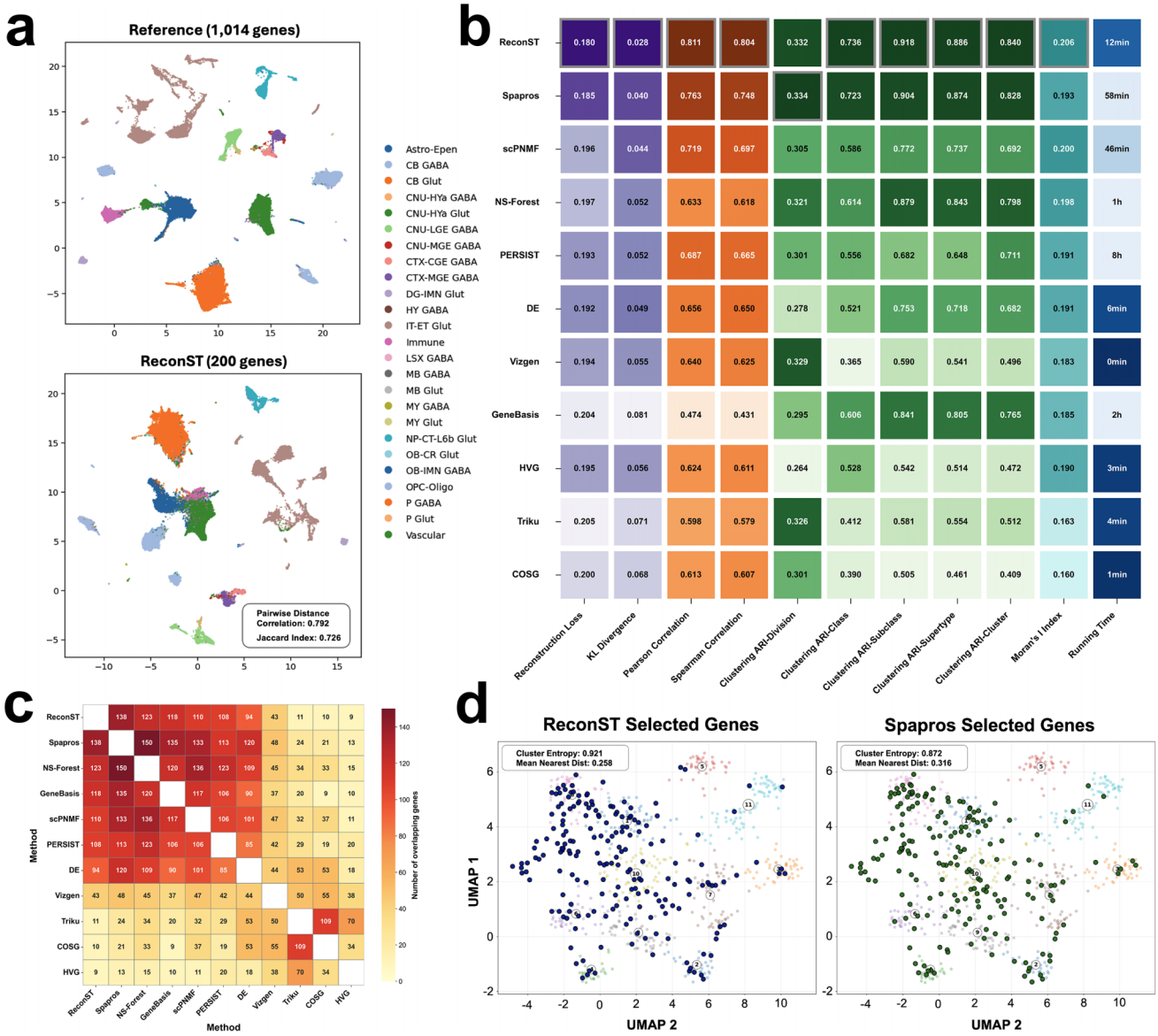


Figure 2. Comparison of the gene panel design by different methods on the mouse brain study. (**a**) UMAP visualization of MERFISH data colored by annotated cell types from the reference dataset. (**b**) Benchmark analysis of different gene panel selection methods using the MERFISH dataset [26] as reference. In each column, the best performer is outlined. (**c**) Gene panel overlap of the methods. (**d**) UMAP showing expression patterns of selected genes by ReconST and Spapros, respectively.

reconstruction or spatial objectives. DE performs noticeably better than HVG in the clustering metrics, especially at the subclass, supertype, and cluster levels, although both remain less competitive in overall transcriptome recovery. The default Vizgen panel, while convenient as a pre-designed solution, shows limited performance with relatively low clustering and Moran's I scores, likely because it was not tailored to this specific dataset. Triku and COSG, despite being among the fastest methods, yield relatively weak spatial preservation and clustering performance. Further, GeneBasis, NS-Forest, and scPNMF show competitive performance in some settings, particularly for clustering at finer annotation levels, but are generally less effective than ReconST and Spapros in jointly balancing transcriptome reconstruction, spatial preservation, and multi-resolution cell-type discrimination. These results demonstrate that ReconST provides the most balanced and consistently strong performance across all evaluated criteria.

