Mapping the Gene Space at Single-Cell Resolution with Gene Signal Pattern Analysis

Aarthi Venkat
Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard
avenkat@broadinstitute.org

Sam Leone
Yale University
sam.leone@yale.edu

Scott Elliot Youlten
Yale University
scott.youlten@yale.edu

Hannah Thomas
Nashua High School South
hannahgthomas8@gmail.com

Eric Fagerberg
Yale University
eric.fagerberg@yale.edu

John Attanasio
Yale University
john.attanasio@yale.edu

Nikhil S. Joshi
Yale University
nikhil.joshi@yale.edu

Michael Perlmutter
Boise State University
mperlmutter@boisestate.edu

Smita Krishnaswamy
Yale University
smita.krishnaswamy@yale.edu

Abstract

In single-cell and spatial sequencing analysis, several computational methods have been developed to map the cellular state space, but little has been done to map or create embeddings of the gene space. Here we formulate the gene embedding problem, design tasks with simulated single-cell data to evaluate representations, and establish relevant baselines. We then present *gene signal pattern analysis* (GSPA), a graph signal processing approach that learns rich gene representations from single-cell data using a dictionary of diffusion wavelets on the cell–cell graph. This approach embeds genes based on their patterning and localization on the cellular manifold, enabling characterization of genes for diverse biological tasks, including identifying gene coexpression modules and gene subnetworks associated with patient phenotypes.

1 Introduction

Techniques to map the cellular state space in single-cell RNA sequencing (scRNA-seq) embed cells in low-dimensional spaces based on transcriptional similarity, revealing clusters of cells or trajectories along phenotypic continuums. Gene expression is also highly organized, coordinated into complexes and pathways, but existing cell embedding methods cannot capture this gene landscape due to biological and technical noise (e.g. dropout due to sampling inefficiency) [1, 2].

We address this by framing genes as *signals* on a cell-cell graph and applying graph signal processing to learn their representations [3]. We first construct a cell-cell graph and diffusion operator $\mathbf{P}$, where powering $\mathbf{P}$ to t gives the transition probabilities of a t-step random walk. We then construct a dictionary of multiscale diffusion wavelets [4], which power $\mathbf{P}$ to multiple t, and decompose each gene into a set of graph diffusion wavelet coefficients. We reduce the dimensionality of this representation with an autoencoder, enabling downstream tasks while preserving gene-gene relationships. We demonstrate this on simulated [5], CD8+ T cell [6], and tumor-associated immune [7] single-cell transcriptomic data, as well as lymph node [8] and hepatocellular carcinoma [9] spatial transcriptomic data. Our approach enables analysis of gene coexpression modules, localized signatures, perturbation-specific networks, spatial variability, cell communication, and patient response. This work extends our recently published study [10] with new comparisons and insights into GSPA for identifying gene subnetworks linked to patient phenotypes from spatial transcriptomics (Figure 1). These results demonstrate the utility of learning multiscale representations of graph signals.

Venkat et al., Mapping the Gene Space at Single-Cell Resolution with Gene Signal Pattern Analysis (Extended Abstract). Presented at the Fourth Learning on Graphs Conference (LoG 2025), Hybrid Event, December 10–12, 2025.

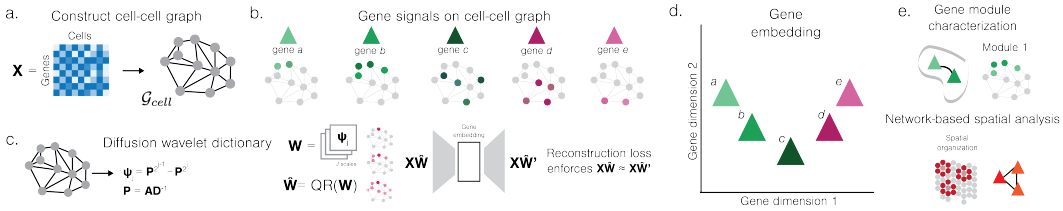

Figure 1: Overview of Gene Signal Pattern Analysis. a. Construction of a cell–cell graph. b. Demonstrative gene signals, where signals are functions defined on nodes of cell–cell graph. c. Construction of diffusion wavelet dictionary $\mathbf{W}$, or QR-factorized dictionary $\hat{\mathbf{W}}$, consisting of diffusion wavelets for scales $1, ..., J$. Gene signals are projected onto dictionary and embeddings are learned via autoencoder. d. Demonstrative gene embedding. e. Example downstream applications.

2 Background

Single-cell transcriptomics measures thousands of genes (features) per cell (observation), yet cellular behavior is constrained to a limited set of fates, motivating the *manifold assumption* that high m-dimensional observations lie upon a low d-dimensional manifold (i.e., a d-dimensional subset of $\mathbb{R}^m$ that is locally equivalent to $\mathbb{R}^d$) for some $d \ll m$ (see [11] for review of manifold learning for single-cell data). To approximate the unknown manifold, most manifold learning approaches first construct a graph $\mathcal{G}_{cell} = (V_{cell}, E_{cell})$ with adjacency matrix $\mathbf{A}$ and diffusion operator $\mathbf{P} = \mathbf{A}\mathbf{D}^{-1}$, where $\mathbf{D}$ is the diagonal degree matrix ($\mathbf{D}_{i,i} = \sum_j \mathbf{A}_{i,j}$, $\mathbf{D}_{i,j} = 0$ if $i \neq j$). $\mathbf{P}$ describes the transition probabilities of a lazy random walker, and powering $\mathbf{P}$ to diffusion time t corresponds to averaging signal $\mathbf{x}$ over t-step random walks.

From these observations, [12] introduced diffusion maps to embed datapoints into a low-dimensional Euclidean space $\mathbb{R}^d$, $d \ll m$, parameterized by scale t. Formally, we take the eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ and corresponding eigenvectors $\{\phi_j\}_{j=1}^N$ of $\mathbf{P}$ and map each point $x_i \in X$ to a d-dimensional vector $\Phi_t(x_i) = [\lambda_1^t \phi_1(x_i), \ldots, \lambda_N^t \phi_d(x_i)]^T$. Small values of t capture local representations and large t capture global representations. Inspired by classical wavelet constructions (e.g., [13]), diffusion wavelets [4] extend diffusion maps for multiscale data representations by computing the difference in powered diffusion operators, e.g. $\mathbf{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j}$. Diffusion wavelets have played a powerful role in graph signal processing [3] and in geometric deep learning [14, 15].

3 Problem Setup

Given a single-cell sequencing dataset consisting of m genes and their measurements in n cells, organized into an $m \times n$ matrix $\mathbf{X}$, with the insight that gene measurements, like cell measurements, may also be compressed into a lower-dimensional space for analyzing gene-gene relationships, we aim to obtain a low-dimensional representation of each gene which preserves the inherent structure of the gene space with respect to the cellular manifold. In particular, we seek a reduced dimensional map $\Theta : \mathbb{R}^n \to \mathbb{R}^d$, $d \ll n$, which satisfies three desired properties: preservation of local and global signal distances, noise robustness, and flexibility to downstream tasks (see Section A.1).

4 Related Work

Techniques to calculate gene-gene relationships, which can be used to construct gene embeddings, fall into (1) diffusion-based manifold learning, (2) optimal transport between gene signals on the graph, and (3) joint gene-cell embeddings (see Section A.2). (1) MAGIC [16] denoises the gene expression matrix by left-multiplying $\mathbf{X}$ to $\mathbf{P}^t$. Eigenscores [17] rank genes by alignment with first $r \ll n$ left Laplacian eigenvectors. These methods map genes by their low-frequency patterning on the cell-cell graph at a single scale. (2) DiffusionEMD [18] computes optimal transport with multiscale diffusion kernels, and GFMMD [19] calculates signal distances by analytically solving for a smooth optimal witness function. (3) siVAE [20] is a variational autoencoder with two encoders (cell-wise and gene-wise) and a combined decoder which outputs each gene's expression for each cell. SIMBA [21] constructs and embeds a heterogeneous graph consisting of gene and cell nodes, where

genes are connected to cells they are expressed in. In experiments, we embed high-dimensional gene representations with an autoencoder for accurate comparison, and we compare GSPA to additional baselines: embedding the raw measurements $\mathbf{X}$ with an autoencoder, and constructing a gene-gene k-NN graph from $\mathbf{X}$ and embedding the graph with Node2Vec [22] or a graph autoencoder [23].

5 Gene Signal Pattern Analysis

5.1 Model overview

To construct the map Θ, we make the critical observation that the expression pattern $\mathbf{X}_i$ for gene i can be described as a signal (function) defined on the nodes of a cell-cell similarity graph $\mathcal{G}_{cell}$. We thus compare how gene expression patterns are similar to and different from each other based on distances along the cellular manifold, achieving gene embedding desiderata (see Section A.3).

5.1.1 Constructing a cell-cell similarity graph from single-cell data

First, we build a graph $\mathcal{G}_{cell} = (V_{cell}, E_{cell})$, where each node in V_{cell} corresponds to a cell, and each edge $E_{v_1 v_2}$ in E_{cell} describes the similarity between cell v_1 and cell v_2. To build $\mathcal{G}_{cell}$, we compute the Euclidean distances between all pairs of cells and apply an α-decaying kernel to calculate affinities. The α-decaying kernel is defined as $K_{k,\alpha}(v_1, v_2) = \frac{1}{2}\exp\left(-\left(\frac{\|v_1 - v_2\|_2}{\varepsilon_k(v_1)}\right)^{\alpha}\right) + \frac{1}{2}\exp\left(-\left(\frac{\|v_1 - v_2\|_2}{\varepsilon_k(v_2)}\right)^{\alpha}\right)$, where v_1 and v_2 are cells $\in V_{cell}$, viewed as points in $\mathbb{R}^m$ corresponding to columns of $\mathbf{X}$, $\varepsilon_k(v_1), \varepsilon_k(v_2)$ are the distance from v_1, v_2 to their k-th nearest neighbors, respectively, and α controls the decay rate [24]. We describe scalability for large graphs in Section A.4.

5.1.2 Building dictionary of graph diffusion wavelets for gene representation

Given cellular graph $\mathcal{G}_{cell}$, we construct a multiscale representation with diffusion wavelets. We first define $\mathbf{P} = \mathbf{A}\mathbf{D}^{-1}$ as the diffusion operator. Then, each wavelet of scale j centered at vertex v can be calculated by $\mathbf{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j}$ for $1 \leq j \leq J$ (and $\mathbf{\Psi}_0 = \mathbf{I} - \mathbf{P}$) and extracting the v-th row via $\delta_v^T \mathbf{\Psi}_j$, where δ_v is the Kronecker delta centered at the v-th vertex. Then $\{\mathbf{\Psi}_j^T \delta_v\}_{v \in V_{cell}, j \in 0,1 \ldots J}$ defines our wavelet dictionary $\mathbf{W}$ of shape $n \times Jn$, where small j capture local representations, and large j capture global. Number of scales J is defined as the log of the number of cells n based on Lemma A.2 introduced and proven in [18]. Because the diffusion operator $\mathbf{P}$ is smoothing, we assume the numerical rank decreases as we take powers of the operator [4], and a small set of large wavelets can describe the graph at a coarse resolution. Therefore, to remove redundant wavelets, we can perform QR factorization and get a compressed wavelet dictionary $\widehat{\mathbf{W}} = \{\widetilde{\mathbf{\Psi}}_j^T \delta_v\}_{v \in V_{cell}, j \in 0,1 \ldots J}$, where for each j, $\widetilde{\mathbf{\Psi}}_j$ is a set of linear combinations of wavelets at j that account for the most variance. For large j, QR factorization naturally computes the numerical rank of $\mathbf{\Psi}_j$ by taking a linear combination to form $\widetilde{\mathbf{\Psi}}_j$ such that the total error in projecting $\mathbf{\Psi}_j$ onto $\widetilde{\mathbf{\Psi}}_j$ is less than some ϵ fraction of the norm of $\mathbf{\Psi}_j$. We evaluate GSPA with (GSPA+QR) and without (GSPA) factorization.

5.1.3 Projecting gene signals onto wavelet dictionary

Each gene signal $\mathbf{X}_i$ of shape $1 \times n$ corresponds to the expression of the gene in the cellular state space. Given all gene signals $\mathbf{X}$ and wavelet dictionary ($\widehat{\mathbf{W}}$ or $\mathbf{W}$), we project $\mathbf{X}$ onto the dictionary ($\mathbf{X}\widehat{\mathbf{W}}$ or $\mathbf{X}\mathbf{W}$). This reveals each gene signal's spatial and frequency information over the corresponding cell-cell graph $\mathcal{G}_{cell}$. Theorem A.1 shows that the wavelet projection $\mathbf{X} \to \mathbf{X}\mathbf{W}$ is continuous with respect to an Unbalanced Diffusion Earth Mover's Distance (UDEMD) [25], which describes how similar two gene signals $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$ are in a manner informed by the geometry of the cellular graph.

