scientific reports

Check for updates

OPEN

CiCLoDS: Joint cell clustering and gene selection for single-cell spatial transcriptomics

Nuozhou Wang[1,3], Yimeng He[2,3], Edward Ray[2], Kyoung Jae Won[2], Curtis L Cetrulo[2], Debiao Li[2], Jason H Moore[2], Shuzhong Zhang[1]✉ & Xiuzhen Huang[2]✉

Imaging-based spatial transcriptomic (ST) assays now profile targeted gene panels for hundreds of thousands of cells while preserving tissue context, yet most downstream analyses still need a compact, interpretable subset. We introduce CiCLoDS (Clustering in Critical & Low-Dimensional Subspace), an unsupervised framework adapted for spatial transcriptomics to jointly optimize clustering and feature selection under a strict, user-defined gene budget. The method returns an explicit feature subset aligned with the discovered partitions, accepts spatial side information via sine–cosine positional encodings, and converges in minutes on commodity CPUs. We benchmark CiCLoDS on three diverse datasets: Vizgen MERFISH mouse liver (1.27 M cells), 10x Xenium human colon (23 K cells), and Human DLPFC. On hepatocyte zonation, CiCLoDS raises the adjusted Rand index by up to +0.36 over PCA and geneBasis, preserves neighborhood structure with a kNN-overlap AUC of 0.89, and detects vessel-associated voids at 92.8% F1. On Human DLPFC (Visium), CiCLoDS demonstrates robust generalization, achieving a mean Adjusted Rand Index (ARI) of 0.40—surpassing BayesCafe (0.32)—and maintaining high accuracy on heterogeneous samples where baselines fail. Furthermore, we reveal a synergistic utility: using CiCLoDS to initialize BayesSpace raises the mean ARI to 0.50, resolving local minima issues to outperform either method alone. These results show that a lightweight, jointly optimized objective can achieve competitive accuracy, robust generalization, and synergistic utility while producing assay-ready feature panels.

Spatial transcriptomics (ST) has revolutionized tissue analysis by mapping gene expression with high sensitivity while preserving native architecture[1,2]. Whether using sequencing-based platforms like 10x Genomics Visium[3] or imaging-based methods like MERFISH and Xenium that resolve transcripts at subcellular granularity[4,5], the fundamental analytical challenge remains the same: extracting meaningful biological structure from high-dimensional, noisy data. Specifically, downstream analysis typically requires identifying distinct cellular neighborhoods (clustering) and the minimal set of genes that define them (feature selection).

Current computational approaches to this problem generally fall into three methodological families, distinguished by how they couple—or decouple—these two tasks.

Sequential Dimensionality Reduction and Selection. The most common workflow treats feature selection and clustering as separate, sequential steps. Standard pipelines often employ Principal Component Analysis (PCA) to reduce dimensionality before applying K-means or Leiden clustering. While computationally efficient, PCA optimizes for global variance preservation, meaning the top principal components may capture technical noise or cell-cycle effects rather than the biological distinctions necessary for clustering. Dedicated selection methods like geneBasis[6] and SCMER[7] improve on this by selecting gene panels that preserve the local data manifold. However, because the selection step remains decoupled from the clustering objective, the chosen features are not guaranteed to be optimal for the subsequent partition of cellular domains.

Deep Embedded Clustering. To overcome the limitations of sequential steps, deep learning approaches such as scDeepCluster[8] and scDCC[9] learn nonlinear embeddings and cluster assignments jointly. By incorporating zero-inflated negative binomial (ZINB) losses or pairwise constraints, these models improve robustness on sparse count data. Graph neural networks like scGNN[10] further extend this by integrating cell–cell connectivity. While effective at capturing complex manifolds, these models typically function as "black boxes": they do not return an explicit, user-controlled gene panel, and feature attribution is often post hoc (e.g., via saliency maps), limiting their utility for panel design and biological interpretability.

[1]University of Minnesota, Minneapolis, Minnesota 55455, USA. [2]Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. [3]Nuozhou Wang and Yimeng He contributed equally to this work. ✉email: zhangs@umn.edu; xiuzhen.huang@cshs.org

Joint Optimization and Subspace Clustering. A third family of methods—to which our approach belongs—seeks to mathematically unify feature selection and clustering within a transparent, linear framework. This lineage includes Sparse K-means[11], which introduces a sparsity penalty to weight features during clustering, and Sparse Subspace Clustering (SSC)[12,13], which posits that data points lie in a union of low-dimensional subspaces. Theoretically, these methods offer the rigor of joint optimization without the opacity of deep learning. However, classical implementations like SSC are computationally prohibitive for large cohorts ($N > 10^5$ cells) and were not designed to incorporate the spatial side information inherent to ST data[14]. Consequently, this rigorous mathematical family has seen limited adoption in the ST field compared to heuristic or deep learning alternatives.

In this paper, we introduce **CiCLoDS** (Clustering in Critical & Low-Dimensional Subspace), a framework that adapts the principles of the subspace clustering family to the specific scale and constraints of spatial transcriptomics. CiCLoDS jointly optimizes clustering and feature selection under a single objective subject to a user-defined feature budget p. Unlike sequential "select-then-cluster" pipelines, CiCLoDS ensures that selected genes directly support the discovered partitions. Furthermore, the formulation naturally admits spatial side information via positional encodings without altering the core optimization. We demonstrate that CiCLoDS achieves higher accuracy than state-of-the-art baselines while identifying compact, spatially coherent gene panels, effectively bridging the gap between statistical rigor and practical scalability.

Methods

The general CiCLoDS framework

We first present CiCLoDS (Clustering in Critical & Low-Dimensional Subspace) as a general mathematical framework that optimizes clustering and feature selection simultaneously under a single objective. This formulation supports a defined feature budget (p), producing clusters and an interpretable feature set in a single pass.

Consider a tensor with genetic data $G = \{g_{i_1 i_2 \ldots i_l j}\}$. Our core premise is that high-dimensional biological data often contains many redundant or noisy features. We seek to select a small proportion of informative features and minimize the total loss within each resulting partition. Formally, we propose the following general optimization model:

$$\min f(\mathcal{K}, S, C) = \sum_{t=1}^{l} \sum_{K_t \in \mathcal{K}_t} \sum_{i_t \in K_t} \sum_{j \in S} \ell(g_{i_1 i_2 \ldots i_l j} - C_{K_1 K_2 \ldots K_l j}) \tag{1}$$

where $\mathcal{K}_t$ is a partition of the m_t class of the t-th block, S is a subset of features (genes) constrained to a budget of size n, $C_{K_1 K_2 \ldots K_l j}$ is a center point of the clusters $K_1, K_2, \ldots, K_l$ in feature j, and ℓ is a loss function.

Optimization model for single-cell spatial transcriptomics

While the general scheme (1) applies to multi-modal tensor data, it is particularly relevant for single-cell spatial transcriptomics (ST), which is typically represented as a cell-by-gene matrix. In this section, we tailor the general scheme to this setting. The objective is to identify the most relevant subset of genes S that best explain the spatial organization of cell clusters. A schematic overview of the CiCLoDS framework and its application to spatial transcriptomics is illustrated in Fig. 1.