Next, we compared the gene panels selected by different methods. Examination of the gene overlap patterns in Fig. 2c reveals key insights into method relationships. The methods are ordered by panel overlap with ReconST in descending order. As expected, ReconST shows the highest overlap with Spapros (138 genes), confirming their convergence on a core set of informative genes that explains their superior performance. ReconST also shares substantial overlap with DE genes (94), suggesting it effectively captures differentially expressed markers while extending beyond simple differential expression. Consistently, ReconST exhibits strong overlap with other high-performing methods such as scPNMF (110), NS-Forest (123), and GeneBasis (118), further supporting that these approaches identify a largely shared set of biologically meaningful genes. In contrast, the poorly performing Triku and COSG show high mutual overlap (109 genes) but limited intersection with ReconST (11 and 10 genes, respectively).

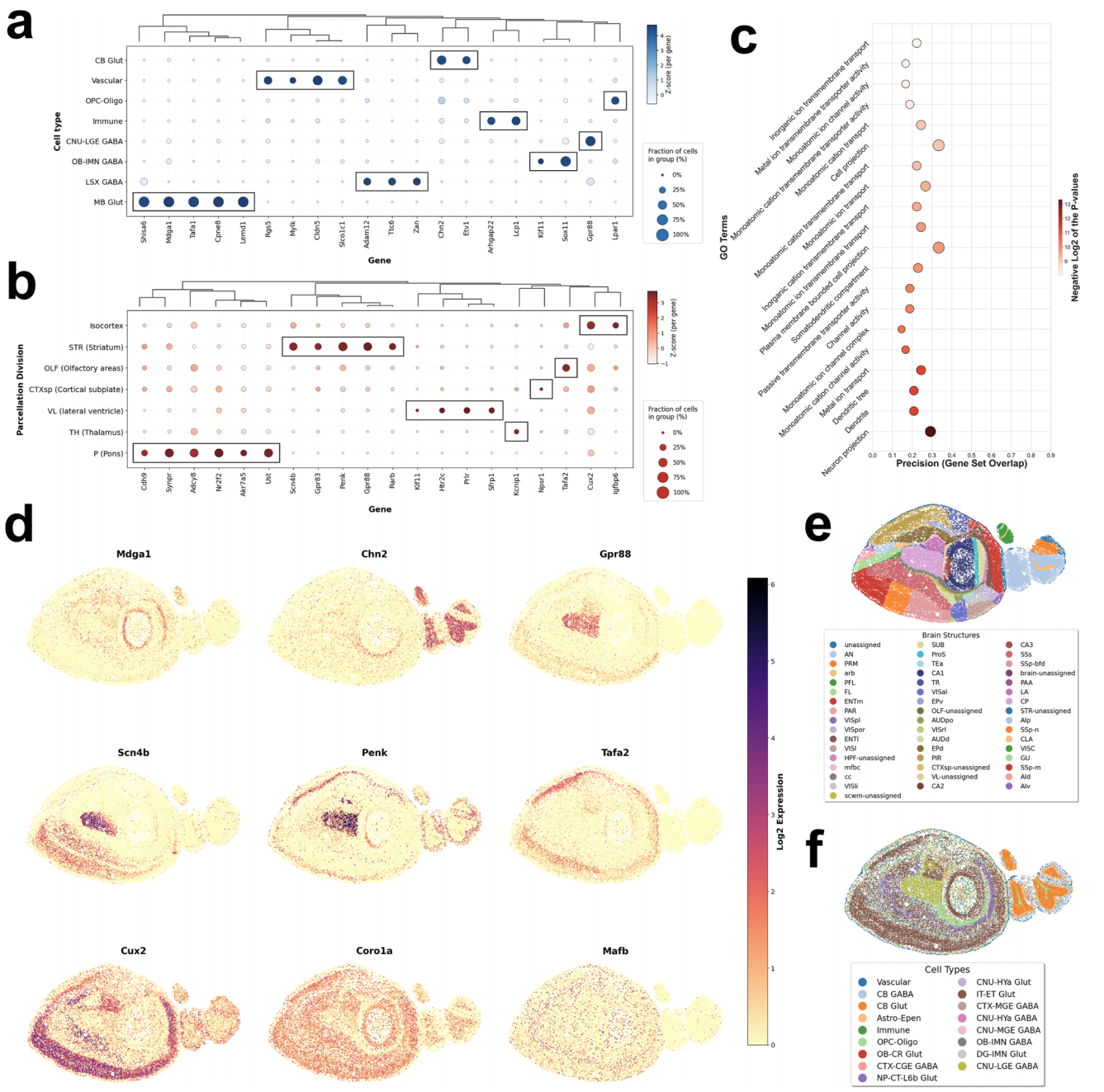


Figure 3. Analysis of the gene panel selected by ReconST on mouse brain MERFISH data. (**a**) Cell-type specificity of the genes from the ReconST panel, with black boxes highlighting blocks of high specificity. (**b**) Anatomical specificity of the genes against parcellation divisions. (**c**) Gene Ontology enrichment analysis of the gene panel uniquely selected by ReconST. (**d**) Spatial patterns of representative genes. (**e**) Spatial patterns of different cell types. (**f**) Spatial patterns of annotated anatomical parcellation divisions.

