Mapping the gene space at single-cell resolution with gene signal pattern analysis

Aarthi Venkat[1] Sam Leone[2] Scott E. Youlten[3] Eric Fagerberg[4] John Attanasio[4] Nikhil S. Joshi[4] Michael Perlmutter[5,6] Smita Krishnaswamy[1,2,3,7]

[1]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT 06511, [2]Applied Math Program, Yale University, New Haven, CT 06511, [3]Department of Genetics, Yale University, New Haven, CT 06511, [4]Department of Immunobiology, Yale University, New Haven, CT 06511, [5]Department of Mathematics, Boise State University, Boise, ID 83725, [6]Program in Computing, Boise State University, Boise, ID 83725, [7]Department of Computer Science, Yale University, New Haven, CT 06511

Abstract

In single-cell sequencing analysis, several computational methods have been developed to map the cellular state space, but little has been done to map the gene space. Here, we present a graph signal processing approach to learning rich gene representations from single-cell data using a dictionary of diffusion wavelets on the cell-cell graph. This allows for the characterization of genes based on their patterning on the cellular manifold. It also enables us to understand how localized or diffuse the expression of a gene is, for which we present a score called the *gene localization score*. We find that highly localized genes can be used to better characterize the cellular space, especially for trajectory-like structure. We formulate the gene embedding problem setup, design tasks with simulated single-cell data to evaluate representations, and establish eight relevant baselines. We also motivate and demonstrate the efficacy of this method for a range of biological datasets and questions, such as identifying gene coexpression modules and perturbation-specific gene-gene interactions, learning active gene signaling networks from single-cell and spatial data, and classifying therapeutic response from patient-specific gene signatures.

1 Introduction

A variety of techniques are used to map the cellular state space for single-cell RNA sequencing (scRNA-seq) analysis, including PCA, t-SNE, UMAP, HSNE, and PHATE [3, 5, 19, 40, 58]. These methods build low-dimensional embeddings based on the similarity between cells across thousands of genes. They reveal the organization of the cell landscape, including clusters of similar cells, trajectories along phenotypic continuums, and multiscale structure fundamental to complex organisms. The expression of genes in cells is also highly organized. Genes are coordinated into complexes, biological processes, and pathways. However, despite the myriad techniques for understanding the cellular state space, it has not been possible to apply them analogously to understand the gene landscape. This is because the amount of noise affecting each gene's measurement depends on its intrinsic level of expression and can vary by an order of magnitude between genes. This affects our ability to quantify gene-gene similarity and construct a low-dimensional gene embedding, obscuring the structure of the gene landscape.

Therefore, rather than building gene embeddings from single-cell data directly, we make the critical framing of genes as *signals* on the cell-cell graph. However, graph feature or graph signal embeddings have been largely understudied compared to graph node embeddings in machine learning, with most advances arising from the field of graph signal processing. Graph signal processing models attributes on graph nodes as signals or functions. Analysis of such attributes extends classical signal processing concepts and tools, such as Fourier and wavelet analysis, to data residing on graphs [42]. For example, [12] shows that diffusion wavelets, through the use of the diffusion operator and dyadic powers, induce multiresolution analysis, useful for capturing local and global structure of the underlying graph. By considering gene expression measurements as signals on the cell-cell graph, we can leverage fundamental ideas from graph signal processing for feature representation learning.

Here, we present gene signal pattern analysis (GSPA), a new method for embedding genes in single-cell datasets using a novel combination of diffusion wavelets and deep learning. First, we build a cell-cell graph (Figure 1a) and define any genes measured as signals on the cell-cell graph (Figure 1b). Then, we decompose the gene signal using a large dictionary of diffusion wavelets of varying scales that are placed at different locations on the graph (Figure 1c). The result is a representation of each gene in a single-cell dataset as a set of graph diffusion wavelet coefficients. We pair this representation with an autoencoder framework to reduce the dimensionality of the gene representation and make it suitable for tasks such as visualization and prediction of gene characteristics (Figure 1d). In addition, we show the utility of ranking genes based on their distance to synthetic signals. This enables cell type association of gene signals, as well as a new type of analysis, which we term *differential localization* (Figure 1e). We highlight the importance of genes that are highly localized on the cellular manifold versus genes that are diffusely expressed, hypothesizing that such genes are beneficial in characterizing different areas of the manifold informatively. We also show that differential localization has several advantages compared to standard cluster-based differential expression pipelines, including reduced false positives [50] and the identification of genes with patterns of localization that do not conform to cluster boundaries [27].

As no previous tasks exist to evaluate algorithms for gene embeddings, we designed two experiments with three simulated benchmarking datasets to test the ability of gene embeddings to (1) preserve gene-gene relationships, and (2) capture the gene localization score. We showcase the superior ability of our approach to preserve coexpression relationships, enable visualization of the gene space, identify salient gene modules, and reveal gene localization.

We further demonstrate the utility of GSPA for biological interpretation in multiple settings (Figure 1f). First, we analyze a newly generated scRNA-seq dataset of CD8+ T cells in response to acute and chronic infection at days 4, 8, and 40. We visualize the gene space and show that localized genes correspond to meaningful biological processes. Additionally, we identify gene modules corresponding to naive, proliferative, effector, exhausted, memory, and interferon-responsive T cell states. We also characterize gene networks perturbed in CD8+ T cells at acute infection day 8 in a newly generated single-cell CRISPR screen. This reveals gene-gene relationships involved in effector differentiation for *Tbx21* knockout and related to memory and effector capacity for *Klf2* knockout.

Next, we analyze the local infiltration of antigen-specific CD8+ T cells in the mouse epidermis [14]. We use GSPA to represent ligand-receptor interactions and pathway-pathway relationships in a cluster and cell type-independent fashion, enabling analysis of cell-cell communication within and across cell types. This

analysis recovers a condition-specific event in which a subset of myeloid cells sends signals to a subset of T cells via the PD-L1 checkpoint.

We also present a version of GSPA for spatial transcriptomic data which integrates both coexpression and spatial colocalization to define gene-gene relationships. We identify gene modules in key lymph node substructures corresponding to the blood vessel, germinal center, T cell zone, B follicle, B plasma cell region, and B interferon signature region [1]. We also characterize microenvironmental signaling events within each gene module between cell types enriched in the substructures.

Finally, we leverage GSPA to map and compare single-cell datasets from patients with melanoma [45]. We calculate gene embeddings for 48 samples from patients pre- and post-immunotherapy, and we use these reduced representations to visualize the patient manifold and predict therapeutic response. We identify genes most predictive of response as associated with T cell progenitor states, and conversely genes predictive of non-response with T cell terminal differentiation states, consistent with prior literature. Code to run GSPA and generate results and figures here is available at https://github.com/KrishnaswamyLab/Gene-Signal-Pattern-Analysis.