We adapt the general model by assuming a 2D matrix structure and employing a squared Euclidean loss (equivalent to K-means). We select a small proportion of genes and minimize the total within-cluster sum of squares:

$$\min f(\mathcal{K}, S, C) = \sum_{K \in \mathcal{K}} \sum_{i \in K} \sum_{j \in S} (g_{ij} - C_{Kj})^2 \tag{2}$$

where $\mathcal{K}$ is an m-class partition of the cells, S is a subset of genes limited to size n, and C_{Kj} represents the centroid of cluster K for gene j.

We solve this using a block coordinate descent method, alternating the optimization of the objective function $\min_C f(\mathcal{K}, S, C)$ with respect to the partition $\mathcal{K}$ and the feature set S.

Once the optimal features and clusters are identified, we can apply standard K-means using only the selected top genes. Furthermore, we can refine these clusters by incorporating spatial neighborhood information or side information (via positional encodings) directly into the feature matrix before optimization.

Considerations for data preprocessing

Primary preprocessing workflow

For all analyses presented, we employed a standardized preprocessing pipeline. First, we applied library size normalization using `scanpy.pp.normalize_total` to scale the total counts per cell to a fixed target sum (10^4), ensuring comparability across cells with varying sequencing depths. We did not apply logarithmic transformation. Subsequently, we performed feature-wise scaling to standardize the variance: genes with zero variance were excluded, and the remaining gene vectors were divided by their sample standard deviation. This ensures that all genes contribute equally to the objective function's variance minimization term, preventing genes with naturally high magnitudes from dominating the feature selection process.

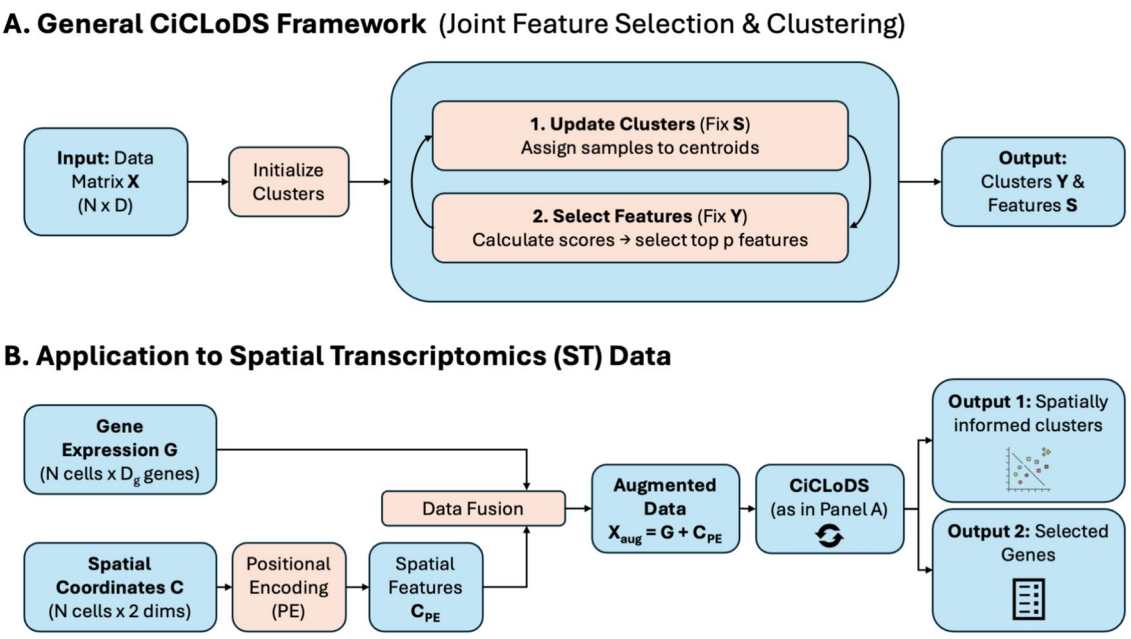

Fig. 1. Overview of the CiCLoDS Framework. (**A**) The iterative block coordinate descent algorithm alternates between cluster updates and feature selection. (**B**) For spatial transcriptomics, gene expression is fused with positional encodings before entering the CiCLoDS optimization loop.

Results

We evaluated CiCLoDS, our joint clustering and feature selection method, using simulated data and two real spatial transcriptomics (ST) datasets generated by advanced submicron-resolution platforms. Below, we introduce the real ST datasets employed in our analysis.

Vizgen mouse liver dataset

The data were acquired from the Vizgen Mouse Liver Map, and were assayed with in situ-based Vizgen MERSCOPE platform. The data comprise 347 targeted genes across four mouse liver samples (two biological replicates each from two specimens). The dataset includes 1.2 billion transcripts across 1.27 million cells, offering single-cell resolution spatial gene expression data along with cell boundary annotations and DAPI/PolyT imaging data for tissue context. We primarily use the `Liver1Slice1` subset in our study, focusing on the expression profiles contained in the `cell_by_gene.csv` file and the associated metadata in `cell_metadata.csv`.

Xenium human colon dataset

The dataset is derived from a human colon adenocarcinoma sample assayed using the in situ-based Xenium platform. This resource contains expression profiles of approximately 5,000 genes measured across 23,478 cells at single-cell resolution, along with cell segmentation boundaries determined by the Xenium Cell Segmentation solution. To maintain a consistent data structure with the mouse liver dataset, we processed expression matrices and spatial coordinates in a similar manner, thereby facilitating direct comparisons between the two platforms.

Identification of zonation-related hepatocyte subpopulations

CiCLoDS achieves competitive clustering performance

To comprehensively evaluate clustering performance for liver zonation, we benchmarked CiCLoDS against three competing methods: geneBasis[6], scDCC[9], and PCA. For methods that do not inherently perform clustering (PCA and geneBasis), we applied a standard k-means algorithm to generate cluster assignments. To ensure a fair comparison, we standardized key parameters across all applicable methods, setting the number of clusters to $k = 10$ and the feature dimensionality to $p = 64$. For PCA, this involved using the top 64 principal components, as using the top 64 most variable genes yielded suboptimal results. A detailed description of all experimental settings is available in Supplementary Section A.1.1.

The clustering performance of each method was quantified using three standard metrics—Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F1-score—on non-overlapping $50\,\mu\mathrm{m}$ spatial tiles. As shown in Table 1, the semi-supervised scDCC method achieved the highest scores across all metrics in both peri-portal and peri-central regions. CiCLoDS consistently ranked as the second-best method, significantly outperforming both PCA and the gene-selection baseline, geneBasis. Specifically, CiCLoDS demonstrated substantial performance gains over PCA, improving the ARI/NMI/F1 scores by +0.16/+0.10/+0.07 in the peri-portal region and by a remarkable +0.36/+0.18/+0.17 in the peri-central region.

Region	Method	Tile-level metrics		
		ARI↑	NMI↑	F1-score↑
Peri-portal	PCA	0.403	0.378	0.799
	CiCLoDS	0.559	0.475	0.872
	geneBasis	0.393	0.368	0.845
	scDCC	**0.618**	**0.569**	**0.889**
Peri-central	PCA	0.119	0.272	0.716
	CiCLoDS	0.477	0.448	0.884
	geneBasis	0.357	0.396	0.841
	scDCC	**0.679**	**0.598**	**0.938**

Table 1. Tile-level metrics on non-overlapping 50 μm tiles. **Bold** = best; underline = second best within each region/metric.