5.1.4 Learning low-dimensional representation with autoencoder

To reduce redundancy and improve computational tractability for downstream analysis, we reduce the dimensionality of $\mathbf{X}\widehat{\mathbf{W}}$ with autoencoder $D \circ E$ where the objective is to minimize the mean squared error. That is, $\mathbf{X}\widehat{\mathbf{W}}' \approx D(E(\mathbf{X}\widehat{\mathbf{W}}))$, so that $\|\mathbf{X}\widehat{\mathbf{W}} - \mathbf{X}\widehat{\mathbf{W}}'\|_2^2 = \sum_{i \in \mathbf{X}} \|\mathbf{X}_i\widehat{\mathbf{W}} - \mathbf{X}_i\widehat{\mathbf{W}}'\|_2^2$ is as small as possible. The latent representation $E(\mathbf{X}\widehat{\mathbf{W}})$ is the embedding we evaluate and characterize in downstream analysis, i.e. $GSPA(\mathbf{X}) = E(\mathbf{X}\mathbf{W})$ and $GSPA\text{+}QR(\mathbf{X}) = E(\mathbf{X}\widehat{\mathbf{W}})$ where $\mathbf{W}$

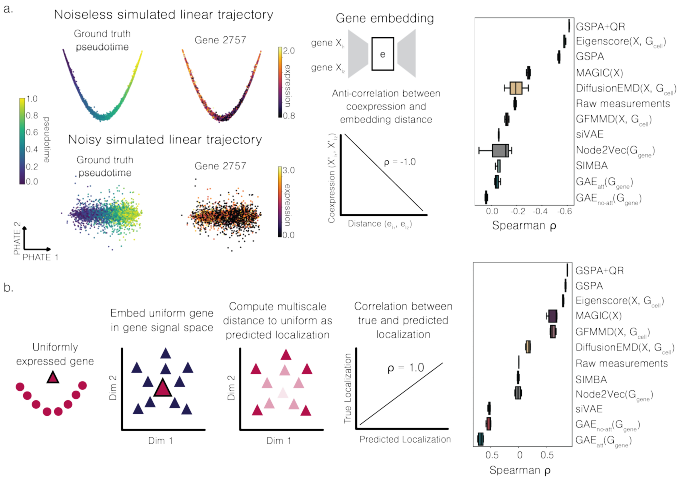

Figure 2: Coexpression and localization experiments over three random seeds. a. Noiseless and noisy cell embeddings of simulated linear trajectory, colored by pseudotime and example gene expression. Experimental set-up and coexpression evaluation. b. Experimental set-up and localization evaluation.

and $\widehat{\mathbf{W}}$ are uncompressed and compressed wavelet dictionaries (respectively), E is the encoder discussed above, and $GSPA$ and $GSPA + QR$ are taken to be the map Θ in the problem setup.

5.2 Differential localization reveals genes specific to populations without prior clustering

Beyond preserving relationships within the gene space, GSPA also preserves distances to any signal defined on the cell–cell graph. This enables ranking genes based on informativeness for characterizing cell-cell variation, termed *differential localization*, which proves useful in cases where it is difficult to assign cluster labels and annotate cell types [26] (see Section A.6). Based on the observation that uniformly expressed genes are least likely to be involved in cell state-defining biological processes [27], the gene localization score is calculated as the distance between each gene embedding and an embedding of a uniform (constant) signal $\mathbf{u} = \frac{1}{\sqrt{n}}\mathbf{1}$ (Definition A.1). Localized genes (with high scores) can thus be used for pathway enrichment analysis and cluster-independent feature selection.

5.3 GSPA for multiple modalities (GSPA-multimodal)

Where we have datasets of the same datapoints with multiple modalities, we can construct a combined representation using integrated diffusion [28], which constructs affinity graphs for each modality (e.g., for two modalities, $\mathcal{G}_1$ and $\mathcal{G}_2$). Then, each graph has associated diffusion filters $\mathbf{P_1}^{t_1}$ and $\mathbf{P_2}^{t_2}$, where t_1 may not equal t_2 due to differing degrees of noise. Finally, the integrated diffusion operator is calculated by multiplying diffusion filters, i.e. $\mathbf{P_{integrated}} = \mathbf{P_1}^{t_1}\mathbf{P_2}^{t_2}$. We can use this operator to construct an integrated wavelet dictionary and project signals for downstream analysis, as described above. For spatial transcriptomic data, each spot has two measurements: spatial coordinates and expression. Thus, using a version of integrated diffusion for spatial transcriptomic data [29], gene embeddings from GSPA-multimodal are informed by both spatial and expression similarity.

6 Experiments

6.1 Comparison to alternative gene mapping strategies

We evaluated embeddings on three simulated single-cell datasets with differing latent structure (linear, 2 branches, and 3 branches), where each dataset had corresponding noisy $\mathbf{X}$ and unseen true (noiseless) $\mathbf{X}'$ counts [5]. Defining coexpression between genes i_1 and i_2 as the correlation of true counts $\mathbf{X'}_{i_1}$ and $\mathbf{X'}_{i_2}$, we learned gene embeddings from the noisy counts $\mathbf{X}$ and compared the anti-correlation between embedding distance and true coexpression for GSPA and baselines (Figures 2a, A.2, Table 1). We next benchmarked embeddings for their ability to capture gene localization, generating simulated signals with "ground truth" localization and calculating each signal's localization score (Figures 2b, A.3, A.4, Table 1). GSPA+QR outperformed baselines irrespective of normalization and graph construction (Figure A.5a-b), and methods using the cell-cell graph showed better overall performance, supporting our assertion that the graph can improve gene-gene analysis (Figure A.5c). Finally, both QR factorization and the autoencoder contributed to performance gains (Table 2).

6.2 Coexpression in CD8+ T cells with gene embeddings

To investigate CD8+ T cell plasticity in response to infection [30], we analyzed a newly developed dataset comprising antigen-specific CD8+ T cells sequenced at three timepoints from acute and chronic LCMV infections [6] (Figure 3a). Cluster-derived differentially expressed genes, including markers of interest, showed expression in multiple clusters, (Figure A.6a-b), motivating mapping the gene space to capture distinct T cell signatures. GSPA+QR gene embeddings organize genes by these signatures, with gene modules corresponding to memory, naiveness, proliferation, effector, exhaustion, and type 1 interferon response (Figure 3b, Figure A.6c). Localized genes within each module showed significantly higher interaction than expected in STRINGdb [31] (Figure 3c, Figure A.6c). Furthermore, re-embedding cells with localized genes preserved overall manifold structure, highlighting localization for topologically-informed feature selection (Figure A.7). To evaluate how well baselines capture subtle but known type 1 interferon signaling in chronic conditions [32, 33], we identified gene modules in each baseline and compared gene set enrichment [34] of the gene module with type 1 interferon marker *Irf7*. Cell clusters and baseline gene modules showed low to no enrichment of type 1 interferon signaling, suggesting GSPA and GSPA+QR uniquely identify this signature. By capturing strong gene coexpression patterns, GSPA additionally enables derivation of perturbation-specific gene-gene networks (Figure A.6d) and shows improved classification of patient immunotherapy response (Figure A.8) compared to traditional single-cell response prediction baselines.

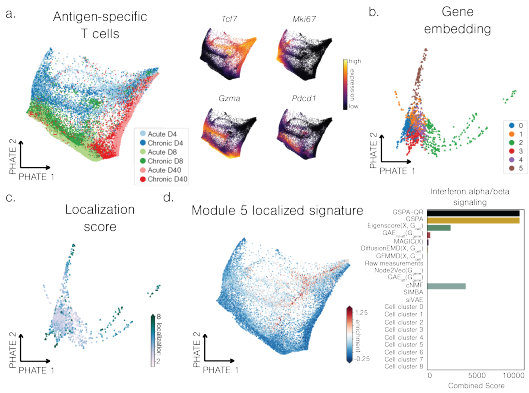

Figure 3: GSPA for CD8+ T cells. a. Antigen-specific CD8+ T cells colored by condition and marker gene expression. b. Gene embedding, colored by gene module assignment. c. Gene embedding, colored by localization score. d. Enrichment of top localized genes of GSPA+QR gene module 5, and type 1 interferon signaling enrichment for top genes from all comparisons.

6.3 Spatially-organized gene networks with GSPA-multimodal

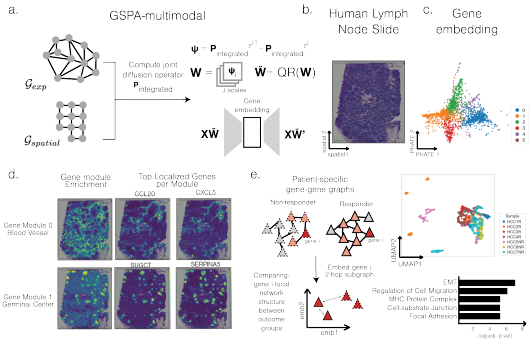

Figure 4: GSPA-multimodal overview. a. Schematic for spatial transcriptomics. b. H&E stain. c. Gene embedding with gene module assignment. d. Visualization of two gene modules and localized genes. e. GSPA-multimodal for HCC response.

GSPA-multimodal derives gene embeddings from multimodal data via an integrated diffusion operator $\mathbf{P}_{integrated}$ [28, 29], such that we can construct an integrated wavelet dictionary informed by all modalities (Figure 4a). Applied to 10x Visium lymph node data (Figure 4b, [8]), GSPA-multimodal calculates gene embeddings and localized genes representing distinct tissue substructures (Figure 4c-d), enabling spatially variable gene and cell communication analysis (Figure A.9).

To assess how gene networks locally rewire with patient phenotypes, we extended GSPA-multimodal by constructing patient-specific gene-gene k-NN networks from GSPA embeddings. Then, for each gene in each network, we extracted the local gene neighborhood and embedded it with attributed Graph2Vec [35], such that each point represents the local network for a given patient gene expression profile. We then rank genes based on the difference in structural similarity within versus between phenotypic groups, identifying genes with stronger inter-group differences (Definition A.2, Figure 4e). Analyzing four responder and three non-responder samples from patients with hepatocellular carcinoma (HCC) [9] (Figure 4e), this revealed genes enriched for adhesion, immune regulation, and plasticity pathways, implicating structural and microenvironmental remodeling in HCC response [34].

Author Contributions

A.V., S.L., S.E.Y. and S.K. developed and implemented GSPA. A.V. and H.T. designed and performed computational analyses, overseen by S.K. E.F. and J.A. performed scRNA-seq and perturbation experiments, overseen by N.S.J. A.V. wrote the abstract with contributions from all authors.

Acknowledgments

A.V. acknowledges funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. M.P. is funded by NSF DMS grant number 2327211 and NSF OIA grant number 2242769. S.K. is funded by NSF Career Grant number 2047856, NSF DMS grant 2327211 and NSF CISE grant 2403317. E.F. is funded by NIAID/NIH grant number T32 AI155387. N.S.J. is funded by the Mark Foundation Emerging Leader Award.

References

[1] D. Grün, L. Kester, and A. van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nat. Methods*, vol. 11, pp. 637–640, June 2014. 1, 9

[2] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nat. Methods*, vol. 11, pp. 740–742, July 2014. 1, 9

[3] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018. 1, 2

[4] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 53–94, July 2006. 1, 2, 3, 24

[5] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell RNA sequencing data," *Genome Biol.*, vol. 18, Dec. 2017. 1, 4, 24, 25

[6] Data from: KLF2 maintains lineage fidelity and suppresses CD8 T cell exhaustion during acute LCMV infection (LCMV DSM scRNA data and ATAC-seq) (Dryad, 2024). https://doi.org/10.5061/dryad.dv41ns27h. 1, 5, 25

[7] M. Sade-Feldman, K. Yizhak, S. L. Bjorgaard, J. P. Ray, C. G. de Boer, R. W. Jenkins, D. J. Lieb, J. H. Chen, D. T. Frederick, M. Barzily-Rokni, S. S. Freeman, A. Reuben, P. J. Hoover, A.-C. Villani, E. Ivanova, A. Portell, P. H. Lizotte, A. R. Aref, J.-P. Eliane, M. R. Hammond, H. Vitzthum, S. M. Blackmon, B. Li, V. Gopalakrishnan, S. M. Reddy, Z. A. Cooper, C. P. Paweletz, D. A. Barbie, A. Stemmer-Rachamimov, K. T. Flaherty, J. A. Wargo, G. M. Boland, R. J. Sullivan, G. Getz, and N. Hacohen, "Defining T cell states associated with response to checkpoint immunotherapy in melanoma," *Cell*, vol. 175, no. 4, pp. 998–1013.e20, Nov. 2018. 1, 21, 25

[8] V1 Human Lymph Node, Spatial Gene Expression Dataset by Space Ranger 1.1.0, 10x Genomics, (2023, August 29). https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Human_Lymph_Node. 1, 5, 25