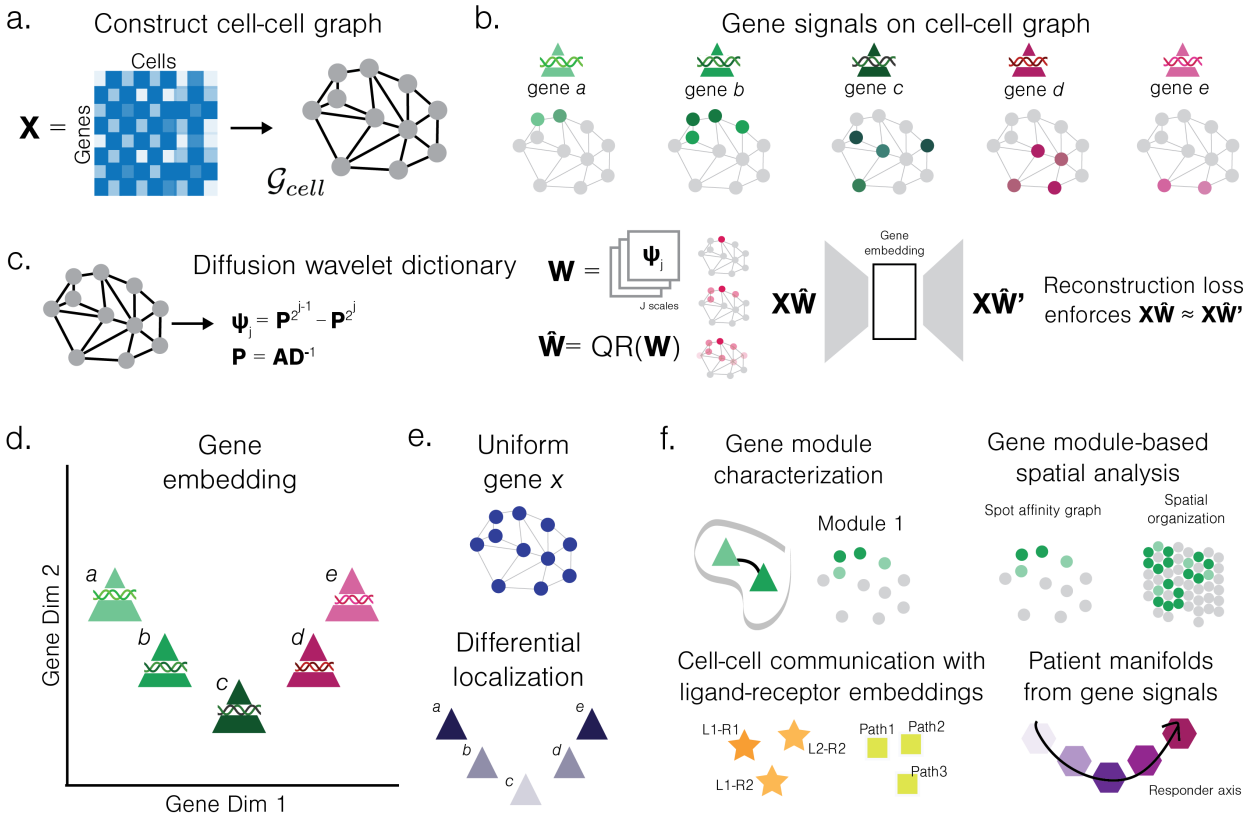


Figure 1. Overview of Gene Signal Pattern Analysis. a. Construction of cell-cell graph based on gene profile affinities. b. Demonstrative gene signals (triangles) defined on nodes of cell-cell graph. c. Diffusion wavelet dictionary defined at J scales, with optional rank-revealing QR factorization, and following autoencoder. d. Demonstrative gene embedding, where similar gene patterns are closer together despite signatures that are not overlapping and not cell type-specific. e. Diagram of differential localization, where a and e are most localized and c is least localized. f. Example downstream applications of gene signal pattern analysis. **P**: diffusion operator; **A**: adjacency/affinity matrix; **D**: degree matrix; L-R: ligand-receptor; Path: pathway

2 Results

2.1 Gene embedding 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 $\mathbf{X}$, our goal is to obtain a low-dimensional representation of each gene which preserves the inherent structure of the gene space. In particular, we seek a dimension reduction map $\Theta : \mathbb{R}^n \to \mathbb{R}^d$, $d \ll n$, which satisfies the desirable properties enumerated below.

Desired Properties:

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 underlying graph geometry, rather than the naive pointwise distance, between gene signals.
2. **Denoising**: In general, high-dimensional measurements are noisy and confounded by biological and technical variations. Therefore, we 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. **Flexible 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.

In the following section, we present our approach, gene signal pattern analysis (GSPA), to achieve these desired properties.

2.2 Gene signal pattern analysis (GSPA) overview

In order to construct the map Θ, our approach, gene signal pattern analysis (GSPA), frames each gene as a signal (function) defined on a cellular graph, where the vertices correspond to cells and edges correspond to cells with similar gene profiles. Here, we focus on gene expression measurements at single-cell resolution with scRNA-seq. We first build a cell-cell similarity graph $\mathcal{G}_{cell}$, which is a standard first step for many single-cell pipelines, including those to map the cellular state space or cluster cells. We then consider the expression of each gene i as a signal $\mathbf{X}_i$ on this graph. In order to create a multiscale representation of these signals, we decompose them using a dictionary of diffusion wavelets [12], optionally compressed via local multiscale orthogonalization which results in the emergence of natural hierarchical coarse-graining of the cell-cell graph (see [12]). Thus, by projecting the gene signals $\mathbf{X}$ onto this wavelet representation, we uncover spatial and frequency information about our signals with respect to $\mathcal{G}_{cell}$. Finally, we learn a meaningful low-dimensional embedding using an autoencoder for reducing the wavelet-based representations (Algorithm 1). This reduced embedding can be used for many downstream analyses including visualization, gene module identification, or differential localization. We expand upon each step of this algorithm in the Methods section.