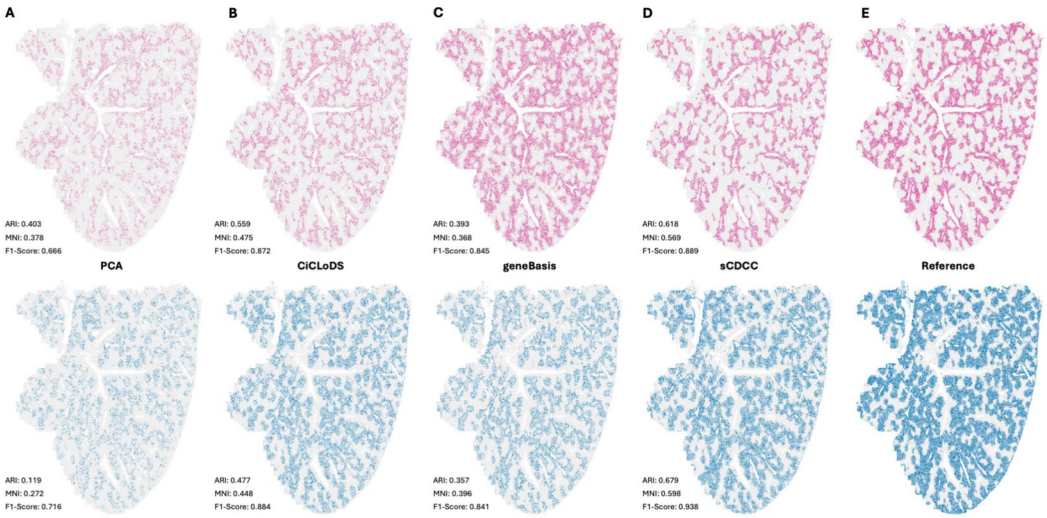

Fig. 2. Comparison of spatial zonation maps generated by different computational methods. The figure displays zonation patterns in peri-portal (top row) and peri-central (bottom row) regions of liver tissue. Each column represents a different method for identifying spatial zones, evaluated against a ground-truth reference map (far right). **(A-D) Method Comparison:** The zonation maps were generated using four different approaches: (A) PCA with k-means clustering (k=10) on the top 64 principal components; (B) CiCLoDS (k=10 clusters, p=64 features); (C) k-means clustering (k=10) on the 64 genes selected by geneBasis; and (D) scDCC, a semi-supervised method using pairwise constraints derived from Leiden clustering (n=10 clusters). **(E) Reference Map:** The reference zonation (far right) was established by classifying hepatocytes from gene expression and then clustering them based on the key peri-central biomarker *Axin2*. **Quantitative Evaluation:** Metrics overlaid on each panel represent the tile-level Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and macro-F1 score (left to right), quantifying the similarity of each method's output to the reference map. All metrics were computed using non-overlapping 50 μm spatial tiles.

Visual inspection of the zonation maps in Fig. 2 corroborates these quantitative clustering results. The cluster assignments produced by scDCC most closely replicate the spatial patterns of the reference map. CiCLoDS generates highly coherent zonal patterns with sharp boundaries and significantly fewer scattered or misassigned tiles compared to PCA and geneBasis, particularly in the peri-central region. In contrast, the maps from PCA and geneBasis appear more fragmented and fail to capture key structural features near blood vessels, consistent with their lower metric scores. Overall, these results indicate that CiCLoDS achieves a compelling balance between interpretability and clustering accuracy, providing clear spatial organization that is highly competitive with the end-to-end clustering baseline.

CiCLoDS identifies biologically valid hepatocyte populations

Having confirmed the strong clustering performance of CiCLoDS, we next sought to determine if it could identify more subtle, biologically meaningful hepatocyte populations. For this fine-grained analysis, we increased the feature space to $p = 128$ genes to maximize our ability to detect novel biological signals. We then benchmarked our clustering results against the established peri-portal and peri-central reference populations defined by Vizgen's *Axin2*-based methodology[15].

Our method demonstrates a remarkable alignment with this biological ground truth, as illustrated in Fig. 3. We identified two specific clusters—Cluster 7 ("Peri-Portal-like") and Cluster 6 ("Peri-Central-like")—that corresponded directly to the Peri-Portal and Peri-Central reference meta-clusters, respectively. The concordance was substantial: 85.4% of cells in the reference Peri-Portal meta-cluster were assigned to our Cluster 7, and 84.9% of cells in the reference Peri-Central meta-cluster were assigned to our Cluster 6.

Crucially, CiCLoDS achieved this alignment without using any prior knowledge of the key biomarker, *Axin2*, or spatial coordinates. This suggests that the expanded gene set selected by CiCLoDS contains novel combinations of genes that define hepatocyte spatial identity, demonstrating its robust ability to uncover biologically meaningful patterns without relying on predefined biomarkers.

(For all analyses, we set $k = 10$. Detailed experimental setups are described in Supplementary Section A.1)