[9] S. Zhang, L. Yuan, L. Danilova, G. Mo, Q. Zhu, A. Deshpande, A. T. F. Bell, J. Elisseeff, A. S. Popel, R. A. Anders, *et al.*, "Spatial transcriptomics analysis of neoadjuvant cabozantinib and nivolumab in advanced hepatocellular carcinoma identifies independent mechanisms of resistance and recurrence," *Genome Medicine*, vol. 15, no. 1, p. 72, 2023. 1, 5, 25

[10] A. Venkat, S. Leone, S. E. Youlten, E. Fagerberg, J. Attanasio, N. S. Joshi, M. Perlmutter, and S. Krishnaswamy, "Mapping the gene space at single-cell resolution with gene signal pattern analysis," *Nature Computational Science*, vol. 4, no. 12, pp. 955–977, Dec. 2024. 1

[11] K. R. Moon, J. S. Stanley, III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy, "Manifold learning-based methods for analyzing single-cell RNA-sequencing data," *Curr. Opin. Syst. Biol.*, vol. 7, pp. 36–46, Feb. 2018. 2, 12

[12] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, July 2006. 2

[13] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999. 2

[14] M. Perlmutter, A. Tong, F. Gao, G. Wolf, and M. Hirn, "Understanding graph neural networks with generalized geometric scattering transforms," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 4, pp. 873–898, 2023. 2, 12

[15] J. Chew, M. Hirn, S. Krishnaswamy, D. Needell, M. Perlmutter, H. Steach, S. Viswanath, and H.-T. Wu, "Geometric scattering on measure spaces," *Applied and Computational Harmonic Analysis*, p. 101635, 2024. 2

[16] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, "Recovering gene interactions from single-cell data using data diffusion," *Cell*, vol. 174, pp. 716–729.e27, July 2018. 2, 9, 10

[17] R. S. Hoekzema, L. Marsh, O. Sumray, T. M. Carroll, X. Lu, H. M. Byrne, and H. A. Harrington, "Multiscale methods for signal selection in single-cell data," *Entropy (Basel)*, vol. 24, p. 1116, Aug. 2022. 2, 10, 24

[18] A. Y. Tong, G. Huguet, A. Natik, K. Macdonald, M. Kuchroo, R. Coifman, G. Wolf, and S. Krishnaswamy, "Diffusion earth mover's distance and distribution embeddings," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 10336–10346, PMLR, 18–24 Jul 2021. 2, 3, 10, 11, 14, 24

[19] S. Leone, A. Venkat, G. Huguet, A. Tong, G. Wolf, and S. Krishnaswamy, "Graph fourier MMD for signals on graphs," *SAMPTA*, 2023. 2, 10

[20] Y. Choi, R. Li, and G. Quon, "siVAE: interpretable deep generative models for single-cell transcriptomes," *Genome Biol.*, vol. 24, p. 29, Feb. 2023. 2, 10

[21] H. Chen, J. Ryu, M. E. Vinyard, A. Lerer, and L. Pinello, "SIMBA: single-cell embedding along with features," *Nat. Methods*, May 2023. 2, 11

[22] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016. 3

[23] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *CoRR*, vol. abs/1611.07308, 2016. 3

[24] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, "Visualizing structure and transitions in high-dimensional biological data," *Nat. Biotechnol.*, vol. 37, pp. 1482–1492, Dec. 2019. 3, 13

[25] A. Tong, G. Huguet, D. Shung, A. Natik, M. Kuchroo, G. Lajoie, G. Wolf, and S. Krishnaswamy, "Embedding signals on graphs with unbalanced diffusion earth mover's distance," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5647–5651, 2022. 3, 11

[26] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell RNA-seq data," *Nat. Rev. Genet.*, vol. 20, pp. 273–282, May 2019. 4

[27] A. Vandenbon and D. Diez, "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data," *Nat. Commun.*, vol. 11, p. 4318, Aug. 2020. 4

[28] M. Kuchroo, A. Godavarthi, A. Tong, G. Wolf, and S. Krishnaswamy, "Multimodal data visualization and denoising with integrated diffusion,'IEE Machine Learning for Signal Processing' Oct 2021. 4, 5

[29] M. Kuchroo, D. F. Miyagishima, H. R. Steach, A. Godavarthi, Y. Takeo, P. Q. Duy, T. Barak, E. Z. Erson-Omay, S. Youlten, K. Mishra-Gorur, J. Moliterno, D. McGuone, M. Günel, and S. Krishnaswamy, "spARC recovers human glioma spatial signaling networks with graph filtering," *bioRxiv*, Aug. 2022. 4, 5

[30] J. R. Giles, S. F. Ngiow, S. Manne, A. E. Baxter, O. Khan, P. Wang, R. Staupe, M. S. Abdel-Hakeem, H. Huang, D. Mathew, M. M. Painter, J. E. Wu, Y. J. Huang, R. R. Goel, P. K. Yan, G. C. Karakousis, X. Xu, T. C. Mitchell, A. C. Huang, and E. J. Wherry, "Shared and distinct biological circuits in effector, memory and exhausted CD8+ T cells revealed by temporal

single-cell transcriptomics and epigenetics," *Nat. Immunol.*, vol. 23, pp. 1600–1613, Nov. 2022. 5

[31] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, Oct. 2014. 5, 20

[32] T. Wu, Y. Ji, E. A. Moseman, H. C. Xu, M. Manglani, M. Kirby, S. M. Anderson, R. Handon, E. Kenyon, A. Elkahloun, W. Wu, P. A. Lang, L. Gattinoni, D. B. McGavern, and P. L. Schwartzberg, "The TCF1-Bcl6 axis counteracts type I interferon to repress exhaustion and maintain T cell stemness," *Sci. Immunol.*, vol. 1, pp. eaai8593–eaai8593, Dec. 2016. 5

[33] F. McNab, K. Mayer-Barber, A. Sher, A. Wack, and A. O'Garra, "Type I interferons in infectious disease," *Nat. Rev. Immunol.*, vol. 15, pp. 87–103, Feb. 2015. 5

[34] Z. Xie, A. Bailey, M. V. Kuleshov, D. J. B. Clarke, J. E. Evangelista, S. L. Jenkins, A. Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, M. Jeon, and A. Ma'ayan, "Gene set knowledge discovery with enrichr," *Curr. Protoc.*, vol. 1, p. e90, Mar. 2021. 5, 20

[35] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal. graph2vec: Learning Distributed Representations of Graphs. *arXiv*, 2017. Available at: https://arxiv.org/abs/1707.05005 5

[36] N. Brugnone, A. Gonopolskiy, M. W. Moyle, M. Kuchroo, D. van Dijk, K. R. Moon, D. Colon-Ramos, G. Wolf, M. J. Hirn, and S. Krishnaswamy, "Coarse graining of data via inhomogeneous diffusion condensation," in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2019. 13

[37] M. Kuchroo, J. Huang, P. Wong, J.-C. Grenier, D. Shung, A. Tong, C. Lucas, J. Klein, D. B. Burkhardt, S. Gigante, A. Godavarthi, B. Rieck, B. Israelow, M. Simonov, T. Mao, J. E. Oh, J. Silva, T. Takahashi, C. D. Odio, A. Casanovas-Massana, J. Fournier, Yale IMPACT Team, S. Farhadian, C. S. Dela Cruz, A. I. Ko, M. J. Hirn, F. P. Wilson, J. G. Hussin, G. Wolf, A. Iwasaki, and S. Krishnaswamy, "Multiscale PHATE identifies multimodal signatures of COVID-19," *Nat. Biotechnol.*, vol. 40, pp. 681–691, May 2022. 13

[38] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, Feb. 2018. 17, 25

[39] D. Kotliar, A. Veres, M. A. Nagy, S. Tabrizi, E. Hodis, D. A. Melton, and P. C. Sabeti, "Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq," *Elife*, vol. 8, July 2019. 20

[40] R. Huang, I. Grishagin, Y. Wang, T. Zhao, J. Greene, J. C. Obenauer, D. Ngan, D.-T. Nguyen, R. Guha, A. Jadhav, N. Southall, A. Simeonov, and C. P. Austin, "The NCATS BioPlanet - an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics," *Front. Pharmacol.*, vol. 10, p. 445, Apr. 2019. 20

[41] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, pp. 2498–2504, Nov. 2003. 20

[42] V. Svensson, S. A. Teichmann, and O. Stegle, "SpatialDE: identification of spatially variable genes," *Nature Methods*, vol. 15, no. 5, pp. 343–346, Mar. 2018. 23

[43] C. Grasso, J. E. G. Roet, C. G. de Graça, J. F. Semmelink, M. de Kok, E. Remmerswaal, A. Jongejan, P. D. Moerland, R. E. Mebius, and L. G. M. van Baarsen, "Identification and mapping of human lymph node stromal cell subsets by combining single-cell RNA sequencing with spatial transcriptomics," *European Journal of Immunology*, vol. 55, no. 6, Jun. 2025. 23

[44] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, and O. A. Bayraktar, "Cell2location maps fine-grained cell types in spatial transcriptomics," *Nature Biotechnology*, vol. 40, no. 5, pp. 661–671, Jan. 2022. 23

[45] D. Türei, A. Valdeolivas, L. Gul, N. Palacio-Escat, M. Klein, O. Ivanova, M. Ölbei, A. Gábor, F. Theis, D. Módos, T. Korcsmáros, and J. Saez-Rodriguez, "Integrated intra- and intercellular signaling knowledge for multicellular omics analysis," *Molecular Systems Biology*, vol. 17, no. 3, p. e9923, Mar. 2021. 23

A Appendix

A.1 Defining desired gene embedding properties

Our goal is to obtain a low-dimensional representation of each gene which preserves the inherent structure of the gene space with respect to the cellular manifold. In particular, we seek a reduced dimensional map $\Theta : \mathbb{R}^n \to \mathbb{R}^d$, $d \ll n$, which satisfies the desired properties enumerated below.

1. **Preserving local and global distances between signals**: A good gene embedding should produce similar representations of genes $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$ (viewed as rows of $\mathbf{X}$) if they have similar measurement profiles. In order to ensure that we capture meaningful information, we aim to preserve distances based on the geometry of the underlying cell-cell graph $\mathcal{G}_{cell}$, rather than the naive pointwise distance, between gene signals.
2. **Noise robustness**: Addressing biological noise, such as cell-to-cell variation, and technical noise, such as dropout, have been longstanding concerns in single-cell analysis and best practices [1, 2, 16]. Due to variability in noise between genes with different expression levels [1], noise robustness is especially relevant for constructing gene embeddings. We thus seek a representation Θ such that $\|\Theta(\mathbf{X}_{i_1}) - \Theta(\mathbf{X}_{i_2})\|_2 \approx \|\Theta(\mathbf{X}_{i_1} + \epsilon_{i_1}) - \Theta(\mathbf{X}_{i_2} + \epsilon_{i_2})\|_2$ where ϵ_{i_1} and ϵ_{i_2} are measurement noise associated with genes $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$.
3. **Flexibility to downstream tasks**: Finally, we want to ensure our embedding Θ is flexibly defined for training on various additional tasks, whether concurrently with the learned embedding or downstream of the embedding.

A.2 Related work and baselines

Here, we describe in detail related work and comparisons for our experiments with GSPA:

- Raw measurements approach embeds $\mathbf{X}$.
- $\text{GAE}_{\text{no-att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$, representing a gene-gene similarity graph based on the scRNA-seq data.
- $\text{GAE}_{\text{att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$.
- Node2Vec($\mathcal{G}_{gene}$) embeds $\mathcal{G}_{gene}$.
- MAGIC($\mathbf{X}$) embeds $\mathbf{X}$ after denoising with $\mathcal{G}_{cell}$.
- DiffusionEMD($\mathbf{X}$, $\mathcal{G}_{cell}$) embeds $\mathbf{X}$ via optimal transport on $\mathcal{G}_{cell}$, representing a cell-cell similarity graph based on the scRNA-seq data.
- GFMMD($\mathbf{X}$, $\mathcal{G}_{cell}$) embeds $\mathbf{X}$ via MMD on $\mathcal{G}_{cell}$.
- Eigenscore($\mathbf{X}$, $\mathcal{G}_{cell}$) embeds $\mathbf{X}$ via alignment to Laplacian eigenvectors of $\mathcal{G}_{cell}$.
- SIMBA co-embeds $\mathbf{X}$ and $\mathbf{X}^T$ via heterogeneous graph embedding.
- siVAE co-embeds $\mathbf{X}$ and $\mathbf{X}^T$ via jointly trained cell-wise and feature-wise VAEs.

A.2.1 Direct embedding of gene expression measurements

The simplest and most intuitive approach to map the gene space is with the original measurements. $\mathbf{X}$ consists of values where each cell is measured as a vector of gene expression counts, so we can consider the case where the genes are observations, and each gene is measured as a vector of expression counts in each cell. We use autoencoder $D \circ E$ to reduce the dimensionality, where $\mathbf{X} \approx D(E(\mathbf{X}))$ and $E(\mathbf{X})$ is the embedding.