Algorithm 1 Gene signal pattern analysis algorithm for learning gene representations

Input: Single-cell dataset $\mathbf{X}$ of shape m x n, containing values for m genes and n cells
Output: Gene embeddings of shape m x d for some $d \ll n$, representing the gene space of $\mathbf{X}$
1. Build cell-cell similarity graph $\mathcal{G}_{cell}$ based on pairwise Euclidean distances between cells
2. Compute dictionary of diffusion wavelets $\mathbf{W}$ of different scales centered at all nodes of the graph $\mathcal{G}_{cell}$.
3. (Optional) Calculate a compressed dictionary $\widehat{\mathbf{W}}$ using rank-revealing QR factorization
4. Project gene signals onto the designed wavelet dictionary
5. Use autoencoder architecture $D \circ E$ to learn low-dimensional embedding of gene space

Constructing a cell-cell similarity graph from single-cell data In scRNA-seq, each cell is measured as a vector of gene expression counts. That is, the output of a scRNA-seq experiment is a matrix $\mathbf{X}$ of shape m x n, where m corresponds to the number of genes, and n the number of cells. The first step of the GSPA algorithm is to build a graph $\mathcal{G}_{cell} = (V_{cell}, E_{cell})$, where each node in V_{cell} corresponds to a cell, and each edge E_{xy} in E_{cell} describes the similarity between cell x and cell y.

Building dictionary of graph diffusion wavelets for gene representation Next, we build a dictionary of diffusion wavelets [12] corresponding to the cellular graph that serve as atoms for the representation of each gene. That is, given the cellular graph $\mathcal{G}_{cell}$, we compute the diffusion operator $\mathbf{P}$ and generate a wavelet of scale j centered at vertex v by constructing the operator $\boldsymbol{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j}$ for $1 \leq j \leq J$ (and $\boldsymbol{\Psi}_0 = \mathbf{I} - \mathbf{P}$) and then extracting the v-th row via $\delta_v^T \boldsymbol{\Psi}_j$, where δ_v is the Kronecker delta centered at the v-th vertex. Thus, the set of functions $\{\boldsymbol{\Psi}_j^T \delta_v\}_{v \in V_{cell}, j \in 0,1 \ldots J}$ may be reshaped into a dictionary of wavelets $\mathbf{W}$ of shape n x Jn.

In general, the numerical rank of $\boldsymbol{\Psi}_j$ decreases as j increases, meaning that the small set of large wavelets can be used to describe the graph at a coarse resolution. To reduce this redundancy, we can leverage rank-revealing QR factorization as in [12]. This results in a compressed wavelet dictionary $\widehat{\mathbf{W}} = \{\widetilde{\boldsymbol{\Psi}}_j^T \delta_v\}_{v \in V_{cell}, j \in 0,1 \ldots J}$, where for each scale, $\widetilde{\boldsymbol{\Psi}}_j$ is a set of linear combinations of those wavelets at scale j that account for the most variance. For large values of j, $\boldsymbol{\Psi}_j \delta_x \approx \boldsymbol{\Psi}_j \delta_y$ if vertex x is near vertex y on the graph. We test gene signal embeddings both with (GSPA+QR) and without (GSPA) compression.

Project signals onto the 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. Importantly, we consider each measured gene feature as a signal on the cellular graph. Given a gene signal $\mathbf{X}_i$ and wavelet dictionary $\widehat{\mathbf{W}}$ (or $\mathbf{W}$), we project $\mathbf{X}_i$ onto the dictionary. For all gene signals, this corresponds to $\mathbf{X}\widehat{\mathbf{W}}$ or $\mathbf{XW}$. Theorem 1 stated below shows that the wavelet projection $\mathbf{X} \rightarrow \mathbf{XW}$ is continuous with respect to an unbalanced earth movers distance, UDEMD [54]. Intuitively, this UDEMD 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.

Learning a low-dimensional representation with autoencoder for downstream analysis As $\mathbf{X}\widehat{\mathbf{W}}$ and $\mathbf{XW}$ correspond to a meaningful representation of genes, we can use this matrix for downstream analyses. However, this representation is unnecessarily high-dimensional, such that each feature corresponds to the gene signal projection of a particular cell at a particular diffusion scale. To reduce the dimensionality of the representation, we leverage an autoencoder, consisting of an encoder E and decoder D such that $D \circ E$ is approximately the identity, in order to learn a low-dimensional, information-rich gene embedding.

The GSPA architecture Combining the above steps, we define gene signal pattern analysis (GSPA) by

$$GSPA(\mathbf{X}) = E(\mathbf{X}\widehat{\mathbf{W}})$$

where $\widehat{\mathbf{W}}$ is the compressed wavelet dictionary, E is the encoder discussed above, and GSPA to taken to be the map Θ discussed in the problem setup.

2.3 Achieving desired representation properties with GSPA

Distance preservation As discussed in Section 2.1, our first desired property is an embedding that preserves distances (quantified in a manner informed by the geometry of the cellular state space). Theorem 1 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 Earth Movers Distance, a distance function designed to capture the intrinsic data geometry. Roughly speaking, this distance views the expression of each gene as "mass distribution" over the cells and calculates how much effort it would take to rearrange one mass distribution into another across the cell-cell graph.

Theorem 1. *The diffusion wavelet projection* $\mathbf{W}$ *is Lipschitz Continuous with respect to UDEMD, i.e., there exists a constant* $C > 0$ *such that*

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \leq C \cdot UDEMD(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}). \tag{1}$$

For a proof of Theorem 1 and a detailed statement and discussion of UDEMD, please see Section 6.2.

Denoising Since our genes are decomposed using wavelets, we can automatically denoise their representation. 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., $\mathbf{P} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}$ where the columns of $\boldsymbol{\Sigma}$ contain orthonormal left and right eigenvectors, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{P}$. Thus $\mathbf{P}^t$ effectively results in powering the eigenvalues $\boldsymbol{\Lambda}_{ii}^t$. Since $\mathbf{P}$ is a Markov matrix, this diminishes the higher frequency-lower valued eigenvalues. Therefore, we can restrict the dictionary to wavelets that decompose only the lower frequencies by initially multiplying each wavelet by $\mathbf{P}^t$. As previously shown [16], denoising the signal via $\mathbf{P}^t\mathbf{X}$ results in smoothed signals preserving the lower frequency components. Thus, diffusion wavelets decompose the gene signals into lower and higher frequency components via diffusion, rendering signal representations more amenable to denoising, such as by removing the high-frequency wavelets.

Flexibility to downstream tasks Finally, we achieve flexibility by leveraging an autoencoder framework to learn a generalizable low-dimensional representation. In the following sections, we highlight this flexibility. First, we describe two gene rankings enabled by GSPA representations. Then, we benchmark GSPA against various baselines for the preservation of gene-gene distances and gene localization on three simulated datasets. Finally, we analyze several biological datasets and diverse research questions to highlight the broad applicability of GSPA representations to a wide array of downstream analyses.