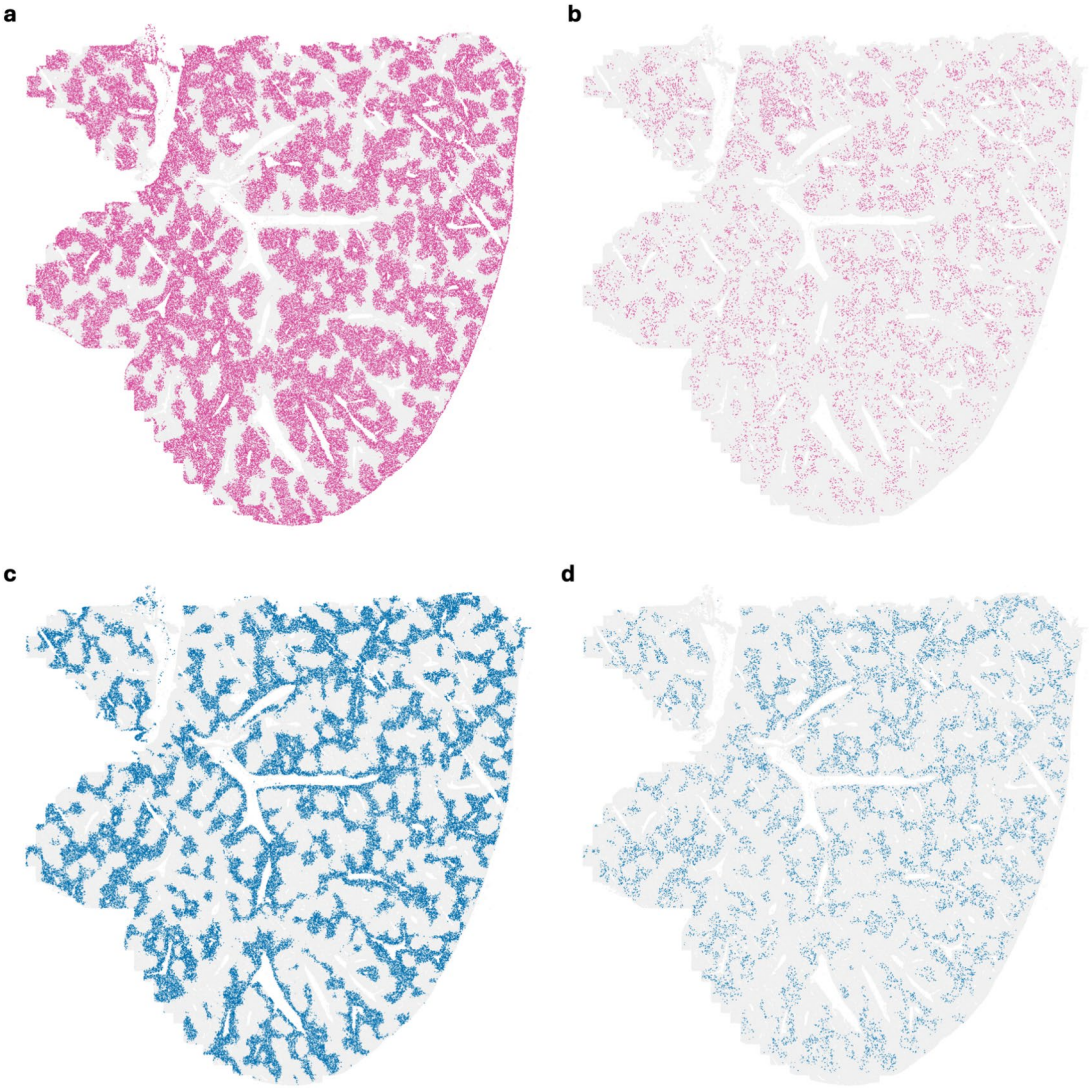

Fig. 3. Visual comparison of clustering results with biological references for the mouse liver dataset. **(a)** Reference Peri-Central Meta-Cluster identified using *Axin2* as a biomarker, following Vizgen's methodology. **(b)** Cluster 6 generated by our method ($k = 10$, $p = 128$), showing alignment with the Peri-Central Meta-Cluster. Our method, which jointly performs clustering and feature selection, was executed ten times with different initialization values to ensure stability and reproducibility. For each cell, the final cluster assignment was determined by the mode of its assignments across all iterations. **(c)** Reference Peri-Portal Meta-Cluster identified using *Axin2* as a biomarker. **(d)** Cluster 7 generated by our method ($k = 10$, $p = 128$), demonstrating alignment with the Peri-Portal Meta-Cluster. Notably, this alignment was achieved without incorporating *Axin2* as prior knowledge or including it in the selected gene set. Specifically, 85.4% of cells in the Peri-Portal Meta-Cluster and 84.9% in the Peri-Central Meta-Cluster matched the reference clusters, underscoring our method's capacity to detect biologically meaningful patterns independent of predefined biomarkers. Here, k represents the number of clusters, and p denotes the number of selected genes used for clustering.

Comparative analysis of blood vessel identification

To benchmark the performance of different computational strategies for identifying blood vessels, we compared our spatially-aware CiCLoDS method against scDCC, PCA, and geneBasis. The performance of each method, with or without the inclusion of positional-encoded spatial data, was evaluated using Precision, Recall, and F1-score against a biomarker-based reference. Detailed method can be found and the key results are summarized in Fig. 4. Detailed description of the experiment can be found in Supplementary Section A.2.1.

The analysis revealed distinct requirements for each method to succeed. The scDCC successfully identified the blood vessel structures from gene expression data alone, achieving a high F1-score of 96.4%. The CiCLoDS achieved an F1-score of 92.8%, but only after its gene expression input was enriched with our positional-encoded spatial information; it failed to identify the structures without this spatial context.

The performance of the other baseline methods was varied. PCA, when applied in its standard configuration using the top 64 principal components, was unable to delineate the vessels. However, a modified approach, where we selected the 64 genes most represented in the top 20 principal components, did succeed in identifying the structures without spatial data, yielding an F1-score of 91.9%. In contrast, geneBasis failed to identify the blood vessels under any condition, and as a result, no metric is reported.

This result highlights our method's unique capability to successfully leverage spatial context when transcriptomic data alone is insufficient. Moreover, the performance of CiCLoDS (F1-score: 92.8%) represents a notable improvement over the modified PCA approach (F1-score: 91.9%), demonstrating its superior accuracy.

GO and KEGG pathway enrichment analyses

We assessed functional coherence of the gene sets using `g:Profiler` (organism: *Mus musculus*) across Gene Ontology (GO) Biological Process (BP), Cellular Component (CC), Molecular Function (MF) and KEGG pathways. Unless noted, we controlled the false discovery rate at FDR $< 0.05$and used the dataset-specific gene universe (all profiled genes after QC) as background[16–18].

Non-spatial signature. The 64-gene set derived from CiCLoDS without spatial information was enriched for hepatocyte metabolic programs consistent with canonical liver zonation. Significant terms (FDR < 0.05) included GO:BP related to lipid and carbohydrate/pyruvate metabolism, and KEGG pathways such as bile secretion and pyruvate metabolism, with leading genes spanning periportal gluconeogenesis (e.g., *Pck1*, *G6pc*) and pericentral xenobiotic/lipid metabolism (e.g., *Cyp2b9*, *Cyp2c23*, *Ugt2b1*, *Fasn*), supporting a periportal $\leftrightarrow$ pericentral axis.

Spatial signature. Including spatial coordinates shifted enrichment to membrane/adhesion and vascular terms. In the GO Cellular Component (CC) branch, we observed significant enrichment for **external side of plasma membrane** (GO:0009897; adjusted $p = 0.0054$; overlap 16/34), **side of membrane** (GO:0098552; adjusted $p = 0.0054$; overlap 18/41), **cell surface** (GO:0009986; adjusted $p = 0.0064$; overlap 21/54), and **cell periphery** (GO:0071944; adjusted $p = 0.0333$; overlap 39/146). Within GO Biological Process (BP), **cell–cell adhesion** (GO:0098609) was enriched (adjusted $p = 0.0169$; overlap 20/44). KEGG analysis highlighted **Lipid and atherosclerosis** (KEGG:05417; adjusted $p = 0.0199$; overlap 7/9 within the background). Collectively, these results indicate a shift from hepatocyte zonation toward cell-surface/adhesion and vascular features in the spatial set, consistent with resolution of blood-vessel niches in the tissue context. Detailed materials can be found in Supplement material 2.