A.2.2 Embedding constructed gene-gene graph

Another approach is to construct a gene-gene k-NN graph $\mathcal{G}_{gene} = (V_{gene}, E_{gene})$ from $\mathbf{X}$, where each node in V_{gene} corresponds to a gene and each edge E_{ij} in E describes the similarity between gene i and gene j based on Euclidean distance. We can then leverage graph representation learning to propagate information between gene-gene relationships and learn node embeddings. We test one shallow embedding Node2Vec($\mathcal{G}_{gene}$), and two graph autoencoder embeddings. The graph autoencoder $D_{no-att} \circ E_{no-att}$ consists of graph convolutional layers, where $\mathcal{G}_{gene} \approx D_{no-att}(E_{no-att}(\mathcal{G}_{gene}))$. The graph autoencoder $D_{att} \circ E_{att}$ consists of graph attention layers, where $\mathcal{G}_{gene} \approx D_{att}(E_{att}(\mathcal{G}_{gene}))$. $E_{no-att}(\mathcal{G}_{gene})$ and $E_{att}(\mathcal{G}_{gene})$ correspond to the embeddings without and with attention, respectively.

A.2.3 Imputing gene signals with cell-cell graph

The above methods do not use information from the cell-cell graph for the computation of gene representations. Based on our desired properties, we hypothesized that incorporating cellular affinities would enable the comparison of non-overlapping gene signals across local and global distances on the cellular manifold.

First, we compare against MAGIC [16], which imputes missing gene expression via data diffusion. MAGIC calculates a diffusion operator $\mathbf{P}$ powered to t, and left-multiplies $\mathbf{P}^t$ to $\mathbf{X}^T$ as a low-pass filter. For comparison, we left-multiply $\mathbf{X}$ to $\mathbf{P}^t$, which practically denoises gene signals and performs comparatively to MAGIC (data not shown). We then employ an autoencoder $D \circ E$, where $\mathbf{X}\mathbf{P}^t \approx D(E(\mathbf{X}\mathbf{P}^t))$ and $E(\mathbf{X}\mathbf{P}^t)$ is the embedding.

A.2.4 Optimal transport distances between gene signals

Due to the relationship between GSPA and Wasserstein distance (i.e. optimal transport), we compare GSPA against approaches for fast optimal transport that have been developed and used for gene signals on the cellular graph.

Diffusion Earth Mover's Distance (DiffusionEMD) [18] computes optimal transport based on multiscale diffusion kernels. Between two genes $i_1, i_2 \in \mathbf{X}$, $\text{DiffusionEMD}_{\beta,J}(i_1, i_2) := \sum_{j=0}^{J} \left\| T_{\beta,j}(i_1) - T_{\beta,j}(i_2) \right\|_1$, where $0 < \beta < 1/2$ is a meta-parameter used to balance long- and short-range distances and J is the maximum scale considered here. $T_{\beta,j}(\mathbf{X}_i) := \begin{cases} 2^{-(J-j-1)\beta}\left(\mu_i^{(2^{j+1})} - \mu_i^{(2^j)}\right) & j < J \\ \mu_i^{(2^J)} & j = J \end{cases}$, where $\mu_i^{(t)} := \frac{1}{n_i}\mathbf{P}^t\mathbf{1}_{\mathbf{X}_i}$ is a kernel density estimate over $\mathcal{G}_{cell}$. Graph Fourier Mean Maximum Discrepancy (GFMMD) [19] is defined via an optimal witness function that is smooth on the graph and maximizes the difference in expectation between the pair of gene distributions. $\text{GFMMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) := \max_{f: f^T\mathbf{L}f \leq 1} \mathbb{E}_{\mathbf{X}_{i_1}}(f) - \mathbb{E}_{\mathbf{X}_{i_2}}(f)$, holding for any construction of a positive semi-definite Laplacian matrix $\mathbf{L}$ and chosen threshold $T = 1$.

For these approaches, multiscale signal features $\widehat{\mathbf{X}}$ are computed prior to distance calculation. We reduce the dimensionality of these features via an autoencoder $D \circ E$, where $\widehat{\mathbf{X}} \approx D(E(\widehat{\mathbf{X}}))$ and $E(\widehat{\mathbf{X}})$ is the embedding.

A.2.5 Computing eigenscores

Eigenscores were proposed as a topologically motivated mathematical method for feature selection, and they were also shown to be useful for mapping the gene space to distinguish cell types [17]. Eigenscores rank signals or genes based on their alignment to low-frequency patterns in the data, identified through spectral decomposition of the graph Laplacian. Specifically, given the first r left eigenvectors of the normalized Laplacian (where $r \ll n$ to preserve low-frequency patterning), $\text{Eigenscore}(i) := \text{concat}(\frac{\langle \mathbf{D}^{1/2}\mathbf{X}_i, e_1 \rangle}{\|\mathbf{D}^{1/2}\mathbf{X}_i\|}, \frac{\langle \mathbf{D}^{1/2}\mathbf{X}_i, e_2 \rangle}{\|\mathbf{D}^{1/2}\mathbf{X}_i\|}, ..., \frac{\langle \mathbf{D}^{1/2}\mathbf{X}_i, e_r \rangle}{\|\mathbf{D}^{1/2}\mathbf{X}_i\|})$. We let $\text{Eigenscore}(\mathbf{X})$, of shape m x r represent the eigenscores for each gene i in $\mathbf{X}$. We finally reduce the dimensionality for gene space mapping via an autoencoder $D \circ E$, where $\text{Eigenscore}(\mathbf{X}) \approx D(E(\text{Eigenscore}(\mathbf{X})))$ and $E(\text{Eigenscore}(\mathbf{X}))$ is the embedding.

A.2.6 Co-embedding of cells and genes

Finally, recent approaches incorporate cell-cell affinities through simultaneously learning embeddings for cells and genes. This methodology has the benefit of *learning* the pairwise similarities between cells, rather than constructing the cell-cell graph *a priori*, and training this module in tandem with gene-gene similarity training. siVAE [20] is a neural network consisting of cell-wise and gene-wise encoder-decoders. The cell-wise encoder takes each cell's measurement across all features and maps cell embeddings similarly to a classical VAE, which computes an approximate posterior distribution over the location of the cell. The gene-wise encoder takes a gene's measurement across all cells and maps gene embeddings. The decoders of both VAEs combine to output the expresion level of

each feature in each particular cell, ensuring that each mapping has semantic structure. SIMBA [21] constructs a heterogeneous graph, where the nodes are cells and genes, and edge type are determined through expression level. SIMBA first bins the continuous gene expression values into a discrete distribution that preserves the shape of the original distribution, then encodes different bins as different relation types. A node embedding for each node in the graph is then learnt via stochastic gradient descent optimization of a link prediction objective. For both procedures, we evaluated only the gene space embeddings in our comparisons.

A.3 Achieving desired gene embedding properties with GSPA

A.3.1 Distance preservation

Our first desired property is an embedding that preserves distances (quantified in a manner informed by the geometry of the cellular state space). Theorem A.1, defined below, shows that GSPA is able to achieve this goal since it guarantees we will have $GSPA(\mathbf{X}_i) \approx GSPA(\mathbf{X}_j)$ whenever $\mathbf{X}_i$ is close to $\mathbf{X}_j$ with respect to the Unbalanced Diffusion Earth Mover's Distance (UDEMD). This distance, a variant of traditional earth mover's distance (EMD), views the signals $\mathbf{X}_i$ (when properly normalized) as probability distributions on the graph.

Earth Mover's Distances (EMDs), alternatively referred to as Monge-Kantorovich or Wasserstein Distances, are a useful way of computing the distances between two signals. In the case where the signals correspond to probability distributions μ and ν, these distances can be thought of as the "cost" of moving a collection of points distributed according to μ to a collection of points distributed according to ν, where the cost of moving each point depends on the distance it must travel (defined with respect to some ground distance). In [25] (see also [18]), it was shown that the Wasserstein distance (with a truncated geodesic distance as ground distance) could be approximated by the Unbalanced Diffusion Earth Mover's Distance (UDEMD) defined below. Here, we will show that the metric induced by our wavelets is continuous with respect to this UDEMD, i.e., $\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \lesssim \text{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$.

In [25], the UDEMD [25] between two signals (genes) $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}$ is defined as

$$\text{UDEMD}_{\beta,J}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) := \sum_{j=0}^{J} \left\|T_{\beta,k}\left(\mathbf{X}_{i_1}\right) - T_{\beta,k}\left(\mathbf{X}_{i_2}\right)\right\|_1, .$$

where $0 < \beta < 1/2$ is a meta-parameter used to balance long- and short-range distances and J is the maximum scale considered here, and $T_{\beta,j}$ is defined by

$$T_{\beta,j}\left(\mathbf{X}_i\right) := 2^{-(J-j)\beta}\left(\mu_i^{\left(2^j\right)} - \mu_i^{\left(2^{j-1}\right)}\right), \quad \mu_i^{(t)} := \mathbf{P}^t\mathbf{X}_i,$$

for $1 \le j \le J$ and $T_{\beta,0}(\mathbf{X}_i) = 2^{-J\beta}(\mathbf{P} - \mathbf{I})\mathbf{X}_i$.

Theorem A.1 *For $0 < \beta < 1/2$, the diffusion wavelet transform $\mathbf{W}$ (with maximal scale J) is Lipschitz continuous with respect to $\text{UDEMD}_{\beta,J}$, that is there exists a constant $C > 0$ (depending on β and J and the ratio between the largest and smallest vertex degrees) such that*

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \le C \cdot \ UDEMD_{\beta,J}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}).$$

Let $\mathbf{X}_{i_1} \neq \mathbf{X}_{i_2}$. (The inequality holds trivially in the case where $\mathbf{X}_{i_1} = \mathbf{X}_{i_2}$.) We may compute

$$\begin{aligned}
\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2^2 &= \|\mathbf{W}^T\mathbf{X}_{i_1}^T - \mathbf{W}^T\mathbf{X}_{i_2}^T\|_2^2 \\
&= \sum_{v=1}^{n}\sum_{j=0}^{J} |\delta_v^T\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \delta_v^T\mathbf{\Psi}_j\mathbf{X}_{i_2}^T|^2 \\
&= \sum_{j=0}^{J} \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_2^2 \\
&\leq \sum_{j=0}^{J} \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_2 \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_1 \\
&\leq C \max_{0\leq j\leq J} \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_2 \sum_{j=0}^{J} 2^{-(J-j)\beta} \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_1 \\
&= C \max_{0\leq j\leq J} \|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_2 \ \mathrm{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}),
\end{aligned}$$

where C is a constant depending on J and β. It follows from Proposition 2.2 of [14] that

$$\|\mathbf{\Psi}_j\mathbf{X}_{i_1}^T - \mathbf{\Psi}_j\mathbf{X}_{i_2}^T\|_2 \leq C\|\mathbf{X}_{i_1}^T - \mathbf{X}_{i_2}^T\|_2,$$

where C is a constant depending only on the ratio between the maximal vertex degree and minimal vertex degree. ([14] considers the wavelets on a weighted inner product space where vertices are weighted by degree. Transferring this result to the unweighted ℓ^2 space induces dependence on the ratio between the maximal and minimal degrees.) Therefore, we have

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2^2 \leq C \cdot \mathrm{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})\|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\|_2,$$

which in turn implies

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \leq C \cdot \mathrm{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) \frac{\|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\|_2}{\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2}.$$

The lower bound in Proposition 2.2 of [14] implies that $\frac{\|\mathbf{X}_{i_1}^T - \mathbf{X}_{i_2}^T\|_2}{\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2}$ is bounded above by a constant (depending on the ratio between the maximal and minimal vertex degrees). Therefore, we have

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \leq C \cdot \mathrm{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$$

as desired.

A.3.2 Noise robustness

Robustness to biological and technical noise is a key feature of diffusion-based single-cell analysis approaches [11]. Note that raising the diffusion operator $\mathbf{P}$ to the power t is equivalent to powering the eigenvalues of the diffusion operator by t, i.e., $\mathbf{P} = \mathbf{\Sigma}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$, where the columns of $\mathbf{\Sigma}$ contain the (right) eigenvectors of $\mathbf{P}$ and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the corresponding eigenvalues. Thus, $\mathbf{P}^t = \mathbf{\Sigma}\mathbf{\Lambda}^t\mathbf{\Sigma}^{-1}$ and powering $\mathbf{P}$ effectively results in powering the eigenvalues contained in $\mathbf{\Lambda}$. The eigenvectors are decreasingly ordered by their "frequency", a notion of how rapidly a signal oscillates over the graph. It is known that $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots$. Therefore, powering $\mathbf{P}$ preserves the lead eigenspace and suppresses the subsequent spaces by a factor of λ_i^t. Acting on a signal $\mathbf{X}_i$ by $\mathbf{P}^t$ preserves the portion of the signal aligned with the first eigenvector and depresses the portion of the signal corresponding to the other eigenvectors by a factor of λ_i^t. As t increases, the high-frequency (small eigenvalue) portion of the signal is suppressed. Naturally occurring signals tend to vary slowly and smoothly over the graph (and thus lie in the low-frequency eigenspaces), whereas noise is not related to the structure of the graph and therefore will often lie in the higher frequencies. In this manner, acting on the signal $\mathbf{X}_i$ by $\mathbf{P}^t$ has a denoising effect since it suppresses the high-frequency (noisy) portion of the signal. Therefore, we can restrict the dictionary to wavelets that decompose only the lower frequencies by initially multiplying each wavelet by $\mathbf{P}^t$. Additionally, we note that the distance preservation result in Theorem A.1 shows that the wavelet projection is continuous with respect to the UDEMD, which may be viewed as a form of noise robustness.