2.4 Capturing rankings based on embedded distances from synthetic signals

Beyond preserving relationships within the gene space, GSPA also preserves distances to other signals one can derive on the shared cell-cell graph. This enables flexible ranking of genes based on distance. Here, we describe two approaches based on this idea for cell type association of gene signals and gene localization.

Cell type association of gene signals In systems where cells naturally can be distinguished by cell types, it is standard to characterize cell types by differentially expressed genes [37]. Gene signal pattern analysis naturally enables this characterization. Given a dataset with clusters or cell types, we can first define an indicator signal for cells belonging to a given cell type and normalize it in the same way we normalize other signals. Then, we can rank genes based on how close they are to the multiscale representation of this signal. Unlike many differential expression tests, which compare mean expression between clusters, this gene ranking prioritizes specificity to the cell type. We highly rank genes that are close to zero expression in all other cell types and close to uniform expression in the cell type of interest. Additionally, conventional methods for detecting differential gene expression between clusters often return inflated p-values because of the double use of gene expression data, first to partition the data into clusters and then to define significance statistics along the same partitions [50]. Consequently, filtering based on p-value alone results in an increased rate of false positives, and some pipelines have turned to gene rankings instead of cutoffs [60]. Therefore, we define cell type association ranking of genes by the following score.

Definition 1. *Given normalized cell type indicator signal* $\mathbf{1}_{celltype}(v) = [v \in celltype]$ *for* v *in* V_{cell}*, and wavelet representation* $\widehat{\mathbf{W}}$*, the cell type association score,* $c(i)$ *for each gene signal* $\mathbf{X}_i$ *is defined as:*

$$c(i) := -\|\mathbf{X}_i\widehat{\mathbf{W}} - \mathbf{1}_{celltype}\widehat{\mathbf{W}}\|_2^2 \tag{2}$$

Differential localization of gene signals on the cell-cell graph With gene signal pattern analysis, one property we can decipher is how pervasively or specifically a gene signal is defined. Characterizing differentially expressed genes between clusters is not feasible for many biological systems. For example, for datasets that do not organize into discrete populations, have trajectory-like structure, or consist of subtypes within cell types [61], there is utility in identifying genes localized to particular areas of the cellular manifold without prior cell type identification.

Addressing these limitations, we present an alternative strategy to calculate the specificity of a gene signal $\mathbf{X}_i$ across the cell-cell graph, termed differential localization. We make the observation that genes that are uniformly expressed across all cells are least likely to be involved in cell state-defining biological processes, in line with previous work [60]. Leveraging this insight, we calculate the distance between each gene signal and a multiscale representation of a uniform (constant) signal u, normalized in the same way we normalize other signals. We then rank genes by gene localization score ℓ, where genes most differentially localized are those farthest from the uniform signal representation. Formally,

Definition 2. *The gene localization score, $\ell(i)$ for each gene signal $\mathbf{X}_i$, with normalized uniform signal u and wavelet representation $\widehat{\mathbf{W}}$, is defined as:*

$$\ell(i) := \|\mathbf{X}_i\widehat{\mathbf{W}} - u\widehat{\mathbf{W}}\|_2^2 \tag{3}$$

2.5 Comparison to alternative gene mapping strategies

We evaluate embeddings on three single-cell benchmarking datasets simulated with Splatter (one trajectory, two branches, and three branches) [66]. Using Splatter allows us to generate noisy gene expression data $\mathbf{X}$ as well as the unseen true (noiseless) counts $\mathbf{X}'$ (Extended Data Figure 5a). We adapt eight baselines for comparisons, and we assess the embeddings for preservation of coexpression, fidelity of visualization, coherence of gene modules, and ability to capture signal localization.

Description of gene mapping strategies adapted for comparison For comparison, we pull from prior literature and design additional strategies based on methodology not previously used for gene embeddings. We summarize and diagram these comparisons in Extended Data Figure 1 and expand upon each approach in the Methods section.

First, we can map the gene space with the original measurements $\mathbf{X}$, using an autoencoder to reduce the dimensionality. We term these embeddings the "raw" measurements. Alternatively, we can construct a gene-gene k-NN graph $\mathcal{G}_{gene} = (V_{gene}, E_{gene})$ from $\mathbf{X}$ and employ three graph representation learning methods $\text{GAE}_{\text{att}}(\mathcal{G}_{gene})$ (a graph autoencoder with graph attention layers) [26], $\text{GAE}_{\text{no-att}}(\mathcal{G}_{gene})$ (a graph autoencoder with graph convolutional layers) and $\text{Node2Vec}(\mathcal{G}_{gene})$ [22]. These approaches do not use the cell-cell graph to compute gene representations, which, as described in our first desired property (Section 2.1), may not sufficiently preserve distances. Therefore, we additionally compare against MAGIC [16], which denoises $\mathbf{X}$ via data diffusion on $\mathcal{G}_{cell}$, then use an autoencoder to reduce the dimensionality of the resulting denoised measurements. We demonstrate how diffusion wavelets differ from and improve upon the diffusion-based approach in MAGIC in Extended Data Figure 2, Extended Data Figure 3, and in the Methods section.

We also compare against techniques that consider genes as signals on $\mathcal{G}_{cell}$. Diffusion Earth Mover's Distance (DiffusionEMD) [55] computes optimal transport, or Wasserstein's distance, based on multiscale diffusion kernels. Graph Fourier Mean Maximum Discrepancy (GFMMD) [36] 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. Finally, eigenscores [23], proposed for feature selection, were shown to be useful for mapping the gene space based on their alignment to low-frequency patterns in the data. We again reduce the dimensionality of these resulting multiscale gene representations via an autoencoder.

In summary, the baselines are as follows:

- Raw measurements approach embeds $\mathbf{X}$.
- $\text{GAE}_{\text{no-att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$.
- $\text{GAE}_{\text{att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$.

- Node2Vec($\mathcal{G}_{gene}$) embeds $\mathcal{G}_{gene}$.
- MAGIC($\mathbf{X}$) embeds X after denoising with $\mathcal{G}_{cell}$.
- DiffusionEMD($\mathbf{X}$, $\mathcal{G}_{cell}$) embeds $\mathbf{X}$ via optimal transport on $\mathcal{G}_{cell}$.
- 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 $\mathcal{G}_{cell}$.

These approaches represent a wide array of techniques adapted from graph signal processing, graph representation learning, and single-cell analysis for fair comparison.