PERSIST, despite using an autoencoder approach like ReconST, shows only moderate overlap (108 genes), highlighting how ReconST's gated architecture identifies a more optimal gene set. The Vizgen default panel's moderate overlap with ReconST (43 genes) without capturing the key genes underscores the advantage of ReconST's data-driven selection over pre-designed panels. Overall, ReconST's strong convergence with the other top-performing methods and selective overlap with DE genes demonstrates its ability to identify some biologically informative genes.

We observe that the genes selected by ReconST display rich, distributed patterns across both cell types and anatomical divisions in the mouse brain MERFISH data. In Fig. 3a and b, the top 20 panel genes ranked by their per-class z-score are shown, with dot color encoding the per-gene z-score and dot size encoding the fraction of cells expressing the gene; black boxes mark contiguous blocks of high specificity. In Fig. 3a, sharp, near-exclusive enrichment in single cell-type classes is evident: Mdga1 marks MB Glut, Chn2 marks CB Glut, and Gpr88 marks CNU-LGE GABA; Coro1a and Mafb similarly show cell-type-restricted expression. In Fig. 3b, anatomically coherent patterns emerge across parcellation divisions: Scn4b and Penk in the striatum, Tafa2 in the olfactory areas, and Cux2 in the isocortex. Further, the top 50 genes selected

by ReconST and by other competing methods are shown in Supplementary Figs S3–S13. Together, these patterns indicate that ReconST captures both transcriptomic and spatial structure across the mouse brain.

To further investigate the difference between ReconST and other methods, we compared more closely the lists of genes selected by ReconST and Spapros, which our quantitative analysis shows performs second best after ReconST. To help visualize the co-expression relationship between different genes, we calculated pairwise correlations among high-variance genes based on their spatial expression patterns, then applied UMAP dimensionality reduction to project genes into two dimensions where proximity reflects expression similarity (Fig. 2d). The distribution of genes is not random but forms several compact clusters, indicating that genes within each cluster have similar expression profiles and therefore contain redundant information. To quantify how well the selected panel represents the full gene panel in this embedding, we computed two similarity metrics: cluster entropy [52] and mean nearest neighbor distance [53]. Cluster entropy is defined as the normalized Shannon entropy of the distribution of selected genes across gene clusters, with larger values indicating broader and more balanced coverage. Mean nearest neighbor distance is defined as the average Euclidean distance from each gene in the full panel to its nearest selected gene, with smaller values indicating better representation of the full panel. These metrics are consistent with the visual pattern in Fig. 2d and further show that ReconST achieves better coverage and representation than Spapros. Specifically, ReconST covers all major clusters, whereas the Spapros panel does not cover cluster 5 (colored red in Fig. 2d), which contains 42 genes in the full MERFISH dataset. By checking these genes in cluster 5, we find that this cluster maps to a vessel-border immune and stromal program: pan-leukocyte PTPRC/CD45, DC MIIC regulators TMEM176A/B, microglia-homeostatic SALL1, plus perivascular-fibroblast ECM markers such as COL15A1. This pattern fits border-associated macrophage/microglia engaging a perivascular fibroblast niche. Traditional panels may miss it because these cells are rare and spatially confined, and the genes are highly co-correlated, which redundancy-penalizing objectives may down-rank.

The exclusive reach into cluster 5 motivates gene-level inspection of ReconST-unique selections. ReconST uniquely selects the two genes, Coro1a and Mafb, that reached cluster 5. Coro1a and Mafb together indicate a border myeloid program at central nervous system (CNS) interfaces rather than a neuronal module. In Fig. 3d, Coro1a shows broad signal that aligns with cortex and major fiber tracts; biologically, Coro1a is a leukocyte actin regulator abundant in microglia, where it limits alpha-synuclein-induced lysosomal stress and inflammasome activation, and it has been validated as a microglial immunohistochemical marker. Mafb appears as sparse, distributed puncta; Mafb is a macrophage-lineage transcription factor that maintains adult microglial identity and homeostasis. Interpreting these expression patterns with the cell-type and region overlays in Fig. 3e and f, the most parsimonious view is that both genes track microglia and border-associated macrophages in meningeal and perivascular niches, key neuroimmune hubs at brain borders. This border module can be under-prioritized by variance- or redundancy-driven panel designs because these populations are small, spatially confined, and their markers are highly correlated. The spatial expression maps of Coro1a and Mafb are shown in Fig. 3d, and they display clear spatial patterns compared with the references in Fig. 3e and f. Together, these two ReconST-unique genes highlight a spatially restricted neuroimmune component at CNS borders, supporting the spatial and biological value captured by ReconST's selected gene set.

To further investigate the biological functions of ReconST-uniquely selected genes, we performed Gene Ontology (GO) enrichment analysis on them, using g:Profiler [54] via the gprofiler-official Python package, with the results shown in Fig. 3c. We see that ReconST-unique genes are strongly enriched for neuronal compartments, with the most significant terms spanning dendrite/dendritic tree, neuron projection, and cell projection. Membrane and signaling categories are also prominent, such as plasma membrane region, channel activity, passive transmembrane transporter activity, and monovalent ion transport, which are consistent with genes that shape neuronal excitability. Synaptic terms appear as well: postsynaptic membrane, neurotransmitter receptor complex, and ionotropic glutamate receptor complex, with some of these showing higher precision and larger intersections, indicating broad coverage of receptor components. To control for baseline enrichment, we compared the GO results of ReconST with a randomly selected gene panel in the Supplementary Fig. S2, showing that ReconST preferentially captures functionally specific neuronal and electrophysiological pathways rather than broad, non-specific terms.