A.3.3 Flexibility to downstream tasks

We demonstrate flexibility through learning a low-dimensional representation generalizable for diverse downstream tasks, as represented in our case studies.

A.4 Runtime analysis and GSPA for large graphs

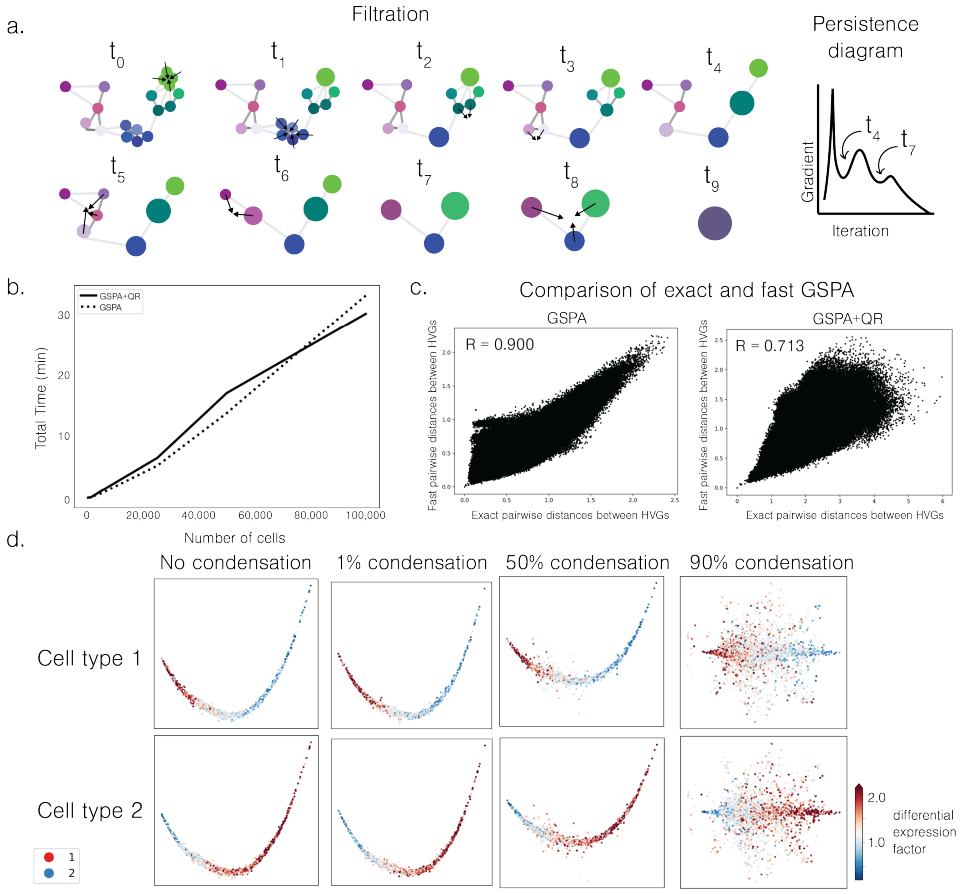

Figure A.1: Scalability and performance with coarse-grained cell-cell graph. a. Diagram of diffusion condensation and identification of persistent resolutions. b. Runtime of GSPA and GSPA+QR. c. Comparison of pairwise gene-gene distances between exact GSPA and fast GSPA embeddings. d. Experiment with simulated dataset comprising of 95% cell type 1 and 5% cell type 2. Gene embedding visualization colored by ground truth differential expression factor, i.e. how associated a gene is with each cell type, for varying degrees of condensation.

For large graphs, Gene Signal Pattern Analysis utilizes diffusion condensation, a coarse-graining process which iteratively condensing datapoints toward local centers of gravity and is shown to approximate heat diffusion over the time-varying manifold [36]. Over the condensation time, the original coordinate functions are smoothed by a cascade of diffusion operators, which adaptively removes high-frequency variations. At each iteration, points closer than a given threshold collapse to the same barycenter. This technique allows GSPA to summarize the underlying topology of the data manifold. We use a version of diffusion condensation designed for single-cell analysis, Multiscale PHATE [37], which uses the potential representation of datapoints from PHATE [24] as the initial features. By smoothly condensing nodes and choosing the resolution with persistent (i.e. stable) condensation, this process better preserves the graph topology, including subtle or rare patterns, and shows improvement over clustering-based approaches while remaining highly scalable [37] (Figure A.1a).

For graphs larger than threshold $n_{condense}$, we use Multiscale PHATE to iteratively condense datapoints to a small number of nodes. GSPA then filters for iterations with $n_{condense}$ or fewer nodes, where each node represents a condensation of one or more cells. Finally, GSPA selects the iteration with a node count closest to $n_{condense}$ to balance coarse- and fine-grained information. This represents a smaller cell-cell graph representing the same underlying manifold as the initial (larger) dataset, and GSPA computes a wavelet dictionary based on this graph. Then, gene signals are defined on the nodes of the condensed graph as the mean expression of all the cells in each node. By default, $n_{condense} = 10,000$ cells. Due to the smaller size of the graph, computation becomes much more tractable (100,000 cells in 33.17 minutes and 30.18 minutes with GSPA and GSPA+QR, respectively) with comparable results, where pairwise distances between genes from exact versus GSPA showed high correlation (R=0.900 for GSPA and R=0.713 for GSPA+QR) (Figure A.1b-c). We show that this condensation process preserves subtle patterns that exist in even a small number of cells. In a simulated dataset with two cell types, where cell type 1 comprises 95% of the cells and cell type 2 comprises 5%, GSPA gene embeddings derived with no condensation, 1%, 50% and 90% condensation all preserve the differential gene signature associated with each cell type (Figure A.1d).

A.5 Choosing the number of scales J

The number of scales for the wavelet dictionary J is defined as the log of the number of cells n based on the following lemma introduced by Tong et al. [18] and proven in the original work:

Lemma A.2 *There exists a $K = O(log|V|)$ such that $\mu_i^{(2^K)} \backsimeq \phi_0$ for every $i = 1 \ldots, n$, where ϕ_0 is the trivial eigenvector of $\mathbf{P}$ associated with the eigenvalue $\lambda_0 = 1$.*

This is based on the reasoning that if the Markov process converges in polynomial time with respect the number of nodes $|V|$, then one can ensure that beyond $O(log|V|)$, all density estimates would be indistinguishable from each other.

A.6 Computation of differential localization

Characterizing differentially expressed genes between clusters is not feasible for many biological systems. For example, for datasets that have trajectory-like structure, consist of subtypes within cell types, or do not organize into discrete populations, there is utility in identifying genes localized to particular areas of the cellular manifold without prior cell type identification. To this end, we naturally extend GSPA to a framework called *differential localization*. We calculate the specificity, termed gene localization score $l(i)$, of a given gene signal i by calculating the multiscale representation of a uniform signal $\mathbf{u}$ and computing the distance between this and each gene signal representation. Genes are then ranked, where those that are most differentially localized are farthest from the uniform signal representation.

The gene localization score, $l(i)$ for each gene $\mathbf{X}_i$, with normalized uniform signal $\mathbf{u} = \frac{1}{\sqrt{n}}\mathbf{1}$ and wavelet representation $\widehat{\mathbf{W}}$, is defined as:

$$l(i) := \|\mathbf{X}_i\widehat{\mathbf{W}} - u\widehat{\mathbf{W}}\|_2^2 \tag{A.1}$$

Genes with a high localization score are considered more relevant for describing cell-cell variation and can be used for feature selection or characterization of gene programs and networks without the underlying assumption of discrete clusters.

A.7 Extended simulated data experiments and robustness to normalization and graph construction

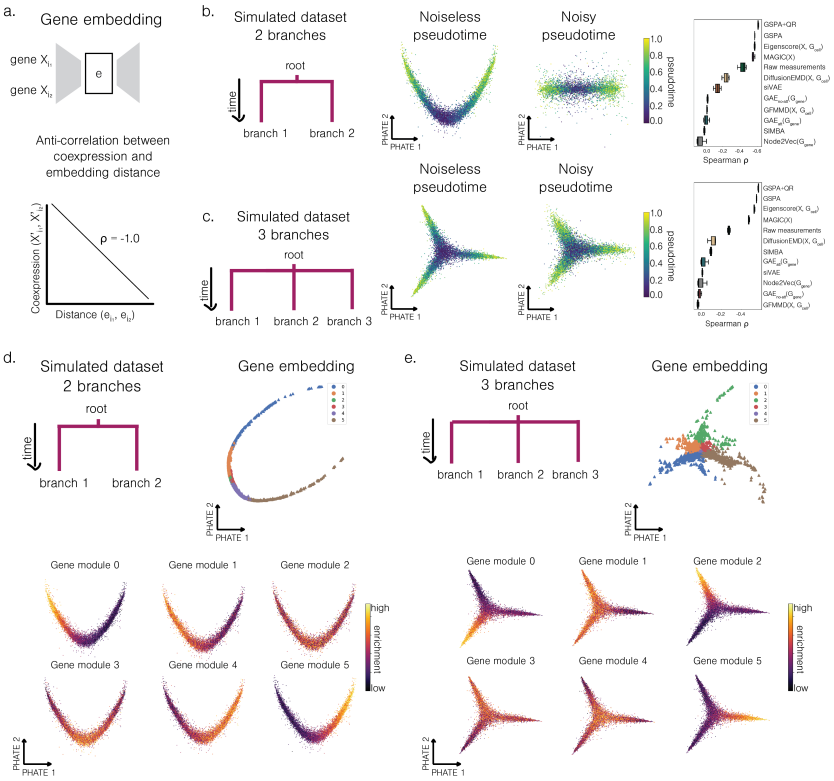

Figure A.2: GSPA preserves coexpression in two alternative single-cell simulations. a. Experimental setup. b. Simulated dataset with two branches schematic. PHATE embedding of cells from noiseless simulation and noisy simulation, colored by pseudotime. Spearman correlation evaluating performance for all comparisons across 3 runs. c. Simulated dataset with three branches schematic. PHATE embedding of cells from noiseless simulation and noisy simulation, colored by pseudotime. Spearman correlation evaluating performance for all comparisons. d. PHATE embedding of genes from two branch simulation, colored by gene module assignments. Cells colored by gene module enrichment score. e. PHATE embedding of genes from three branch simulation, colored by gene module assignments. Cells colored by gene module enrichment score.

For the coexpression experiment with a linear trajectory (Figure 2a) and two and three branches (Figure A.2), we generated simulated data, then defined signals as the gene features from the simulation experiment. Because of the simulation design, this meant we have both noisy $\mathbf{X}$ and noiseless $\mathbf{X}'$ versions of the same gene signals. This allows us to compute "ground truth" coexpression as the Spearman correlation between all noiseless pairs of genes. Given the large number of genes and the nature of biological data, the large majority of gene-gene pairs had a near-zero correlation. The correlation also was associated with the library size of the genes in the pair. Therefore, we stratified the labels based on correlation and the mean library size of the pair within each correlation bin. We learned unsupervised gene embeddings for all comparisons as described above, then, for an equal number of pairs per stratification bin, we computed the distance between gene embedding pairs and the anti-correlation with the true coexpression.

For the localization experiment with a linear trajectory (Figure 2b) and two and three branches (Figure A.4), we generated simulated data as previously described. However, instead of using the genes as signals, we designed signals with "ground truth" localization labels (Figure A.3). We intuited that more localized signals are not defined by where they are enriched in the trajectory, but rather by how spread out that enrichment is. Thus, we aimed to constrain the size of the region where each signal

could be defined, termed “window", where the window can be defined anywhere on the trajectory and is only defined by its size.