Preservation of coexpression relationships in gene embeddings For our first task for comparison, we recognized that genes that are coexpressed should be close in a low-dimensional embedding and genes that are expressed in very different cells should be far apart. To assess this, we first define the coexpression between genes i_1 and i_2 as the correlation of the true counts $\mathbf{X'}_{i_1}$ and $\mathbf{X'}_{i_2}$. Then, we learn gene embeddings e_{i_1} and e_{i_2} from the noisy counts $\mathbf{X}$ and compute their distance. Finally, we use the correlation between this distance and their coexpression as a metric for evaluation (Figure 2a, Extended Data Figure 4a). GSPA+QR performed best on all three benchmarking datasets, followed by GSPA on average (Figure 2b, Extended Data Figure 4b, Extended Data Figure 4c). Furthermore, methods that use $\mathcal{G}_{cell}$ to compute gene embeddings on average have a higher rank (rank=3.75) than methods that do not (rank=8.13), reifying the distance preservation and denoising properties of using the cell-cell graph (Extended Data Table 1).

Visualization not impeded by technical factors of gene variation We next assessed the ability to visualize the gene embeddings, which remains a useful tool for characterization and hypothesis generation [5, 40, 59]. We visualized each gene embedding with PHATE [40] and determined that, for the majority of methods, the major axis of variation captured by visualization is strongly correlated with the number of cells the gene is expressed in (Figure 2c, Extended Data Figure 5b, Extended Data Figure 5c). This is despite normalizing each gene signal (see Methods). By contrast, GSPA+QR and GSPA gene visualizations are not impeded by this confounding factor of gene expression.

Coherence of gene modules on the cellular manifold Finally, we evaluated if clustering the gene embeddings results in coherent gene modules. After clustering the gene embedding from GSPA+QR, we visualize the gene module enrichment on the cellular manifold, and we find that genes within each module show similar activation. We also see relationships across modules, where genes are organized by when they peak on the cellular trajectory (Figure 2d, Extended Data Figure 4d, Extended Data Figure 4e). By contrast, clustering the gene embedding derived from the raw measurements reveals gene module enrichment scores that are not interpretable and do not show a relationship with time (Figure 2e).

Gene localization assessment from embeddings Beyond gene-gene coexpression, we benchmarked embeddings for the ability to capture how differentially localized each gene is. For this task, we generated and embedded artificial signals of varying localization, where more localized signals were defined in smaller windows throughout the trajectory (Figure 3a). Then, we embedded the uniform signal. We assess localization using the anti-correlation between the distance and the size of the window as a metric for evaluation (as described in Definition 2) (Figure 3b). GSPA+QR outcompeted baselines for all three simulated datasets, followed by GSPA (Figure 3c, Extended Data Figure 6). Again, methods that use the cell-cell graph for signal embeddings ranked higher (rank=3.50) than methods that do not (rank=8.50) on average, demonstrating the informativeness of the cell-cell graph for signal embeddings (Extended Data Table 1).

2.6 Gene embeddings reveal coexpression relationships in CD8+ T cells

We first chose to study CD8+ T cell differentiation in response to infection. T cells adopt a range of differentiation states in acute and chronic viral infections, creating diverse and heterogeneous substates differing by phenotype, physiological role, morphology, proliferative potential, and location [10]. Such heterogeneity has been established as a key feature of T cell immunity, where cells can differentiate between states in a

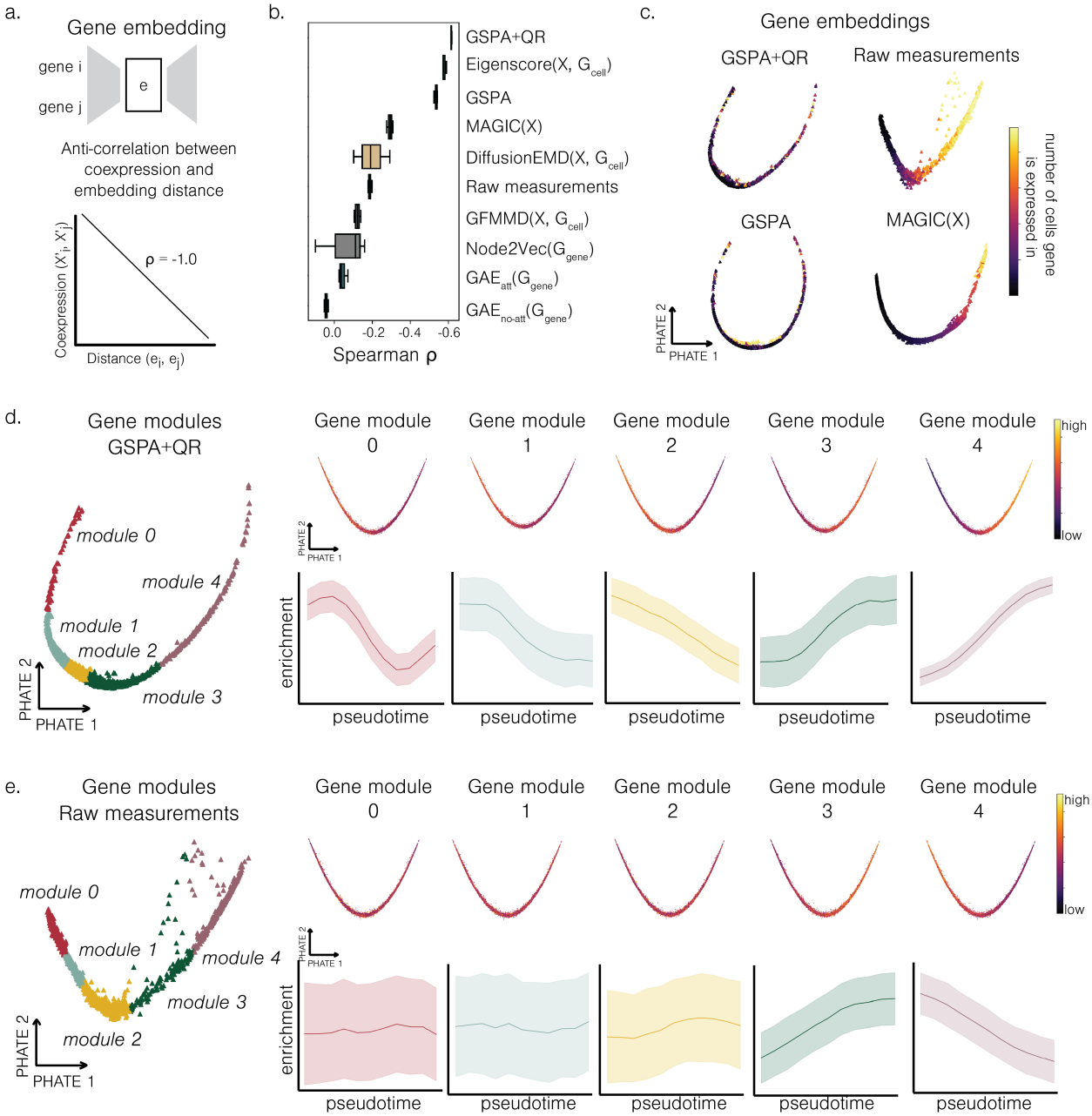


Figure 2. Gene signal pattern analysis preserves coexpression between genes and captures coherent visualization and gene modules. a. Experimental setup. b. Spearman correlation evaluating performance on task for all comparisons across 3 runs. c. PHATE visualization of gene embeddings for GSPA+QR, GSPA, and two comparisons d. Module enrichment and dynamics for GSPA+QR and e. comparison.