In this analysis, we restrict ourselves to the curated MERFISH gene panel because evaluation metrics, Moran's I Index, require spatially resolved expression, which is only available for this targeted gene set. To assess robustness beyond this constraint, we repeated the analysis using genome-wide scRNA-seq input with minimal filtering, where ReconST maintains its performance advantage, as shown in the Supplementary Fig. S1. To further assess if there are any biases for or against certain cell type classes, we also compute the metric separately within each cell type class in the Supplementary Tables S1 and S2, where the results also show ReconST performs the best overall under the per-cell-type metrics and does not bias toward certain cell types. To assess whether our results depend on the choice of training data, we extend the analysis from the original single-region scRNA-seq training subset (31 299 cells) to a stratified multi-region training set by sampling cells across 24 brain regions of over 4 million cells, providing a more representative and large-scale input. ReconST yields consistent performance and stable gene selection under this setting, indicating robustness to data selection, and the details are shown in the Supplementary Table S3.

Performance evaluation and benchmark in a fetal lung study

To further assess the generalizability of ReconST in a developmental tissue context, we conducted an additional benchmark using a fetal lung spatial transcriptomics dataset [35]. In this study, lung tissue cores were profiled with Xenium using a custom 343-gene default panel, as well as scRNA-seq data with full genes. After preprocessing for quality control, 1 630 319 segmented cells and 299 018 086 transcripts were retained for downstream analysis. Details of preprocessing are provided in the "Materials and methods" section. The large sample size, targeted gene panel design, and cell-level spatial resolution make this dataset well suited for evaluating gene panel design strategies in a realistic experimental setting.

We compare ReconST against the ten benchmark methods as detailed in the "Materials and methods" section and one default Xenium gene panel used in this fetal lung dataset study itself [35]. From the scRNA-seq data, each method selects a gene panel with 343 genes to fairly compare with default gene panel. For this study, since the 343-gene default Xenium panel is already small and further selection will lead to a too-small panel, ReconST was evaluated mainly on scRNA-seq dataset, and the Xenium dataset is only used to illustrate some selected genes. The methods are evaluated with a comprehensive set of quantitative metrics as detailed in the "Materials and methods" section. We note that Moran's I index is no longer considered here since the evaluation is on scRNA-seq and thus there is no spatial information for the computation of Moran's I index.

The quantitative results are shown in Fig. 4a, with the methods ordered by their average rank across all performance metrics, excluding the running time. It demonstrates that ReconST is the top-performing method overall based on the average rank of the evaluation metrics. Further, it achieves the lowest reconstruction loss, together with the second-best clustering ARI, Pearson, and Spearman correlations, indicating superior recovery of global transcriptomic structure and distributional alignment. While several competing approaches, including GeneBasis, NS-Forest, and scPNMF, attain relatively strong correlation values, they fall short of ReconST either in reconstruction accuracy or distributional fidelity. Spapros remains a close competitor, showing solid performance across metrics, but a clear and consistent gap persists in favor of ReconST. The default panel achieves moderate correlation and reconstruction performance, outperforming weaker baselines such as HVG, but still lags behind the leading methods, particularly in distributional alignment, as reflected by its higher KL divergence. In contrast, methods such as HVG and PERSIST exhibit substantially weaker performance, especially in correlation-based measures, highlighting their limited ability to capture global transcriptomic structure under the same panel size constraint. This trend is further supported by the cell clustering concordance. To this end, we applied Leiden [43] method to identify cell clusters based on each selected gene panel, versus the annotated cell types from the dataset. ReconST's high ARI indicates the agreement with annotated cell types. While NS-Forest and Spapros also yield relatively strong ARI values, other approaches show more noticeable degradation in cell-type recovery.

The pairwise overlap analysis is shown in Fig. 4b, with methods ordered by panel overlap with ReconST in descending order. It further reveals that ReconST shares moderate overlap with several high-performing methods, including NS-Forest, Spapros, and Triku, while still maintaining a distinct subset of uniquely selected genes. In contrast, some methods either exhibit high redundancy with others or capture more limited portions of the gene space. Collectively, these results indicate that ReconST provides the most balanced performance across reconstruction accuracy, distributional alignment, and preservation of biologically meaningful cell-type structure, with Spapros and NS-Forest representing the closest alternatives.

To further assess how well each panel represents the full gene-expression landscape, we overlaid the genes selected by ReconST and the default Xenium panel onto a UMAP of high-variance genes constructed from pairwise spatial expression correlations, as shown in Fig. 4c. In this fetal lung dataset, the gene structure is less distinctly separated than in the previous dataset, so the visual contrast is somewhat weaker. Even so, ReconST shows broader and more balanced coverage across the full UMAP, successfully covering all clusters, whereas the default panel completely misses cluster 10 and provides only minimal coverage of cluster 2 with a single gene. This pattern is supported by the summary metrics: ReconST achieves higher normalized cluster entropy, 0.940 versus 0.855, and a smaller mean nearest distance, 0.270 versus 0.492, indicating more even representation across clusters and better overall coverage of the full gene set. Together, these results show that ReconST provides a more representative panel than the default design in this developmental tissue benchmark.