Comparison of selected genes with competing methods

Comparison with PCA

To evaluate the feature selection capability of CiCLoDS, we conducted experiments on both spatial transcriptomics datasets: the Vizgen Mouse Liver Dataset and the Xenium Human Colon Dataset. We evaluated our model's feature selection ability by comparing the selected genes with those identified by the first principal component (PC1) of PCA. The comparison focuses on genes ranked as top contributors to PC1, as PCA is commonly used

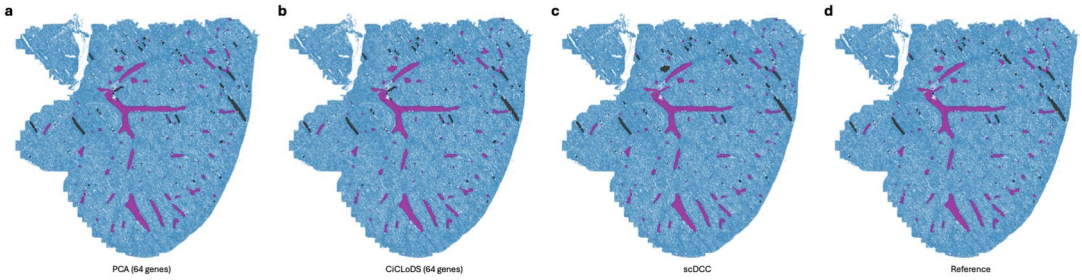

Fig. 4. Blood vessel vs. artifact classification in mouse liver. Panels (**a**–**d**) show PCA (64 genes), CiCLoDS (64 genes), scDCC (n=10, pairwise constraints from Leiden clusters), and the reference map, respectively. Geometrically detected void regions are colored as blood vessels (magenta) or artifacts (black) over the tissue background (blue); dashed red boxes highlight typical disagreements with the reference. All methods use the same two-step procedure: (i) detect voids by geometry; (ii) cluster cells within each void and label a region as a vessel if any member cell belongs to the vessel-associated cluster, otherwise artifact. Numbers printed in each panel report region-level Precision, Recall, and F1 for the blood-vessel class (positive class) against the VWF-derived reference.

to capture the most important features in high-dimensional data. To quantify the agreement between the two methods, we calculated the overlap percentage using the formula:

$$\text{Overlap Percentage} = \frac{|S_{\text{method}} \cap S_{\text{PCA}}|}{|S_{\text{PCA}}|} \times 100$$

where S_{method} represents the set of genes selected by our method with high frequency across runs, and S_{PCA} represents the set of top-ranking genes in PC1.

Before applying CiCLoDS, we performed dataset-specific preprocessing to ensure data quality. For the Vizgen mouse liver dataset, variance normalization was applied to adjust for differences in gene expression variability. For the Xenium human colon dataset, we excluded genes with no expression in more than 90% of cells to focus on biologically meaningful features.

We tested our model with $k = \{3, 5\}$, representing the number of clusters, and $p = \{16, 32, 64, 128\}$, representing the number of selected genes. Testing involved evaluating the overlap percentage between the genes selected by our method and those identified by PCA under these settings.

The results of the overlap analysis, shown in Fig. 5, reveal that the overlap with PCA-selected genes is consistently high across the Vizgen mouse liver dataset. As the number of selected genes increases, the consistency with PCA improves, particularly for the higher cluster settings. This suggests that including more genes enhances the stability of feature selection.

In contrast, the Xenium human colon dataset exhibits greater variability in overlap percentages. While the general trend of increased consistency with PCA as the number of selected genes grows holds, the variability is more pronounced, especially at smaller gene sets. This indicates that feature selection stability in the human dataset is more sensitive to the number of genes selected.

Comparison with competing methods

We evaluated CiCLoDS against three baselines—multi-PC PCA, geneBasis, and a random gene set (seed 42) using the Vizgen mouse-liver. Gene budgets were set to $p \in \{16, 32, 64, 128\}$ with $k \in \{10\}$ clusters. Following Kuemmerle *et al.*[19], we report: (i) kNN-overlap AUC measures how faithfully the selected genes preserve the original single-cell neighborhood structure—higher values mean the same cells remain neighbors after dimensionality reduction. (ii) RF accuracy gauges biological discriminability: random-forest classifiers trained only on the panel must recover the reference hepatocyte subclasses; the higher the accuracy, the more informative the panel. (iii) $1 - \mu|r|$ captures redundancy: it is one minus the average absolute correlation among the panel's genes, so larger scores indicate a more compact, non-redundant set. As shown in Fig. 6, CiCLoDS rapidly outperforms all baselines in neighborhood preservation: its kNN-overlap AUC climbs from 0.07 at $p = 16$ to 0.58 at $p = 64$ and reaches 0.89 at $p = 128$—more than double the next best method (geneBasis, 0.41). At very small budgets CiCLoDS favors highly correlated markers ($1-\mu|r| = 0.12$ for $p = 16$), but redundancy improves as the budget grows, rising to 0.63 at $p = 128$. Although this still trails geneBasis and PCA (0.73–0.84), CiCLoDS maintains a decisive lead in graph fidelity. geneBasis and PCA achieve the top random forest accuracies (0.92–0.93 across budgets), yet their AUC gains plateau at 0.33–0.41. The random baseline lags on nearly every metric. Exact numeric results appear in Supplementary Table 1.

Benchmarking generalization and initialization synergy on human DLPFC

To assess generalization across biological heterogeneity, we benchmarked our method on three samples (151507, 151669, 151673) obtained from three distinct donors in the human dorsolateral prefrontal cortex (DLPFC) dataset. We evaluated CiCLoDS in two contexts: (1) as a standalone spatial clustering tool with positional

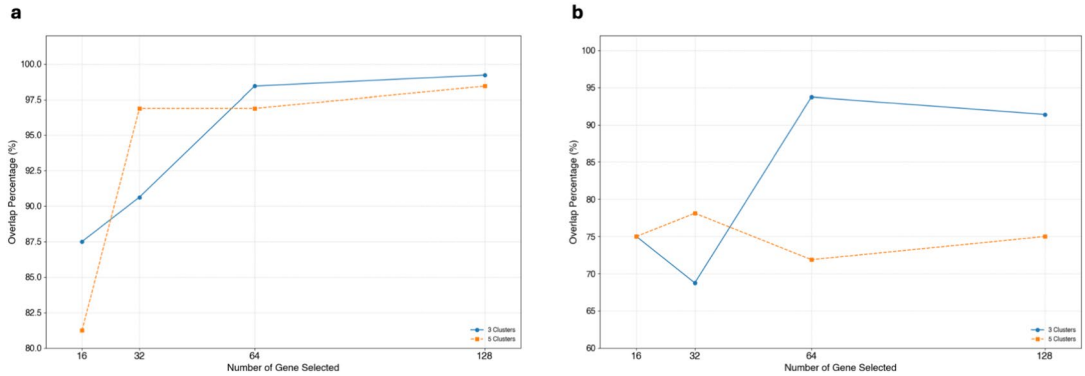

Fig. 5. Overlap of genes selected by our method with those identified by PCA across varying numbers of selected genes (p) and cluster configurations (k). **(a)** Vizgen Mouse Liver Dataset. **(b)** Xenium Human Colon Dataset. Blue and orange lines represent overlap percentages for cluster configurations with $k = 3$ and $k = 5$, respectively. The overlap percentage is calculated as the proportion of genes selected by our method that are also among the top-ranking genes in the first principal component (PC1) identified by PCA. Higher overlap percentages indicate greater agreement between our feature selection method and PCA.