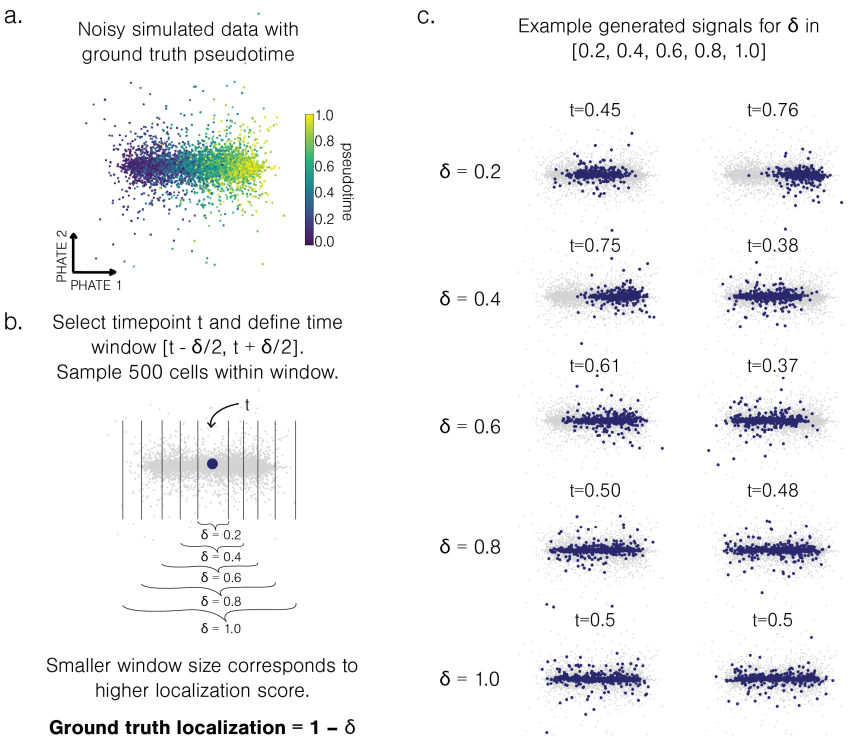

Figure A.3: Schematic of generation of signals for localization experiment. a. Noisy simulated data with pseudotime. b. Selection of windows of size δ, where ground truth localization is $1 - \delta$. c. Examples of generated signals of different δ.

To generate signals and associated localization scores with these properties, we used the ground truth pseudotime label (provided by Splatter) scaled to be between 0 and 1, and we defined window size δ. Then, we randomly selected a timepoint t between $[\delta/2, 1 - \delta/2]$ and defined a pseudotime window $[t - \delta/2, t + \delta/2]$. Next, we sampled 500 cells from all cells within this pseudotime window, and we let the signal equal 1 on these cells and 0 on all other cells (Figure A.3).

For each of five window sizes $\delta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, we generated 50 signals, resulting in 250 signals total. As smaller δ corresponds to a higher localization score, we defined the true localization score for each signal to be $1 - \delta$. This score is unrelated to where the signal is defined based on randomly selected t. Furthermore, all signals are defined on exactly 500 cells, so the localization score is not associated with the number of cells expressing the gene.

For GSPA+QR, GSPA, and MAGIC, computing signal localization involved projecting the uniform signal onto the cell representation/dictionary and calculating the distance between the projected uniform signal and all other projected signals. Eigenscore and GFMMD defined a version of this localization based on the L2 norm of their embeddings, so we evaluated localization using this measure. For DiffusionEMD, we learned a multiscale representation of the uniform signal, and we computed the distance to all other signals before dimensionality reduction. For the raw measurements, we took the distance of the uniform signal to all other signals before dimensionality reduction. For Node2Vec and the GAE approaches, we built a signal-signal graph with the uniform signal and embedded these graphs, then computed the L2 distance between the uniform embedding and the other signals. For SIMBA and siVAE, which learn a low-dimensional representation of the genes directly, we learned a low-dimensional embedding of the uniform signal and computed the distance to all other signals in this latent space.

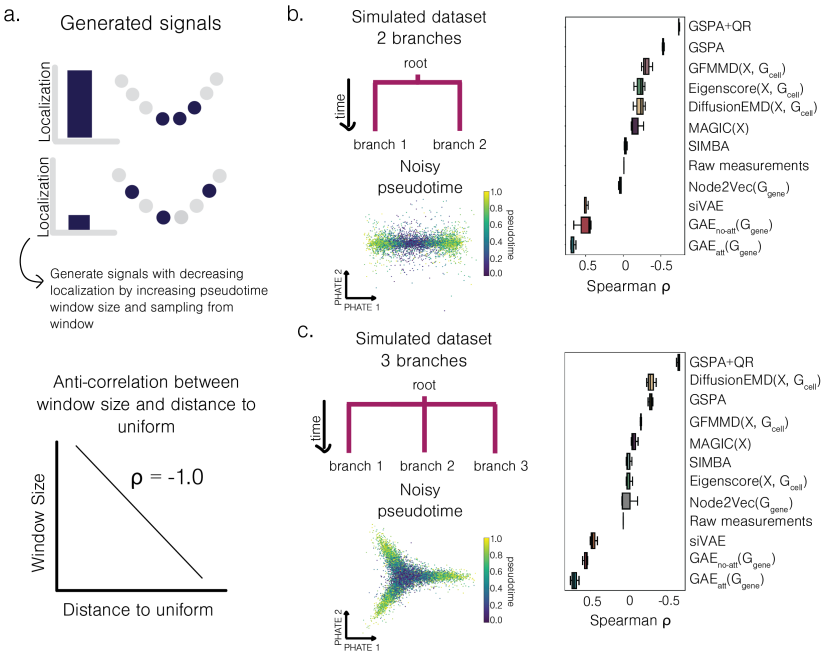

Figure A.4: GSPA captures localized genes in alternative single-cell simulations. a. Diagram of generated signals based on pseudotime window and anti-correlation between window size and localization. b. Two branch noisy simulated dataset, visualized with PHATE and colored by pseudotime. Spearman correlation evaluating performance for all comparisons. c. Three branch noisy simulated dataset, visualized with PHATE and colored by pseudotime. Spearman correlation evaluating performance for all comparisons across 3 runs.

To evaluate robustness to steps to process the cellular measurements before mapping the gene space, we ran our coexpression experiment and localization experiment for all comparisons over each combination of the following: two runs, two single-cell dataset transformations (`log` and `sqrt`), four choices for k in construction of k nearest neighbors graph (5, 15, 25, 50), and three choices for construction of nearest neighbors graph (k nearest neighbors (kNN), shared nearest neighbors (SNN), and construction with an adaptive α-decaying kernel). Together, this resulted in 48 runs for each method (Figure A.5a).

On average across all hyperparameters and preprocessing choices, GSPA and GSPA+QR outperformed all other approaches (Figure A.5b). Furthermore, despite potential sensitivity to graph construction, approaches that leveraged the cell-cell graph to calculate gene-gene relationships outranked approaches that used pointwise gene measurements on both experiments (Figure A.5c). For the coexpression experiments, approaches with the cell-cell graph had an average rank of 2.929, and approaches without the cell-cell graph had an average rank of 8.071. For the localization experiments, approaches with the cell-cell graph had an average rank of 2.686, and approaches without the cell-cell graph had an average rank of 8.314. This result reinforces the desired distance preservation and noise robustness properties garnered from using the cell-cell graph and further supports our assertion that considering genes as signals on the cell-cell graph can improve analysis of gene-gene relationships. Additionally, as most single-cell sequencing analysis tools and pipelines construct a cell-cell graph, including for visualization, clustering, and trajectory inference [38], using the same graph can ensure consistent biological analysis with GSPA.

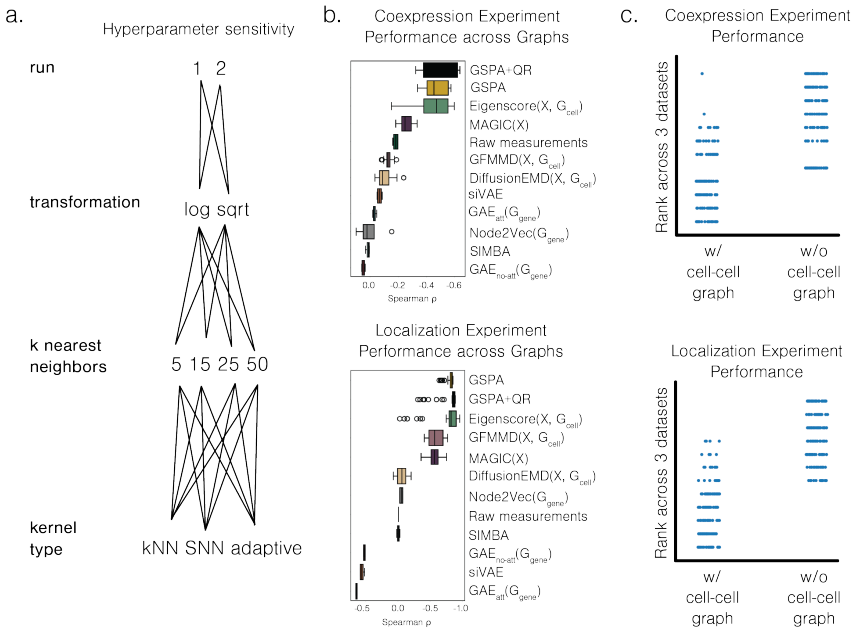

Figure A.5: GSPA robust to transformation and graph construction. a. Schematic of grid search of 2 transformations, 4 kNN choices, 3 kernels, and 2 replicates (48 runs total). b. Coexpression and localization experiment performance across all runs. c. Comparison of performance rank of methods that use cell-cell graph versus without cell-cell graph.

A.8 Evaluating autoencoder contribution to GSPA

The median performance on the three datasets and two benchmarks is as follows:

	linear		**2 branches**		**3 branches**	
model	**coex. (↓)**	**loc. (↓)**	**coex. (↓)**	**loc. (↓)**	**coex. (↓)**	**loc. (↓)**
% improvement over best baseline	7.33%	10.55%	7.05%	151.02%	8.64%	134.88%
% improvement over best non-graph baseline	230.65%	12471.40%	37.02%	2852%	105.82%	4225%
GSPA+QR	**-0.615** ± 0.003	**-0.880** ± 0.004	**-0.607** ± 0.006	**-0.738** ± 0.006	**-0.566** ± 0.004	**-0.660** ± 0.022
GSPA	-0.539 ± 0.009	-0.843 ± 0.008	-0.569 ± 0.001	-0.529 ± 0.013	-0.547 ± 0.002	-0.294 ± 0.032
Eigenscore	-0.573 ± 0.010	-0.796 ± 0.014	-0.567 ± 0.006	-0.232 ± 0.074	-0.521 ± 0.005	0.020 ± 0.043
MAGIC	-0.296 ± 0.016	-0.678 ± 0.106	-0.541 ± 0.017	-0.122 ± 0.090	-0.467 ± 0.008	-0.051 ± 0.047
DiffusionEMD	-0.192 ± 0.094	-0.187 ± 0.046	-0.247 ± 0.043	-0.233 ± 0.080	-0.141 ± 0.044	-0.281 ± 0.063
Raw	-0.186 ± 0.010	-0.007 ± 0.000	-0.443 ± 0.052	-0.007 ± 0.000	-0.275 ± 0.014	0.071 ± 0.000
GFMMD	-0.122 ± 0.017	-0.603 ± 0.055	-0.019 ± 0.004	-0.294 ± 0.075	0.031 ± 0.009	-0.157 ± 0.007
siVAE	-0.059 ± 0.003	0.518 ± 0.024	-0.144 ± 0.045	0.516 ± 0.028	-0.015 ± 0.005	0.475 ± 0.047
Node2Vec	-0.113 ± 0.138	0.006 ± 0.060	0.081 ± 0.060	0.032 ± 0.019	0.019 ± 0.050	0.083 ± 0.117
SIMBA	-0.068 ± 0.021	0.014 ± 0.019	0.009 ± 0.008	-0.025 ± 0.019	-0.100 ± 0.010	0.016 ± 0.039
GAE (att)	-0.037 ± 0.025	0.674 ± 0.053	-0.005 ± 0.035	0.692 ± 0.037	-0.010 ± 0.040	0.716 ± 0.057
GAE (no-att)	0.045 ± 0.010	0.510 ± 0.046	-0.026 ± 0.007	0.455 ± 0.127	0.017 ± 0.020	0.556 ± 0.034

Table 1: Median performance ± 1 standard deviation across 3 runs for coexpression and localization experiments (visualized in figures). Top performance bolded, second best underlined.

To assess how the autoencoder component of GSPA contributes to performance gains, we repeat our coexpression experiment on three simulated benchmarks (linear trajectory, two branch trajectory, and three branch trajectory) with two ablations: GSPA+QR without the autoencoder (GSPA +QR -AE) and GSPA without the autoencoder (GSPA -QR -AE). We replace the autoencoder with SVD to maintain the same low dimensionality across all comparisons. We report the median correlation

between gene-gene coexpression and gene-gene distance (where -1.0 is optimal performance) over three seeds.

model	linear coex. (↓)	2 branches coex. (↓)	3 branches coex. (↓)
GSPA +QR +AE	**-0.615** ± 0.003	**-0.607** ± 0.006	**-0.566** ± 0.004
GSPA +QR -AE	-0.562 ± 0.001	-0.566 ± 0.003	-0.554 ± 0.001
GSPA -QR +AE	-0.539 ± 0.009	-0.569 ± 0.001	-0.547 ± 0.002
GSPA -QR -AE	-0.505 ± 0.003	-0.556 ± 0.003	-0.547 ± 0.002

Table 2: QR factorization and autoencoder (AE) ablation study of coexpression experiment.