To investigate the biological functions uniquely preserved by ReconST, we performed GO enrichment analysis on ReconST-unique genes from clusters that were poorly represented by the default panel, with the results shown in Fig. 4d and e. The clearest signal emerged in cluster 10, for which the default panel selected no genes. ReconST-unique genes in this cluster showed strong enrichment for cytoplasmic translation, translation initiation and elongation, ribosomal subunit formation, and related protein synthesis pathways, indicating that ReconST captures a coherent biosynthetic program absent from the default design. In cluster 2, where the default panel selected only a single gene, ReconST-unique genes were enriched mainly for G1/S transition regulation, cell-cycle checkpoint signaling, and DNA integrity control, consistent with a proliferative cell-state program coupled with genome maintenance. Together, these findings show that the broader coverage achieved by ReconST reflects a recovery of biologically meaningful functional programs.

Next, we examined Cluster 2 and 10 more closely because they are only well-represented in the gene panel selected by ReconST. Cluster 10 mainly corresponds to a protein production program. These genes are involved in helping cells build proteins, which is one of the most fundamental activities needed for growth and function. The GO terms consistently point to translation, including translation initiation, elongation, and ribosome-related processes, and the gene set strongly matches that biology. **EIF3E** is part of the machinery that starts protein synthesis [55], **YBX1** is an RNA-binding protein with important post-transcriptional and translational regulatory functions [56]. Further, **COX6C** is linked to mitochondrial energy generation [57], which provides the fuel needed to sustain active biosynthesis, and it shows a clear spatial pattern as shown in Fig. 4g in comparison with the image of original tissue in Fig. 4f. Thus, this cluster appears to represent cells with strong biosynthetic activity, combining high translational demand with metabolic support.

The GO enrichment results suggest that Cluster 2 mainly represents a cell division and quality-control program. These genes are associated with cells preparing to copy their DNA and move through the cell cycle, while also checking that the DNA is intact before division continues. This is supported by the key genes in the cluster. The **CCND1** gene is known to be responsible for pushing cells from the G1 phase into DNA replication [58], and it shows a strong spatial pattern as shown in Fig. 4h in comparison with the image of original tissue in Fig. 4f. Further, **NUCKS1** is involved in regulating cell-cycle progression [59], and **PRKDC** plays an important role in repairing DNA damage and maintaining genome stability [60]. Together, these findings suggest that ReconST successfully captured a spatially organized population of genes