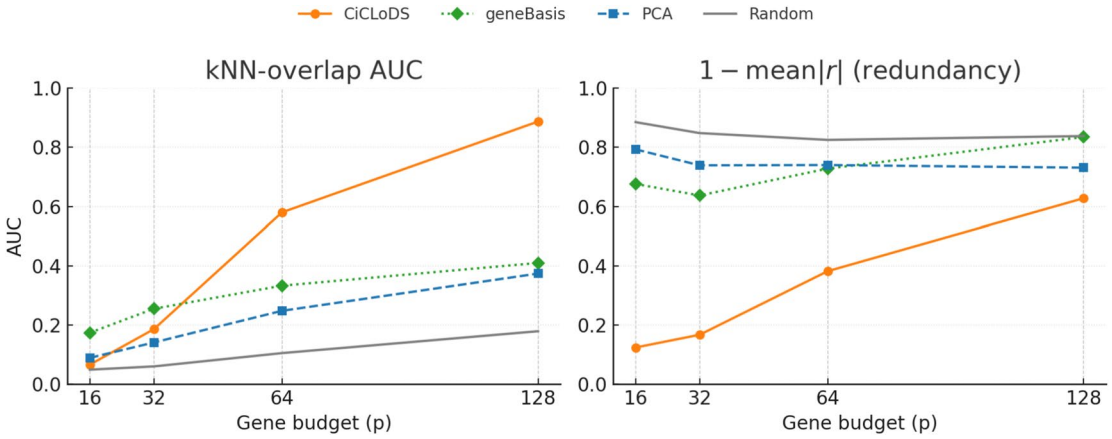

Fig. 6. Evaluating Structural Preservation versus Information Efficiency in Gene Selection Methods on mouse liver. Each curve follows one method—CiCLoDS (orange), PCA (blue), geneBasis (green), and a random panel (grey)—across four gene budgets ($p = 16, 32, 64, 128$). **(Left) kNN-overlap AUC:** the fraction of cell-to-cell neighbours preserved when the data are restricted to the selected genes (higher = better). **(Right) Redundancy**$1 - \mu|r|$: one minus the mean absolute correlation among the selected genes (higher = less redundancy). Values are evaluated one slice of mouse liver data. CiCLoDS overtakes all baselines on neighborhood preservation from $p \geq 64$ while steadily reducing redundancy, illustrating that larger, spatially informed panels can be both structurally faithful and information-efficient. Exact numbers—including random-forest accuracy used to assess predictive power—appear in Supplementary Table 1.

encodings (PE), and (2) as an initialization strategy for the probabilistic framework BayesSpace[20], termed the "Hybrid" approach. To ensure fair comparison, we matched the number of genes used in each method (i.e., 2000). A detailed description of the experimental settings can be found in Supplementary Section A.5.1.

Robust generalization across distinct biological donors

First, we compared CiCLoDS against a standard PCA baseline and two specialized spatial methods: BayesCafe[21] and BayesSpace. As shown in Fig. 7a, baseline performance varied significantly across donors. While BayesCafe performed well on Sample 151507 (ARI=0.49), it struggled to generalize to Sample 151669 (ARI=0.19). Similarly, BayesSpace showed high variance, dropping to an ARI of 0.29 on Sample 151669. In comparison, CiCLoDS achieved highly competitive performance, surpassing BayesCafe on average (Mean ARI 0.37 vs. 0.32) and performing comparably to BayesSpace (Mean ARI 0.40). Notably, on the challenging Sample 151669, CiCLoDS achieved an ARI of 0.52, significantly outperforming both BayesCafe (0.19) and BayesSpace (0.29). These results demonstrate that CiCLoDS is a capable standalone alternative to current state-of-the-art spatial clustering methods, offering distinct advantages in structural preservation for specific tissue architectures.

Enhancing probabilistic model stability via hybrid initialization

Second, we investigated whether the stability of CiCLoDS could benefit probabilistic models like BayesSpace, which are sensitive to initialization. By using CiCLoDS-derived partitions to initialize BayesSpace (the "Hybrid" approach), we observed a substantial improvement in overall robustness. The Hybrid method achieved the highest performance across all metrics (Mean ARI 0.50), effectively mitigating the local minima issues observed when BayesSpace is run with standard initialization. Visual inspection of Sample 151673 (Fig. 7b) confirms that the Hybrid method yields spatially coherent laminar structures that most closely match the Ground Truth annotations (ARI=0.56). This indicates that CiCLoDS can serve as a synergistic "warm-start" module, enhancing the reliability of downstream spatial refinement algorithms.

Robustness

Given the iterative nature of CiCLoDS, we assessed its stability in feature selection by performing gene selection on both Vizgen mouse liver and Xenium human colon datasets.

Stability was evaluated using runs initialized with different starting points. Consistency was quantified by the Jaccard Index, which measures the overlap of selected genes across runs, ranging from 0 (no overlap) to 1 (complete overlap). Higher values indicate greater agreement and reproducibility. To ensure robustness, the algorithm was executed 10 times, and the two runs with the largest loss values were excluded to minimize the impact of outliers. Stability was assessed by varying the number of selected genes ($p = 16,\ 32,\ 64,\ 128$) for each fixed number of clusters ($k = 3,\ 5,\ 10$).

Results, shown in Fig. 8, indicate that stability improved consistently with larger p values and varied based on k. In the mouse liver dataset, the Jaccard Index was consistently high, exceeding 0.85 when $p \geq 64$, regardless of k. These results demonstrate strong agreement between runs and the robustness of feature selection in this dataset (Metrics are displayed in the Supplementary section A.3).

In the human colon dataset, stability was lower at smaller p values and higher k values, reflecting the greater heterogeneity and complexity of this dataset. Specifically, at $p = 16$ and $k = 10$, the Jaccard Index was below