GSPA +QR +AE performs the best on all three benchmarks. The no-AE versions of both methods overall perform worse than the +AE versions, and the +QR versions overall outperform the -QR versions. These results demonstrate that both QR factorization and the autoencoder contribute to the strong performance of GSPA. Additionally, all of the GSPA variants outperform "Raw measurements", for which we take the gene expression matrix and reduce dimensionality also with matrix factorization and an autoencoder, but without taking into the cell-cell graph structure. We note that the localization experiment computes localization using the wavelet dictionary, not the autoencoder, so performance is the same on this experiment with and without it.

A.9 Extended CD8+ T cell experiments

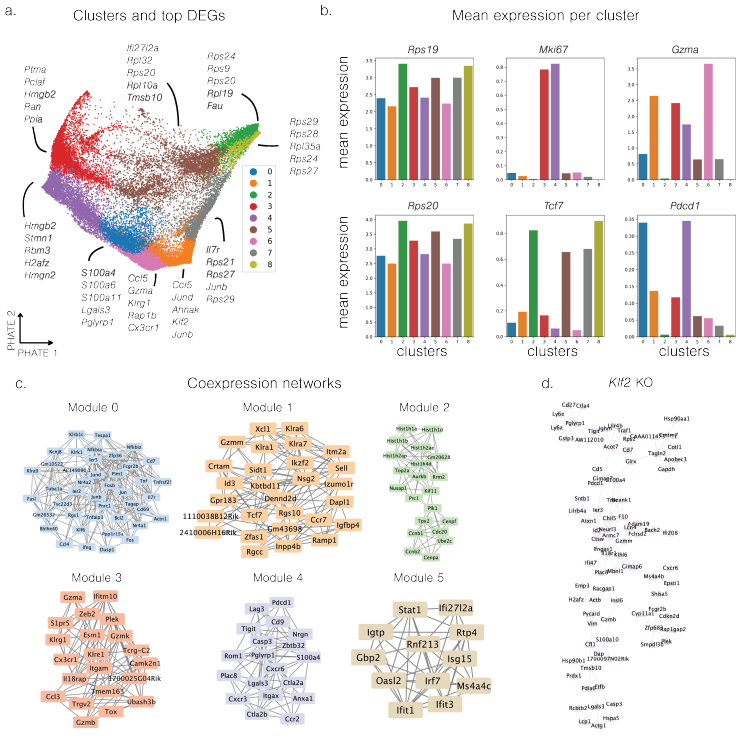

Figure A.6: Extended CD8+ T cell analysis. a. Cells clustered with top DEGs identified. b. Mean expression of top DEGs *Rps19* and *Rps20* and key CD8+ T cell marker genes per cluster, highlighting how differentially enriched genes and T cell markers do not show cluster-specific expression. c. Coexpression networks of top localized genes in each gene cluster. d. *Klf2* KO network, constructed from gene-gene k-NN graph of negative control and *Klf2* KO, then visualizing only those connections that do not exist in the KO network.

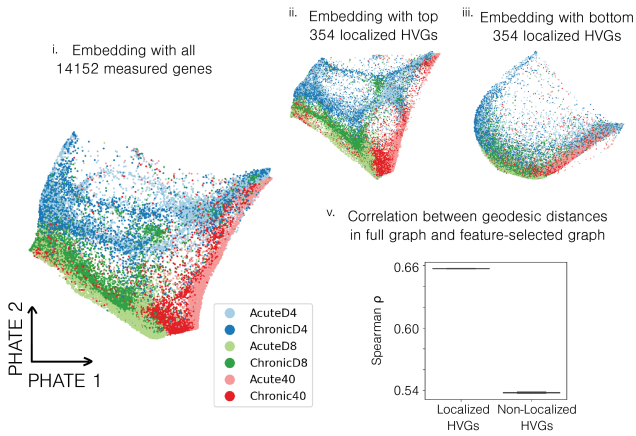

Figure A.7: Biologically-relevant gene patterns from data with GSPA. Cell embeddings visualized with PHATE, with (i) all genes (ii) top 25% localized highly variable genes, and (iii) bottom 25% localized highly variable genes. (v). Spearman correlation between random subset of 100,000 pairwise geodesic distances (chosen in two runs) in full cell-cell graph and feature-selected cell-cell graphs.

A.9.1 Computing and comparing type 1 interferon signaling signature

To determine the enrichment of the type 1 interferon signaling gene signature, we constructed gene embeddings for all approaches designed to map the gene space. Then, we identified gene modules using Leiden clustering, chose the gene module containing canonical type 1 interferon marker *Irf7*, and selected the top 10% localized genes within the gene module. This allowed us to choose genes that were both related to type 1 interferon signaling, through similarity to *Irf7*, but were also unbiasedly selected based on the calculated gene modules and localization score. We next wanted to add additional comparisons to other canonical approaches for identifying gene signatures. To compare against analysis done by clustering cells and identifying differentially expressed genes, we selected the top 100 DEGs from each cell cluster. Finally, to compare against factor analysis approach cNMF [39], we extracted the gene program for which *Irf7* had the highest loading, then selected the genes with the highest 10% loading score to that program. To compare the biological relevance of selected genes from each comparison, we performed gene set enrichment analysis using Enrichr [34] and the BioPlanet gene set resource [40], and visualized enrichment scores for a type 1 interferon-related gene set.

A.9.2 Building module-specific gene coexpression networks

While gene modules group genes based on relatively similar expression profiles, the localization score determines how specific that expression profile is. For example, *Rps20* and *Tcf7* both belong to gene module 1, but, because *Rps20* shows high expression in other cells, whereas *Tcf7* shows almost no expression in other cells, *Tcf7* has a higher localization score. Therefore, to build module-specific gene coexpression networks, we identified the top 10% localized genes in each gene module, then built a k-NN graph with $k = 5$ from the GSPA+QR gene representations. Networks were then visualized with Cytoscape [41]. We performed protein-protein interaction analysis with STRINGdb [31] by testing if each module showed significantly higher interaction than expected for a random set of proteins of the same size and degree distribution.

A.9.3 Building perturbation-specific gene coexpression networks

First, we identified genes that were in the top 25% localized in both the negative control and the knockout (KO). Then, we built a k-NN graph with $k = 5$ for the negative control genes, and a k-NN graph with $k = 100$ for the KO genes from the GSPA+QR representations. We subtracted the KO adjacency matrix from the negative control adjacency matrix and built a new graph from the positive entries, visualizing this graph with Cytoscape. This effectively identifies coexpression edges that are in the negative control that are not in the KO gene-gene graph. Notably, the difference in k was in

order to emphasize connections that were very similar in the negative control and very different in the KO. For visualization, we removed disconnected subgraphs consisting of 2 or fewer nodes.

A.10 Predicting immunotherapy treatment response with GSPA gene embeddings

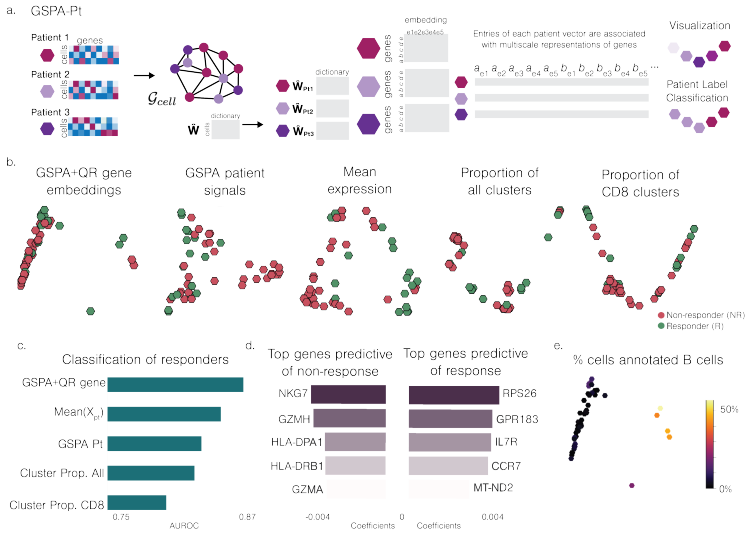

Figure A.8: Response trajectories and biomarkers revealed by multiscale GSPA patient manifold. a. Schematic of GSPA-Pt. b. PHATE visualization of patient embeddings based on GSPA+QR gene embeddings and comparisons. c. AUROC evaluation of response classification (logistic regression). d. Top genes predictive of response and non-response based on highest and lowest logistic regression coefficients. e. Patient embedding colored by percent of total cells annotated as B cells.

We hypothesized GSPA gene embeddings could enable response prediction by capturing key features of the gene coexpression and coregulation in the tumor microenvironment. Given immune cells from 48 melanoma patients from [7] treated with immune blockade checkpoint inhibitors, we tested the ability to classify if the patient is a responder or a non-reponder. We first derive an approach to construct patient vectors from GSPA gene embeddings, where, as features of the patient vector correspond to genes, we can explore genes predictive of response (Figure A.8a). We compared this approach to GSPA embeddings of patient set indicator signals on the cell-cell graph, as well as standard single-cell patient comparison approaches: patient vectors based on cell type proportions, CD8+ T cell subtype proportions, pseudobulked mean gene expression. We then trained a logistic regressor to classify responders versus non-responders. GSPA gene embeddings achieved the highest classification performance (Figure A.8b-c). As patient embeddings comprise gene features, the coefficients of the logistic regressor reflect the importance of different genes for prediction (Figure A.8d). Many important genes were related to T cell function, reflecting their role in tumor recognition and control. Genes most associated with non-response include NKG7 (rank 1), GZMA (rank 5) and CD38 (rank 28), resembling known terminal differentiation programs. Genes associated with response include IL7R (rank 3), CCR7(rank 4) and TCF7 (rank 16), linked to T cell progenitor states such as stemness, memory, activation and survival, and reflecting the known role of progenitor T cell states as immunotherapy targets. While mean expression-based embeddings showed comparable gene rankings for some markers (NKG7 (rank 3), IL7R (rank 17) and CCR7 (rank 3)), other markers are ranked lower, including GZMA (rank 115), CD38 (rank 176) and TCF7 (rank 499). Finally, GSPA gene embeddings reveal a distinct group of patients with significantly higher proportion of B cells, revealing information beyond T cells not captured by other methods (Figure A.8e). This result demonstrates that using GSPA gene embeddings enables interpretable and improved prediction on perturbation-based independent benchmarks.

A.10.1 GSPA-Pt framework

In the GSPA-Pt framework, we first consider $\mathbf{X}_{Pt_p}$ as a single-cell dataset for patient p for $p \in 1...P$. We then concatenate all samples to build a shared cell-cell graph $\mathcal{G}_{cell}$, which we use to build the wavelet dictionary $\widehat{\mathbf{W}}$ as before. As each entry in $\widehat{\mathbf{W}}$ is associated with a patient $p \in 1...P$, we can split $\widehat{\mathbf{W}}$ into patient-specific dictionaries $\widehat{\mathbf{W}}_{Pt_1}$, $\widehat{\mathbf{W}}_{Pt_2}$,...,$\widehat{\mathbf{W}}_{Pt_P}$. Then, for each p, we project $\mathbf{X}_{Pt_p}$ onto $\widehat{\mathbf{W}}_{Pt_p}$ and learn a reduced patient-specific gene representation. Each patient is represented by a gene embedding, which is flattened into a vector for downstream analysis.

We performed PCA with 5 components and flattened these gene representations into a single vector of size 1 x $5m$ to represent the patient. We used the first five PCs to represent the patient rather than the autoencoder embedding (as in previous analysis) because the PCs allowed for more interpretable analysis of the coefficients of the classifier. A single dimension of the latent space of the autoencoder may not necessarily capture the major axes of variation for a gene, but the first dimension of the gene PC definitionally captures the major (linear) axis of variation.

For comparison, we performed GSPA using the patient indicator signals on the cell-cell graph. We also computed the mean expression across all cells for each patient. Finally, we computed the proportion of all clusters (representing immune cell types) and all CD8 clusters (representing CD8 cell states). Using these as unsupervised patient representations, we then classified response using a ridge classifier, comparing based on AUROC of classification. Given that the ridge classifier is a linear model, the coefficients represent the features of the patient representation most important for prediction. The features correspond to five components for each gene, so we can map the coefficients to genes relevant for prediction. We visualize all patient embeddings with PHATE.