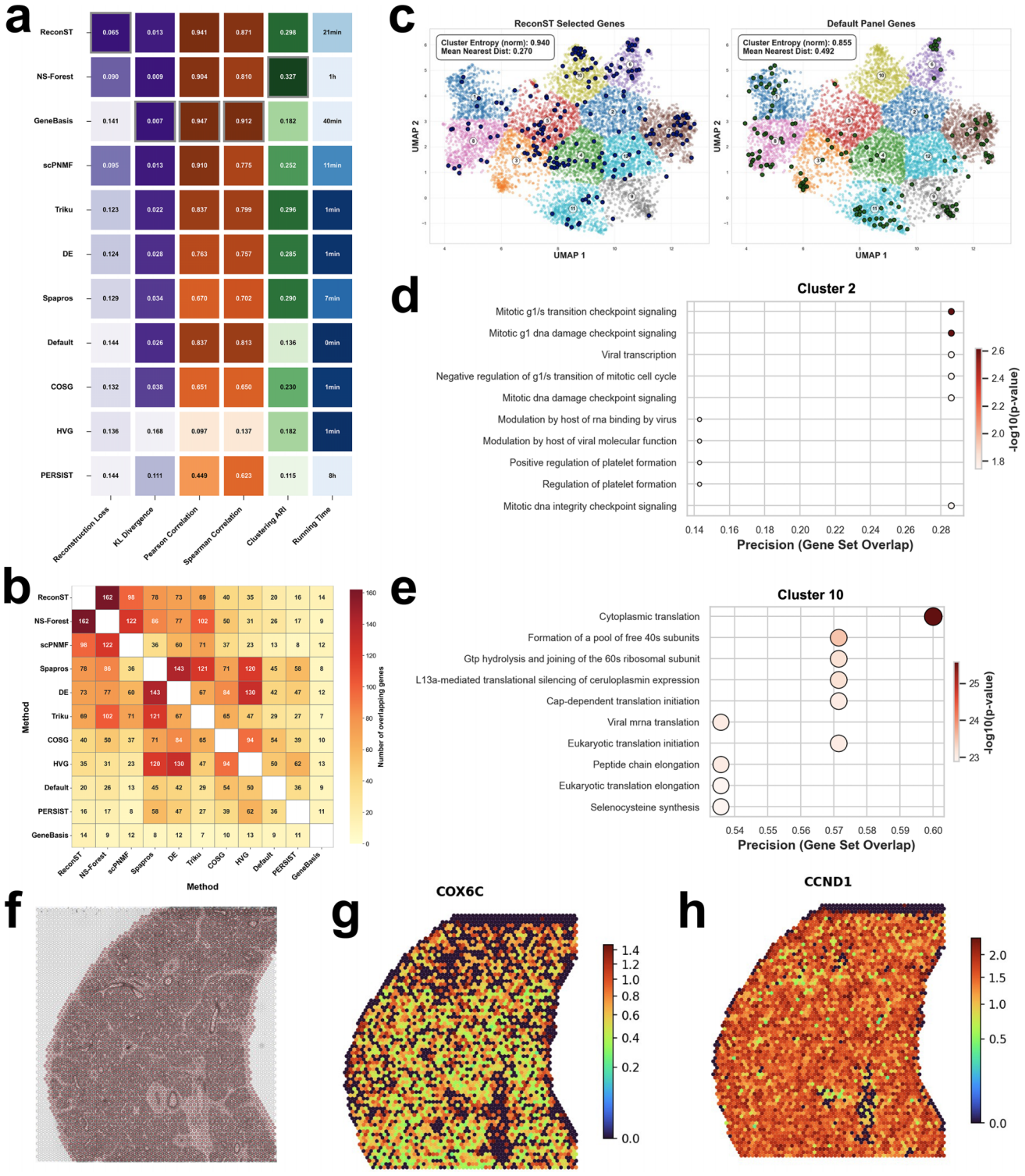


Figure 4. Comparison and analysis of the gene panel designed by different methods on the fetal lung study. (**a**) Benchmark analysis of different gene panel selection methods using the fetal lung dataset. In each column, the best performer is outlined. (**b**) Gene panel overlap of the comparing methods. (**c**) UMAP showing expression patterns of selected genes by ReconST and the default panel, respectively. (**d**, **e**) GO enrichment analysis of the ReconST unique genes in clusters 2 and 10, respectively. (**f**) Image of original fetal lung tissue sample from [35]. (**g**, **h**) Spatial patterns of representative genes.

related to cell proliferation and cell-cycle control in the fetal lung.

Discussion

We have developed ReconST, a novel method for unbiased gene panel design in spatial transcriptomics analysis. ReconST uses a gated autoencoder model to select the subset of genes that optimally reconstructs whole transcriptomic information. The gated autoencoder clearly distinguishes selected from unselected genes by assigning nonzero weights to the selected genes. If the number of selected genes is predetermined, this step can be customized by further thresholding the nonzero weights. Our benchmark analysis shows that ReconST achieves higher reconstruction accuracy, stronger spatial autocorrelation, and clearer cluster resolution compared to existing methods. Further, we established a systematic benchmark that connects training on scRNA-seq data with validation on high-resolution spatial transcriptomics data (MERFISH) to evaluate how gene panel selection methods perform in realistic spatial contexts. Unlike evaluations confined to the single-cell domain, our framework measures recovery of transcriptomic content and spatial organization in actual ST data. The set of metrics (expression reconstruction, spatial autocorrelation, clustering concordance, and runtime) enable fair comparison across baselines and reveal tradeoffs between transcriptomic fidelity and spatial structure. This framework demonstrates the robustness of ReconST and provides a reusable pipeline for future method development and comparison. We have made ReconST available as a Python package. The entire pipeline can be conveniently executed with one function, and it also includes a tutorial example for detailed training and selection steps.

ReconST is practical for experimental design: starting from a single-cell dataset, it outputs a compact panel ready for targeted spatial transcriptomics, reducing sequencing cost while keeping downstream analyses such as trajectory inference and differential expression reliable. By leveraging the representation learning power of deep neural networks, ReconST not only selects genes but also learns how to encode complex transcriptomic structures. Through end-to-end training, it simultaneously optimizes both feature extraction and gene selection, and the integration of data-driven gated selection with deep representation learning makes it uniquely capable of identifying the most informative and efficient gene subset for reconstruction while maintaining biological relevance. Enrichment analysis confirms that the chosen genes align with pathways that matter for the profiled tissue, and spatial maps of representative genes display coherent patterns that span several anatomical divisions. Despite these constraints, ReconST offers an efficient and flexible route to information-rich panel design in spatial transcriptomics.

ReconST's unique selections highlight biologically consequential immune programs at brain borders. Notably, Coro1a and Mafb are specifically captured and together delineate a border-associated myeloid module that tracks microglia and border-associated macrophages in meningeal and perivascular niches, neuroimmune hubs that are often underrepresented in variance- or redundancy-driven designs. Coro1a reflects leukocyte actin regulation with a broad spatial signal, while Mafb marks macrophage lineage identity and microglial homeostasis, enabling a sensitive readout of surveillance, cytokine exchange, and microglia-associated remodeling processes. Alongside these immune markers, ReconST retains neuronal and membrane signaling genes with coherent laminar and regional biases, supporting integrated analyses of circuit level gradients and immune stromal interactions *in situ*. By optimizing for global information reconstruction rather than a fixed catalog of markers, the panel improves detection of rare but functionally important immune states while preserving neuronal organization, thereby facilitating hypothesis generation in development, plasticity, and neurodegeneration.

The fetal lung benchmark further supports the biological relevance of ReconST in a developmental context. The genes uniquely selected by ReconST capture coherent functional programs, including protein synthesis and cell-cycle regulation, and provide broader coverage of the gene-expression landscape. These signals are spatially organized and biologically interpretable, reflecting key processes in tissue growth and differentiation. Notably, such programs may be underrepresented in panels designed primarily for marker-based selection. This result highlights that optimizing for global transcriptomic reconstruction enables the recovery of functionally important biological signals beyond classical markers.