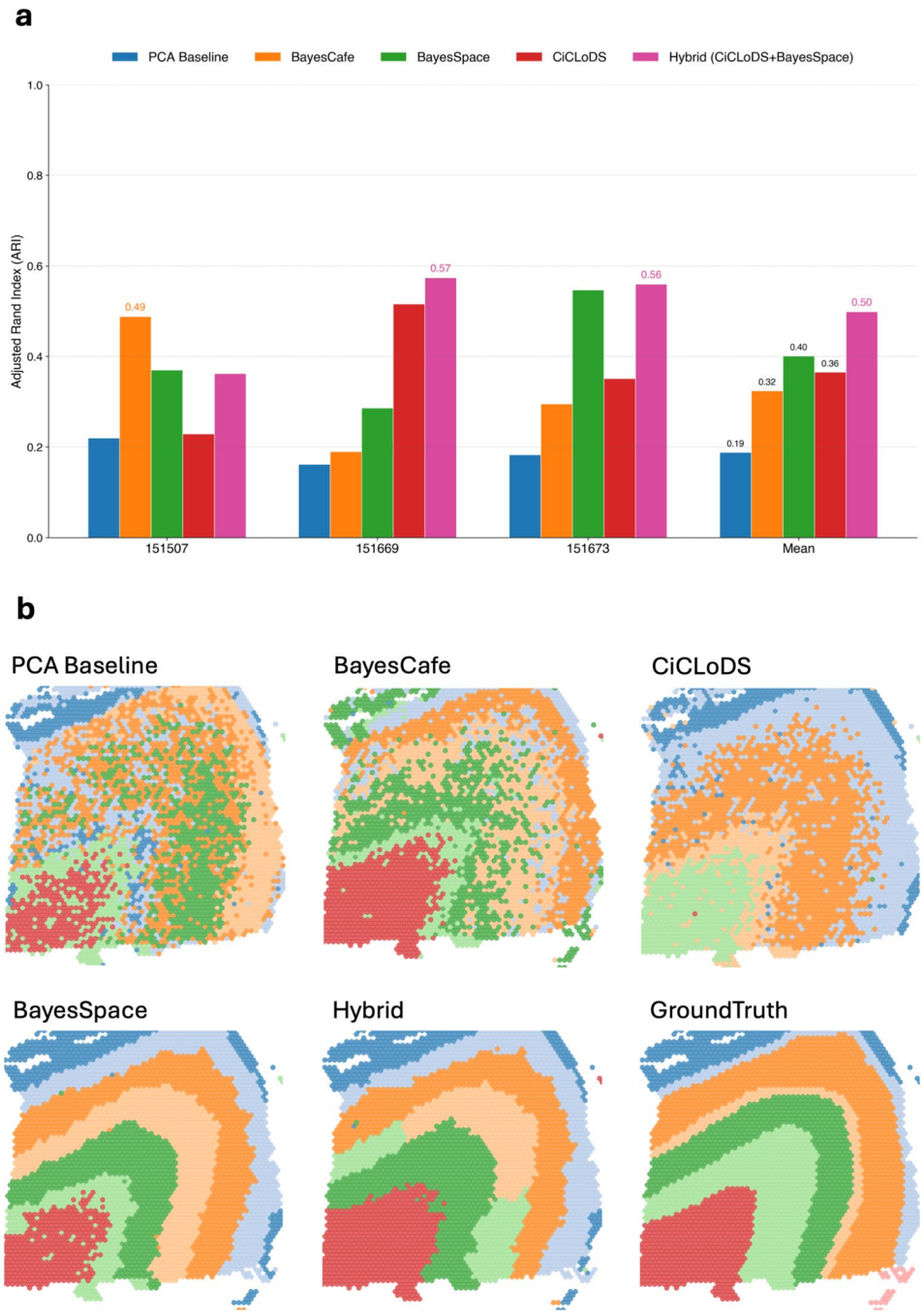

Fig. 7. Benchmarking clustering robustness and generalization on human dorsolateral prefrontal cortex (DLPFC). **(a)** Quantitative comparison of Adjusted Rand Index (ARI) scores across three distinct donors (Samples 151507, 151669, and 151673). We evaluated five methods: a PCA Baseline, BayesCafe, BayesSpace, CiCLoDS (ours), and a Hybrid approach (CiCLoDS initialization refined by BayesSpace). While specialized spatial methods like BayesCafe and BayesSpace exhibit performance fluctuations across donors (e.g., BayesCafe drops to ARI=0.19 on Sample 151669), the Hybrid method demonstrates superior stability, achieving the highest mean ARI of 0.50. **(b)** Visual comparison of spatial domains identified in Sample 151673. The PCA Baseline produces fragmented clusters that fail to capture tissue organization. While BayesSpace recovers the general laminar structure, the Hybrid method yields spatially coherent domains that most closely resemble the Ground Truth manual annotations (ARI=0.56).

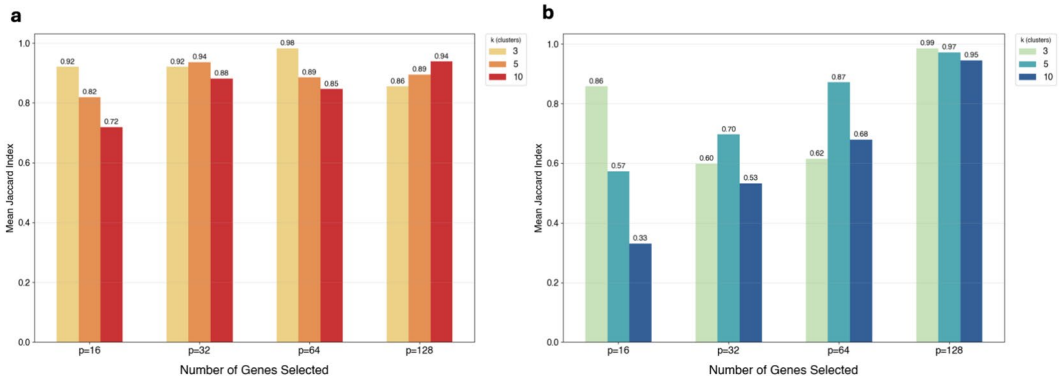

Fig. 8. Stability of feature selection across different numbers of selected genes (p) and cluster configurations (k) for two spatial transcriptomics datasets. **(a)** Vizgen Mouse Liver Dataset: Mean Jaccard Index measuring the overlap of selected genes across multiple runs with varying numbers of selected genes ($p = 16, 32, 64, 128$) and cluster configurations ($k = 3, 5, 10$). **(b)** Xenium Human Colon Dataset: Mean Jaccard Index under the same parameter settings. Blue and orange lines represent cluster configurations with $k = 3$ and $k = 5$, respectively. Higher Jaccard Index values indicate greater consistency and reproducibility in feature selection across different runs.

0.6, indicating moderate agreement. However, stability improved substantially with larger p, reaching values comparable to the mouse liver dataset at $p = 128$. This trend highlights the importance of selecting a sufficient number of genes to mitigate instability introduced by increased clustering complexity.

Overall, the Jaccard Index values across different p and k settings demonstrated the robustness of our method. The results underscore its ability to maintain stability across datasets with varying levels of complexity, while offering flexibility in parameter selection to optimize performance.

Model efficiency

In addition to robustness, we evaluated the computational efficiency of our method to ensure its practicality for large-scale spatial transcriptomics analyses. Model efficiency was assessed on two different hardware configurations to demonstrate versatility and scalability.

For both the Vizgen Mouse Liver and Xenium Human Colon datasets, our method converged within 20 iterations and completed the analysis in under 5 minutes. These experiments were conducted on the following systems:

- **Apple M3 Pro System**: Equipped with 36 GB of memory.
- **Dell Inspiron 15 3520**: Powered by a 12th Gen Intel(R) Core(TM) i5-1235U processor (1300 MHz, 10 Core(s), 12 Logical Processors, with 16 GB of memory).

The consistent convergence time and low computational overhead across these different hardware setups demonstrate the efficiency and scalability of our approach. This makes our method suitable for processing large-scale spatial transcriptomics data, providing reliable performance without significant computational demands.

Discussion

We view CiCLoDS primarily as a general mathematical model for structure discovery in high–dimensional measurements rather than as a task–specific pipeline. Conceptually, the model is close in spirit to principal component analysis (PCA): it posits that observations concentrate near a low–dimensional structure and seeks a parsimonious representation. Unlike PCA, CiCLoDS incorporates joint sparsity and clustering within a single objective. The sparsity promotes concise feature sets and interpretability, while the clustering term partitions the latent structure into coherent groups. This coupling avoids the disconnect that arises in sequential "select–then–cluster" or "cluster–then–select" procedures and yields representations that are simultaneously compact and discriminative.

This perspective explains the method's behavior across settings. When spatial information is absent, the model reduces to learning a low–dimensional, sparse subspace that captures dominant biological variation (e.g., hepatocyte zonation in liver). When spatial coordinates are available, we append positional encodings (PE) to the expression features, which steers the same objective toward structures consistent with tissue architecture (e.g., vascular niches) without changing the optimization itself. No task–specific rules are required; the same objective interpolates between expression–only and spatially aware regimes. In this sense, CiCLoDS inherits PCA's generality while adding an interpretable selection mechanism and a principled handle to incorporate side information.