A.11 Human lymph node spatially variable genes and cell-cell communication

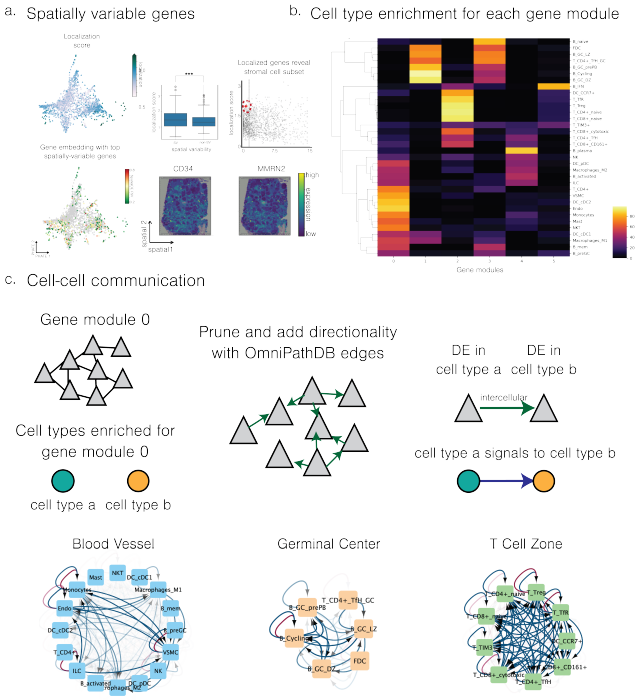

Figure A.9: Expanded case study of human lymph node to spatial transcriptomics tasks. a. Alignment between spatially localized genes and spatially variable genes, with opportunity for new insight into stromal cell subset. b. Cell2location-based mapping of Visium spots to cell types based on enrichment of cell type signature from reference map. Visualization map of gene modules enriched for each cell type. c. Cell-cell communication based on GSPA and cell type mapping.

Often we aim to identify genes with spatial expression variation across the tissue, where analysis based on cell annotations or clusters only detect variations between discrete groups and often do not incorporate spatial information [42]. This has motivated the development of approaches, including SpatialDE, to identify spatially variable genes, or genes that are localized in space. GSPA naturally lends itself to identifying spatially variable genes through computing localization scores informed by both expression and spatial distribution of expression (Figure A.9a). We show spatially variable genes identified by SpatialDE have a significantly higher localization score than non-spatially variable genes. Additionally, by using both the expression graph and the spatial graph to determine localization, GSPA is further empowered to identify relevant biology. Two example genes that are considered localized but insignificant by SpatialDE, CD34 and MMRN2, are enriched in the adventitia of the vasculature and have been previously implicated as progenitor cells that may give rise to other fibroblast subsets [43].

We next show we can use GSPA gene modules to perform gene module-based cell-cell communication analysis. As the human lymph node data is from 10X Visium, each spot is not at single-cell resolution and instead consists of 3-10 cells, which may comprise multiple cell types. Thus, we first map the spatial data to cell types using a reference lymph node atlas with Cell2location [44]. We then visualize how each cell type is enriched for each gene module (Figure A.9b). This reveals expected patterns of colocalized cell types in space based on shared gene modules. For example, gene module 1, which captures the germinal center based on the tissue structure in the H&E, is enriched in B_GC_LZ (B cells in the germinal center light zone), B_GC_DZ (B cells in the germinal center dark zone), T_CD4+_TfH_GC (CD4+ T follicular helper cells in the germinal center), and cycling B cells.

We can thus leverage this to predict cell type to cell type signaling (Figure A.9c). For genes within a given gene module, we construct a k-NN gene-gene network and determine the cell types enriched for the gene module. Then, we prune the network to only those edges with high confidence in OmniPathDB [45], a resource that captures prior knowledge interactions from multiple intercellular and intracellular databases, and we add directionality between genes and annotation of whether the edge is intracellular or intercellular. We repeat the following for each gene module graph: for each directed edge (gene_s, gene_t), for all pairs of cell types enriched for the gene module (cell type_a, cell type_b), if gene_s is differentially expressed in cell type_a and gene_t is differentially expressed in cell type_b, we add a directed edge from cell type_a to cell type_b. We finally visualize intercellular communication edges in blue and intracellular communication edges within the same cell types (that is, (gene_s, gene_t) is intracellular and cell type_a = cell type_b) in red. This captures complex, multicellular networks.

A.12 Identifying gene subnetworks associated with patient phenotypes

We hypothesized GSPA would be useful to explore how the local gene interaction landscape changes between patient phenotypic groups from spatial transcriptomic data. As these datasets are independent and not batch-integrated, we could not directly compare the gene embeddings derived from each patient dataset. Thus, we devised a new approach for constructing patient-specific gene-gene networks and comparing these networks directly.

First, for each patient dataset, we compute gene embeddings with GSPA and subsetted to highly expressed genes across all patients, such that analyses are driven by network-based differences rather than the presence or absence of genes. Then, we construct gene-gene k-NN graphs from the gene embeddings for each patient.

As each patient-specific gene-gene network consists of the same genes with different edges, we next aimed to identify genes with the most consistent local network structure within phenotypic groups and most dissimilar network structure between phenotype groups. For example, in the case study of spatial transcriptomic data from patients with hepatocellular carcinoma, we aimed to identify genes with consistent local network structure within responders and within non-responders to cancer immunotherapy, but strong dissimilar structure between responders and non-responders. These genes represent biomarkers for predicting immunotherapy responders and understanding the molecular pathways underlying response.

To this end, for each network, we extracted ego subgraphs centered at each gene and comprised of its 2-hop neighbors. This results in $n_{samples}$ x n_{genes} subgraphs overall, such that each subgraph corresponds to a gene from a given sample. We then applied attributed Graph2Vec to compute embeddings for all ego subgraphs. To compare similarity of local network structure between genes,

we computed the differential phenotype score, where, given two phenotypes $phen_1$ and $phen_2$ and a gene g, let $E_g^{phen_1} = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ and $E_g^{phen_2} = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ be the embeddings of subgraphs centered at gene g for $phen_1$ and $phen_2$ patients, respectively. Then, we computed the mean pairwise distance between all embeddings within $E_g^{phen_1}$, within $E_g^{phen_2}$, and between $E_g^{phen_1}$ and $E_g^{phen_2}$ to get S_{phen_1}, S_{phen_2}, and $S_{phen_1,phen_2}$, respectively. Then, the differential phenotype score for gene g is as follows:

$$\text{Differential Phenotype Score}(g) = \frac{1}{2}\left(S_{phen_1} + S_{phen_2}\right) - S_{phen_1,phen_2} \tag{A.2}$$

A negative score indicates that the gene's subgraph is more consistent within each group than it is between the groups. As a result, we selected the genes with a negative differential score for analysis of enriched molecular pathways.

A.13 Training Details

A.13.1 Default GSPA hyperparameter selection and training details

The cell-cell graph was built with PHATE using default parameters (k=5 and α decay=40) from the PCA space, as common for cell-cell graph construction. The power was set by default to 2 to mimic the dyadic scales in [4] and J was set by default to $log(n)$ based on [18] (see Lemma A.2 and surrounding discussion above). For GSPA+QR, the epsilon parameter was set to $1e$-3. The data was first dimensionality reduced with PCA to 2048 components (which captures the majority of variation), and then an autoencoder nonlinearly reduced the dimensionality further to latent dimension of 128. The autoencoder was designed with 2 layers with bias in the encoder and decoder, with a relu activation function between layers. The models were trained for an MSE objective with an Adam optimizer with learning rate of 0.001 for 100 epochs, with early stopping (patience of 10) using the loss of a validation set 5% of the size of the training set. For all analyses, signals are first L2 normalized before projection.

A.13.2 Comparison hyperparameter and training details

For method comparisons in Figure 2, we ran each method three times, including reconstructing the graph with new seeds. All signals were first L2 normalized, and, where applicable, dimensionality reduced using PCA with 2048 components and an autoencoder (AE) with latent dimension of 128 (PCA+AE; same configuration as for GSPA). For raw measurements, we ran PCA+AE on $\mathbf{X}$. For MAGIC($\mathbf{X}$), we compute the diffusion operator with default parameters. We then project the signals onto this diffusion operator and run PCA+AE. We compute eigenscores based on the approach described in [17], then dimensionality-reduced with PCA+AE. We learned multiscale representations with DiffusionEMD and GFMMD, then dimensionality reduced with PCA+AE. For signal-signal graphs, k-NN graphs were generated from the signals with $k = 5$. Node2Vec was run on this graph with latent dimensionality of 128, walk length of 80, and 10 walks. $\text{GAE}_{\text{no-att}}$ was run with graph convolutional layers, and GAE_{att} was run with graph attention layers on this graph. The GAE configuration matched the previous AE configuration. For SIMBA, we constructed a heterogeneous cell-gene graph using default parameters, without highly variable genes. We then trained the graph embedding with 128 dimensions, auto-estimating weight decay. For siVAE, we constructed the encoder-decoder architecture with the same number and size of layers as our GSPA autoencoder. We additionally employed 2000 iterations, mb_{size} of 0.2, $l2_{scale}$ of 1e-3, learning rate of 1e-4, decay rate of 0.9, and early stopping with a patience of 100 iterations. We used relu activations between layers.

A.14 Datasets and Pre-processing

A.14.1 Simulated datasets with Splatter

Three datasets were simulated using Splatter [5] with one (linear) trajectory, two branches, and three branches. All datasets were simulated with 10,000 cells and 10,000 genes, where cells were distributed equally between branches (where applicable). The dropout probability was set to 0.95 to generate "noisy" datasets, and each dataset had associated "true" noiseless counts from the same experiment. This dropout level is comparable to true single-cell data (84.3–85.3% sparse). After simulation, genes expressed in less than 50 cells were removed, and the matrix was L1 normalized for

library size and square-root transformed (or log-transformed for robustness analysis). This resulted in 8821 genes in the linear simulation, 8820 genes in the two-branch simulation, and 8823 genes in the three-branch simulation. Cells were then visualized with PHATE.

For the rare cell type experiment, we simulated a dataset with two clusters, one abundant (95%) and one rare (5%), using Splatter [5] and preprocessed the data the same as above.

A.14.2 CD8+ T cell scRNA-seq dataset

Mice were infected with lymphocytic choriomeningitis virus (LCMV) Armstrong (Acute) and Clone 13 (Chronic), and CD8+ CD44+ Tetramer+ T cells were FACS sorted prior to 10X Chromium 5p single-cell RNA sequencing at day 4, day 8, and day 40 [6]. 3-5 mice were infected for each timepoint/condition in a staggered manner to enable same day take down of each timepoint. Spleens from mice were pooled for each timepoint/condition and sorted prior to their loading on the Chromium instrument. 10,000 cells were loaded into a lane of the instrument for each timepoint/condition. The resulting 10X libraries were sequenced on an Illumina NovaSeq with an approximate read depth of 20,000 reads per cell. We then processed the data using CellRanger before further filtering. Cells expressing less than 200 genes, with less than 500 counts or more than 25000 counts, were removed. Genes expressed in less than 3 cells were removed. Cells with mitochondrial percentage greater than 6% were removed. We then L1 normalized for library size, log-transformed, and clustered cells using Leiden clustering, removing contaminating populations enriched for non-CD8+ T cell markers. The acute and chronic datasets were combined, and highly variable genes were detected as the top 10% of genes using `scprep` (https://scprep.readthedocs.io/en/stable/). This resulted in 14,152 genes and 39,704 cells detected across datasets, with 6,811 cells from Acute Day 4; 7,418 cells from Acute Day 8; 6,740 cells from Acute Day 40; 6,205 cells from Chronic Day 4; 7,553 cells from Chronic Day 8; and 4,977 cells from Chronic Day 40. The combined datasets were then visualized with PHATE, and key marker genes were visualized on the PHATE embedding with MAGIC. Graphs for PHATE and MAGIC were built with default parameters, except k for the k-NN graph construction was set to 30 due to the larger number of cells.

A.14.3 Immunotherapy response in melanoma patients scRNA-seq dataset

We obtained pre-processed scRNA-seq data with annotated cell types and other relevant metadata (e.g. sample labels, patient response) from [7] and the Single Cell Portal (https://singlecell.broadinstitute.org/single_cell). From this data, there were 48 samples, which corresponds to 19 pre-therapy samples and 29 post-therapy samples, as well as 31 nonresponder samples and 17 responder samples. There were 15,300 cells and 12,364 genes detected across all samples, with 10,190 cells from nonresponders and 5,110 from responders.

A.14.4 10x Human Lymph Node Spatial Transcriptomics Dataset

10x Genomics data was obtained from the 10x website [8] and downloaded via the `scanpy` package [38]. According to their website, 10x obtained fresh frozen human lymph node tissue from BioIVT Asterand Human Tissue Specimens. The tissue was embedded and cryosectioned as described in the Visium Spatial Protocols Tissue Preparation Guide (Demonstrated Protocol CG000240). Tissue sections of 10 µm thickness were placed on Visium Gene Expression Slides. We removed spots with less than 5000 counts, more than 35000 counts, and over 20% mitochondrial counts. We removed genes detected in fewer than 10 cells, L1 normalized for library size, and log-transformed the data. After the above pre-processing, there were 3861 spots and 19,685 genes detected. The top 2000 highly variable genes were selected following the `scanpy` tutorial for this dataset.

A.14.5 Hepatocellular carcinoma Spatial Transcriptomics Dataset

To compare responder and nonresponder HCC microenvironments, we collected ST data from a study conducted by [9] that investigated the effects of two cancer treatment drugs, cabozantinib and nivolumab, in neoadjuvant therapy on advanced HCC. Specifically, we obtained 10x Visium data from seven patient samples (4 responder, 3 nonresponder) that were acquired through R0 resection. For pre-processing, we removed spots with less than 200 detected genes, applied cell-depth normalization and log transformation.