Acknowledgements

Author contributions: Haoran Lu (Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Luyang Fang (Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Orlando Zeng (Software [supporting]), Wenxuan Zhong (Funding acquisition [equal], Methodology [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Guo-Cheng Yuan (Funding acquisition [equal], Methodology [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Ping Ma (Funding acquisition [equal], Methodology [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This research was partially supported by the U.S. National Science Foundation under grants NSF DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809; the U.S. National Institutes of Health under grant R01GM152814 and RF1MH133703. Funding to pay the Open Access publication charges for this article was provided by U.S. National Science Foundation (DMS-2319279) and U.S. National Institutes of Health (RF1MH133703).

Data availability

The software of the ReconST method is developed in a python package submitted as Supplementary file and is publicly available at https://haoranlustat.github.io/ReconST/ and https://doi.org/10.5281/zenodo.20434940 with a tutorial.

References

1. Zhang M, Eichhorn SW, Zingg B *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* 2021;598:137–43. https://doi.org/10.1038/s41586-021-03705-x
2. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;19:534–46. https://doi.org/10.1038/s41592-022-01409-2
3. Du J, Yang Y-C, An Z-J *et al.* Advances in spatial transcriptomics and related data analysis strategies. *J Transl Med* 2023;21:330. https://doi.org/10.1186/s12967-023-04150-2
4. Rao A, Barkley D, França GS *et al.* Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;596:211–20. https://doi.org/10.1038/s41586-021-03634-9
5. Marx V. Method of the year: spatially resolved transcriptomics. *Nat Methods* 2021;18:9–14. https://doi.org/10.1038/s41592-020-01033-y
6. Janesick A, Shelansky R, Gottscho A *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* 2023;14:8353. https://doi.org/10.1038/s41467-023-43458-x
7. Lu Y, Liu M, Yang J *et al.* Spatial transcriptome profiling by MERFISH reveals fetal liver hematopoietic stem cell niche architecture. *Cell Discov* 2021;7:47. https://doi.org/10.1038/s41421-021-00266-1
8. Zhu Q, Shah S, Dries R *et al.* Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data. *Nat Biotechnol* 2018;36:1183–90. https://doi.org/10.1038/nbt.4260
9. Wang Y, Liu B, Zhao G *et al.* Spatial transcriptomics: technologies, applications and experimental considerations. *Genomics* 2023;115:110671. https://doi.org/10.1016/j.ygeno.2023.110671
10. Rodriques SG, Stickels RR, Goeva A *et al.* Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;363:1463–7. https://doi.org/10.1126/science.aaw1219
11. Ståhl PL, Salmén F, Vickovic S *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;353:78–82. https://doi.org/10.1126/science.aaf2403
12. 10X Genomics. *Visium Spatial Platform*. https://www.10xgenomics.com/platforms/visium (1 April 2026, date last accessed).
13. Merritt CR, Ong GT, Church SE *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol* 2020;38:586–99. https://doi.org/10.1038/s41587-020-0472-9
14. Liu Y, Yang M, Deng Y *et al.* High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* 2020;183:1665–81. https://doi.org/10.1016/j.cell.2020.10.026
15. Kuemmerle LB, Luecken MD, Firsova AB *et al.* Probe set selection for targeted spatial transcriptomics. *Nat Methods* 2024;21:2260–70. https://doi.org/10.1038/s41592-024-02496-z
16. Chen KH, Boettiger AN, Moffitt JR *et al.* Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;348:aaa6090. https://doi.org/10.1126/science.aaa6090
17. Lubeck E, Coskun AF, Zhiyentayev T *et al.* Single-cell *in situ* RNA profiling by sequential hybridization. *Nat Methods* 2014;11:360–1. https://doi.org/10.1038/nmeth.2892
18. Eng C-HL, Lawson M, Zhu Q *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 2019;568:235–9. https://doi.org/10.1038/s41586-019-1049-y
19. 10X Genomics. *Xenium In Situ Platform*. https://www.10xgenomics.com/platforms/xenium (1 April 2026, date last accessed).
20. Wang X, Allen WE, Wright MA *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;361:eaat5691. https://doi.org/10.1126/science.aat5691
21. Lim HJ, Wang Y, Buzdin A *et al.* A practical guide for choosing an optimal spatial transcriptomics technology from seven major commercially available options. *BMC Genomics* 2025;26:47. https://doi.org/10.1186/s12864-025-11235-3
22. Dries R, Chen J, Del Rossi N *et al.* Advances in spatial transcriptomic data analysis. *Genome Res.* 2021;31:1706–18. https://doi.org/10.1101/gr.275224.121
23. Song D, Li K, Hemminger Z *et al.* scPNMF: sparse gene encoding of single cells to facilitate gene selection for targeted gene profiling. *Bioinformatics* 2021;37:i358–66. https://doi.org/10.1093/bioinformatics/btab273
24. Aevermann B, Zhang Y, Novotny M *et al.* A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Res* 2021;31:1767–80. https://doi.org/10.1101/gr.275569.121
25. Missarova A, Jain J, Butler A *et al.* geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq. *Genome Biol* 2021;22:1–22. https://doi.org/10.1186/s13059-021-02548-z
26. Yao Z, van Velthoven CT, Kunst M *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* 2023;624:317–32. https://doi.org/10.1038/s41586-023-06812-z
27. Robbins H, Monro S. A stochastic approximation method. *Ann Math Statist* 1951;22:400–7. https://doi.org/10.1214/aoms/1177729586
28. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
29. Scardapane S, Comminiello D, Hussain A *et al.* Group sparse regularization for deep neural networks. *Neurocomputing* 2017;241:81–9. https://doi.org/10.1016/j.neucom.2017.02.029
30. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* 1995;14:1137–45.
31. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313:504–7. https://doi.org/10.1126/science.1127647
32. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44. https://doi.org/10.1038/nature14539
33. Ng AY. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In: *Proceedings of the twenty-first international conference on Machine learning*. Banff, Alberta, Canada, 2004, 78.
34. Yang M, Lim MK, Qu Y *et al.* Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction. *Expert Syst Appl* 2023;213:118873. https://doi.org/10.1016/j.eswa.2022.118873
35. Quach H, Farrell S, Wu MJM *et al.* Early human fetal lung atlas reveals the temporal dynamics of epithelial cell plasticity. *Nat Commun* 2024;15:5898. https://doi.org/10.1038/s41467-024-50281-5
36. Vizgen. Vizgen Gene Panels. 2025. https://vizgen.com/gene-panels/ (3 April 2025, date last accessed).
37. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics bulletin* 1945;1:80–3. https://doi.org/10.2307/3001968
38. Covert I, Gala R, Wang T *et al.* Predictive and robust gene selection for spatial transcriptomics. *Nat Commun* 2023;14:2091. https://doi.org/10.1038/s41467-023-37392-1
39. M Ascensión A, Ibáñez-Solé O, Inza I *et al.* Triku: a feature selection method based on nearest neighbors for single-cell data. *Gigascience* 2022;11:giac017. https://doi.org/10.1093/gigascience/giac017