"plastic" way. However, the gene signaling pathways and relationships that characterize these transitions are not fully known and can provide new insights for therapeutic intervention.

We therefore investigated a newly developed dataset comprising 39,704 sorted CD8+ Tetramer+ T cells, sequenced at three timepoints (Day 4, Day 8, and Day 40) from two conditions (acute LCMV and chronic LCMV infections). We preprocessed this data and visualized the cellular manifold with PHATE, revealing distinct structure that does not separate into clear clusters or trajectories, as expected due to the dynamic and multiscale nature of T cell heterogeneity. Visualizing known markers of T cell differentiation further

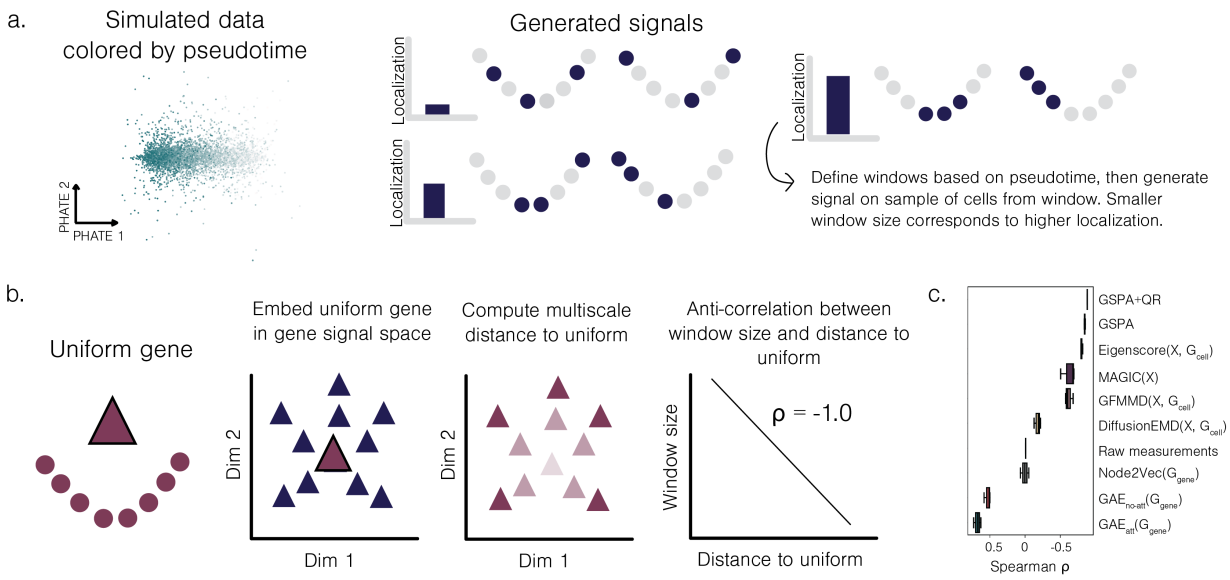


Figure 3. GSPA enables analysis of differential localization of signals. a. Left: Simulated linear trajectory with noise, colored by pseudotime (given by simulation) visualized with PHATE. Right: Diagram of generation of signals, where the window is defined by pseudotime range, and signals are generated in differently sized windows. b. Diagram of how localization reveals genes that are most distant from uniform. c. Spearman correlation evaluating performance for all comparisons across 3 runs.

highlights that key genes are not necessarily specific to a particular cluster, and some gene signatures, such as genes related to proliferation (e.g. *Mki67*), overlap others, such as genes related to activation (e.g. *Gzma*) (Figure 4a). This motivates mapping the gene space to capture key signatures across conditions and timepoints.

We computed representations for the top highly variable genes using GSPA+QR. We then visualized the gene embedding and clustered genes into six modules (Figure 4b, Extended Data Table 2). Next, we computed the localization score of each gene and visualized this score on the gene PHATE (Figure 4c). Genes that had higher localization scores were on the periphery of the embedding, which fits our intuition, as genes that are more spread out are likely to be less committed to one particular population or region of the cellular graph. Furthermore, we saw little relationship between the localization score and number of cells expressing each gene (Spearman $\rho = -0.159$) or the mean overall expression for each gene (Spearman $\rho = -0.108$), indicating that the score is capturing a meaningful structural property of the signal.

To better compare and contrast localization and differential expression in real-world data, we clustered the cells and computed a rank based on the differential expression score in all clusters (Figure 4d). This ranking was slightly positively correlated with the localization score ($\rho = 0.327$). Genes that had high cluster-based rank and high localization included genes strongly enriched in a particular cluster (e.g. *S1pr5*). Genes that had low cluster-based rank and low localization included genes with low or non-specific enrichment overall (e.g. *Trav13d-1*). Other genes, such as *Rps20*, were ranked highly by the cell-clustering rank but lowly by the localization rank due to high mean expression in one cluster, but high overall expression and spread throughout the manifold. Finally, some genes, including *Tox*, were ranked highly by localization but lowly by cell-clustering rank, due to their varied but specific enrichment across the T cell manifold. Together, this demonstrates the ability for localization to not only capture cluster-specific signatures without prior clustering but also prioritize more subtle, lowly-expressed signatures and deprioritize ubiquitously expressed genes. In addition, we show that re-embedding cells with the top localized genes better preserves the overall manifold structure than re-embedding cells with the bottom localized genes and a random subset of highly variable genes, highlighting the utility of localization for topologically-informed feature selection (Extended Data Figure 7).