Our benchmarking on human DLPFC data further illuminates the dual nature of CiCLoDS within the spatial analysis ecosystem: it is both competitively standalone and cooperatively synergistic. As a standalone tool, CiCLoDS matches or exceeds the performance of specialized probabilistic models like BayesCafe and BayesSpace, demonstrating that a rigorous linear objective can capture complex spatial domains without the computational

overhead of full probabilistic inference. Simultaneously, it functions cooperatively as a deterministic "warm-start" module. By providing a stable, spatially coherent initialization, CiCLoDS mitigates the sensitivity to local minima often observed in probabilistic frameworks. This "Hybrid" capability suggests that CiCLoDS need not be viewed solely as a replacement for existing methods, but also as a foundational stabilizer that enhances the reliability of sophisticated downstream refinement algorithms.

A second advantage of the formulation is stability and scalability. Although the joint optimization is non-convex, the problem decomposes into block-wise sub-problems that admit efficient, exact solutions, so the method converges quickly on standard hardware. The induced sparsity controls variance and reduces redundancy, which improves downstream interpretability (e.g., GO/KEGG analyses) without excessive tuning. As with all unsupervised models, hyperparameters such as the number of clusters and the feature budget influence granularity: small budgets emphasize core signals, larger budgets capture finer structure. In practice, we find that choosing the feature budget in a modest range (e.g., dozens to low hundreds) and selecting the cluster number to match expected biological complexity is sufficient; performance curves across settings illustrate a broad plateau of competitive accuracy rather than a single narrow optimum. Because the objective is modality–agnostic within a matrix setting, CiCLoDS applies directly to transcriptomic data with or without spatial coordinates. In the spatial case, we incorporate position by appending positional encodings (PE) to the expression features. This preserves local contiguity while maintaining global separability, enabling joint recovery of molecular programs and anatomical compartments without altering the core objective.

There are limitations. CiCLoDS is primarily linear with sparse loadings; strongly nonlinear manifolds may not be captured without kernelized or deep variants. As a general mathematical method, it does not prescribe a fixed preprocessing pipeline. In our analyses we use a minimal scheme with normalization and feature-wise scaling by dividing each gene by its sample standard deviation, but alternative choices (e.g., different normalizations or variance-stabilizing transforms) can change relative weights and thus outcomes. Hyperparameters—the number of clusters k and the feature budget p—control granularity and redundancy; different settings may yield different partitions, so the selection protocol and ranges should be reported. Spatial information is incorporated via positional encodings (PE); their basis and scaling affect locality bias and should be documented. Finally, pathway enrichment depends on background definition, identifier mapping, and annotation coverage, so enrichment results should be interpreted as corroborative rather than definitive. We document all settings and data processing in Methods to aid reproducibility.

In summary, CiCLoDS is best understood as a general, PCA–adjacent model for simultaneous representation learning, feature selection, and clustering. Its combination of simplicity, speed, and interpretability positions it as a versatile foundation for spatial analysis: it serves effectively as both a competitive standalone tool and a synergistic stabilizer for complex probabilistic frameworks. By providing a clean pathway to anatomical specificity—whether through direct analysis or as a robust initialization—CiCLoDS offers a scalable default for structure discovery in high–dimensional biological data.

Data availability

The datasets used and/or analyzed during the current study are available at the following sources: https://info.vizgen.com/mouse-liver-data; https://www.10xgenomics.com/datasets.

Code availability

https://github.com/hym97/CiCLoDS/.

Received: 21 April 2025; Accepted: 3 February 2026
Published online: 09 February 2026

References

1. Marx, Vivien. Method of the Year: spatially resolved transcriptomics. *Nature methods* **18**(1), 9–14 (2021).
2. Tian, Luyi, Chen, Fei & Macosko, Evan Z. The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41**(6), 773–782 (2023).
3. 10x Genomics Knowledge Base. What is the spatial resolution and configuration of the capture area of the Visium v1 Gene Expression Slide? https://kb.10xgenomics.com/hc/en-us/articles/360035487572. (2023).
4. Kok Hao Chen et al. Spatially resolved, highly multiplexed RNA profiling in single cells. In: *Science* 348 6233, aaa6090 (2015).
5. 10x Genomics. Xenium Analyzer Specifications. Tech. rep. CG000630 Rev F. 10x Genomics, https://cdn.10xgenomics.com/image/upload/v1721758529/support-documents/CG000630_XeniumAnalyzer_Specifications_RevF.pdf (2024).
6. Missarova A, et al. geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq. In: *Genome biology* **22**(1), 333 (2021)
7. Liang, Shaoheng et al. Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nature computational science* **1**(5), 374–384 (2021).
8. Tian, Tian et al. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence* **1**(4), 191–198 (2019).
9. Tian, Tian et al. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature communications* **12**(1), 1873 (2021).
10. Wang J, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. In: *Nature communications* **12**(1), 1882 (2021)
11. Witten, Daniela M. & Tibshirani, Robert. A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**(490), 713–726 (2010).
12. Elhamifar, Ehsan & Vidal, René. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence* **35**(11), 2765–2781 (2013).
13. Chang, Wei-Chien. On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **32**(3), 267–275 (1983).

14. Ng A, Jordan M, & Weiss Y. On spectral clustering: Analysis and an algorithm. In: *Advances in neural information processing systems* **14,** (2001).
15. Sun, Tianliang et al. AXIN2+ pericentral hepatocytes have limited contributions to liver homeostasis and regeneration. *Cell stem cell* **26**(1), 97–107 (2020).
16. Kanehisa, Minoru et al. KEGG: biological systems database as a model of the real world. *Nucleic acids research* **53**(D1), D672–D677 (2025).
17. Kanehisa, Minoru & Goto, Susumu. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**(1), 27–30 (2000).
18. Kanehisa, Minoru. Toward understanding the origin and evolution of cellular organisms. *Protein science* **28**(11), 1947–1951 (2019).
19. Kuemmerle, Louis B. et al. Probe set selection for targeted spatial transcriptomics. *Nature methods* **21**(12), 2260–2270 (2024).
20. Zhao, Edward et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology* **39**(11), 1375–1384 (2021).
21. Li H, et al. An interpretable Bayesian clustering approach with feature selection for analyzing spatially resolved transcriptomics data. In: *Biometrics* **80**(3), ujae066 (2024).

Acknowledgments

This work was partially supported by the National Institute of Health grants from the National Institute on Aging (NIH U01AG066833) and the National Library of Medicine (NIH 5R13LM014478).

Author contributions

Author contributions. Oversaw the project: XH and SZ. Conceived and designed the experiments, performed the experiments: XH, SZ, NW and YH. Worked on numerical testing and analyzed the data, contributed reagents/materials/analysis tools: NW, YH, ER, KJW, CC, DL, JM, SZ and XH. Wrote the paper, revised and finalized the paper: NW, YH, XH, SZ, and all other authors.

Declarations

Competing interests

The authors declare no competing interests. **Acknowledgements**. This work was partially supported by the National Institute of Health grants from the National Institute on Aging (NIH U01AG066833) and the National Library of Medicine (NIH 5R13LM014478).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-39168-1.

Correspondence and requests for materials should be addressed to S.Z. or X.H.