40. Dai M, Pei X, Wang X-J. Accurate and fast cell marker gene identification with COSG. *Brief Bioinform* 2022;23:bbab579. https://doi.org/10.1093/bib/bbab579
41. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Statist* 1951;22:79–86. https://doi.org/10.1214/aoms/1177729694
42. Hubert L, Arabie P. Comparing partitions. *J Classification* 1985;2:193–218. https://doi.org/10.1007/BF01908075
43. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;9:1–12. https://doi.org/10.1038/s41598-019-41695-z
44. Moran PA. Notes on continuous stochastic phenomena. *Biometrika* 1950;37:17–23. https://doi.org/10.1093/biomet/37.1-2.17
45. Xu B, Wang N, Chen T *et al.* Empirical evaluation of rectified activations in convolutional network. In: *Deep learning workshop at International conference on machine learning*. Lille, France, 2015.
46. Srivastava N, Hinton G, Krizhevsky A *et al.* Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 2014;15:1929–58.
47. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *International conference on learning representations*. San Diego, California, United States, 2015.
48. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15. https://doi.org/10.1186/s13059-017-1382-0
49. McInnes L, Healy J, Melville J *et al.* Umap: uniform manifold approximation and projection. *JOSS* 2018;3:861. https://doi.org/10.21105/joss.00861
50. Lee JA, Verleysen M. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing* 2009;72:1431–43. https://doi.org/10.1016/j.neucom.2008.12.017
51. Venna J, Kaski S. Neighborhood preservation in nonlinear projection methods: an experimental study. In: *International conference on artificial neural networks*. Vienna, Austria, 2001, 485–91.
52. Shannon CE. A mathematical theory of communication. *The Bell system technical journal* 1948;27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
53. Clark PJ, Evans FC. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 1954;35:445–53. https://doi.org/10.2307/1931034
54. Raudvere U, Kolberg L, Kuzmin I *et al.* g: profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;47:W191–8. https://doi.org/10.1093/nar/gkz369
55. Brito Querido J, Sokabe M, Díaz-López I *et al.* The structure of a human translation initiation complex reveals two independent roles for the helicase eIF4A. *Nat Struct Mol Biol* 2024;31:455–64. https://doi.org/10.1038/s41594-023-01196-0
56. Kwon E, Todorova K, Wang J *et al.* The RNA-binding protein YBX1 regulates epidermal progenitors at a posttranscriptional level. *Nat Commun* 2018;9:1734. https://doi.org/10.1038/s41467-018-04092-0
57. Wang C, Lv J, Xue C *et al.* Novel role of COX6c in the regulation of oxidative phosphorylation and diseases. *Cell Death Discov* 2022;8:336. https://doi.org/10.1038/s41420-022-01130-1
58. Musgrove EA, Caldon CE, Barraclough J *et al.* Cyclin D as a therapeutic target in cancer. *Nat Rev Cancer* 2011;11:558–72. https://doi.org/10.1038/nrc3090
59. Hume S, Grou CP, Lascaux P *et al.* The NUCKS1-SKP2-p21/p27 axis controls S phase entry. *Nat Commun* 2021;12:6959. https://doi.org/10.1038/s41467-021-27124-8
60. Blackford AN, Jackson SP. ATM, ATR, and DNA-PK: the trinity at the heart of the DNA damage response. *Mol Cell* 2017;66:801–17. https://doi.org/10.1016/j.molcel.2017.05.015

Received: November 6, 2025. **Revised:** April 28, 2026. **Accepted:** May 25, 2026