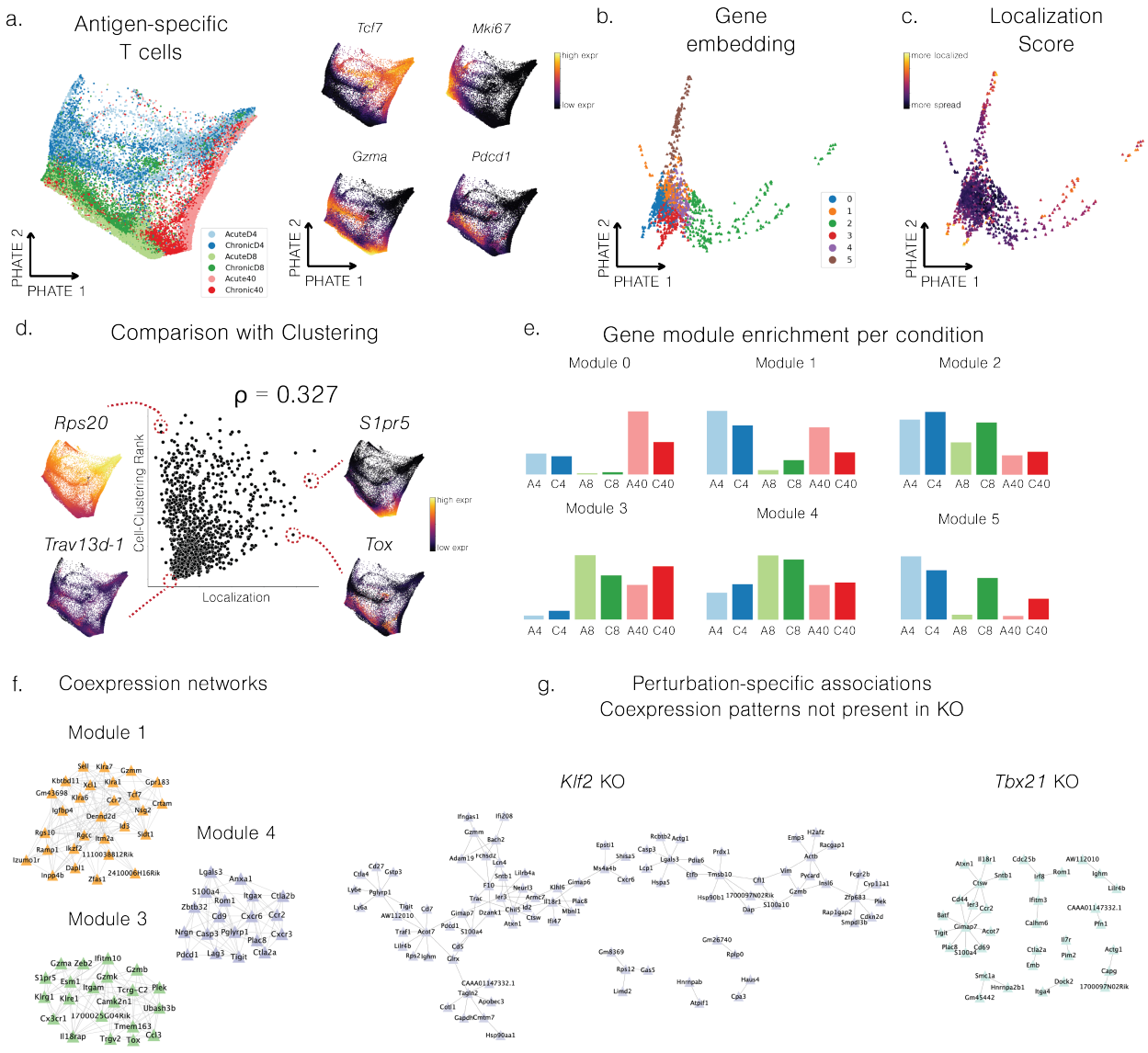


Figure 4. T cell gene signals reveal key signatures of differentiation. a. PHATE embedding of antigen-specific CD8+ T cells from six experimental conditions (left) and marker genes visualized (right). b. Gene embedding visualized with PHATE, colored by gene module assignment. c. Gene embedding visualized with PHATE, colored by computed localization score. d. Cell clustering rank versus localization score, with representative genes visualized to demonstrate similarities and differences. e. Gene module enrichment per condition. f. k-NN graphs of top localized genes from three modules. g. k-NN graphs of gene-gene coexpression relationships that were knocked out in *Klf2* and *Tbx21* knockouts.

Next, we characterized each gene module based on marker genes and gene set enrichment scores, which did not conform to distinct cluster-like shapes (Extended Data Figure 8). We identified cells with strong enrichment for each gene module and compared the proportion of these cells from each condition, which revealed many expected patterns from existing literature [15, 20]. Gene module 1, enriched for naive and memory genes, had a higher proportion of cells from Acute Day 4 and Day 40 timepoints. Module 2, with proliferation genes, was enriched in Acute Day 4 and Chronic Day 4. Module 3, with effector genes, was enriched in Acute Day 8 and Chronic Day 8, and module 4, with late effector and exhaustion genes, was enriched in Acute Day 8, Chronic Day 8, and Chronic Day 40 (Figure 4e). We also identified two other gene

modules, modules 0 and 5. Module 0 seemed to capture a memory-specific signature as it was most highly enriched in the Acute Day 40 timepoint and contained hallmark genes such as *Il7r* and *Bcl2* [21]. Module 5 captured an interferon response signature, which was interestingly present in both Acute and Chronic Day 4, but only seemed to persist past that timepoint in the Chronic setting. This is in line with findings that showed type 1 interferon can enforce T cell exhaustion in chronic infection settings [39, 52, 63].

For each gene module, we selected the most highly localized genes and built a k-NN graph, highlighting the strong gene-gene coexpression relationships in each module (Figure 4f). For example, the naive/memory gene-gene graph contained edges between *Tcf7*, *Ccr7*, *Sell*, *Gpr183*, and other known markers including those associated with acute and chronic progenitor cell states such as *Id3* [64] and *Xcl1* [13]. The effector module contained edges between *S1pr5*, *Zeb2*, *Gzma*, *Gzmk*, and others, and the exhaustion module contained edges between *Pdcd1*, *Cxcr6*, *Lag3*, *Tigit*, and others. Protein-protein interaction analysis with STRINGdb [53] revealed that all modules showed significantly higher interaction than expected for a random set of proteins of the same size and degree distribution ($p < 1.0e$-16 for all modules), suggesting each gene module represents a biological signature validated in prior literature.

Finally, to further test the utility of gene-gene k-NN graphs built from the gene representations, we compared the gene-gene graphs for a negative control of Acute Day 8 cells versus cells from a *Klf2* knockout (KO) and a *Tbx21* KO experiment, visualizing only gene-gene interactions that were knocked out in the experiment but existed in the negative control (Figure 4g). This uncovered some interesting gene networks lost in the *Klf2* KO such as those around *Bach2* and *Id2*, which have been implicated in memory and effector differentiation respectively [49, 64]. Interestingly, in the *Tbx21* KO network, *Cd69* and *Batf* appear, which may be reflective of *Tbx21*'s role in enforcing effector differentiation [25]. Such analysis highlights the ability of gene signal pattern analysis to capture gene coordination events in perturbation-specific ways.

2.7 GSPA captures cluster-independent patterns of cell-cell communication in tissue-resident immune cells

One of the key benefits of learning informative gene features is the ability to characterize patterns of gene signals beyond the boundaries of cell states, including identifying signatures within or between cell states. This is especially useful for characterizing patterns of cellular communication, as these interactions can involve multiple cell states or a small subset of cells that are responding strongly to a stimulus. In canonical cell-cell communication analysis, which calculates an enrichment score based on how specific an interaction is between two cell states, both of these types of interactions are not likely to be prioritized [4].

Here, we describe how to leverage gene signal pattern analysis to identify patterns of ligand-receptor communication. We employ the CellChatDB [24] intercellular communication database, which contains triplets (ligand, receptor, pathway), where the ligand is the source gene, receptor is the target gene, and pathway is an attribute defining to which communication pathway the interaction belongs. Then, we run GSPA+QR on the ligands and receptors separately, and we concatenate the embeddings to learn a representation for every pair. We do this because two communication patterns a and b should be represented similarly if ligand_a and ligand_b show similar expression profiles and receptor_a and receptor_b show similar expression profiles. So, mapping the concatenated representations identifies groups of ligand-receptor (LR) pairs with shared patterning on the cellular manifold across and within cell types. Finally, to further summarize these relationships, we build a k-NN graph of the LR pair representations, and then we consider each pathway attribute as an indicator signal on the LR graph, where the signal is 1 if the LR pair belongs to that pathway and 0 if not. Using GSPA, we map the pathway space (Figure 5a).

We tested this approach in a recently published work examining the role of PD-1 in skin tolerance and potential mechanisms of checkpoint inhibitor-driven immune-related adverse events [14]. The data supported the hypothesis that PD-1 expression by myeloid cells functions to maintain healthy tissue through interaction with local CD8+ T cells, and PD-1-blocking immunotherapies result in adverse events. Given this interaction was specific to antigen expression and functionally validated, we sought to recover the PD-1/PD-L1 interaction axis using our pipeline.

The scRNA-seq dataset from this work comprises 21,178 skin cells from three conditions: antigen off (NO_AG), antigen-expressing (AG), and antigen-expressing treated with checkpoint inhibitors (AG_CPI) (Figure 5b). This dataset was previously annotated with four cell types (Figure 5c), where a subset of myeloid cells are enriched for PD-L1, and a subset of CD8+ T cells are enriched for PD-1 (Figure 5d). As

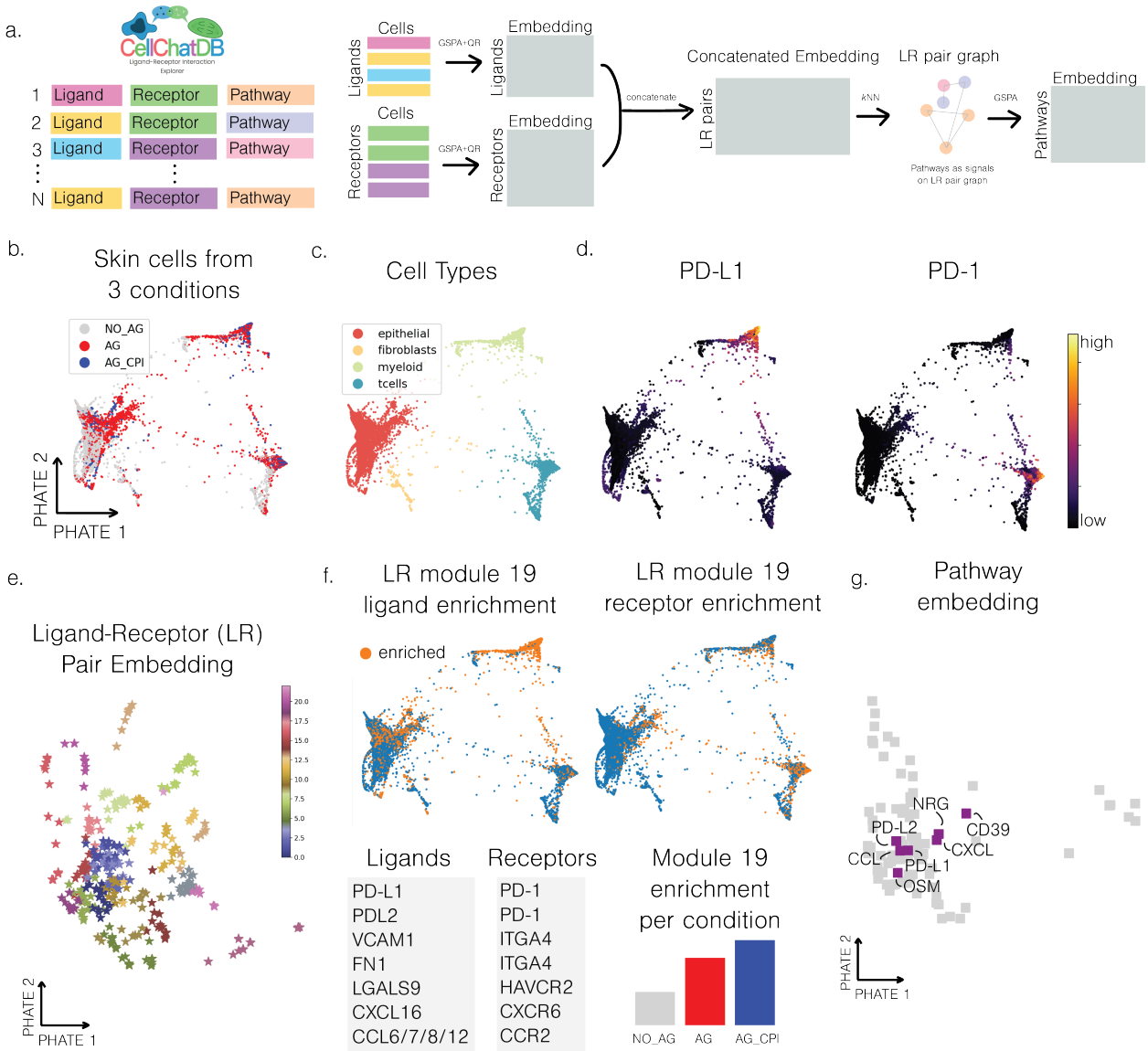


Figure 5. Identification of patterns of ligand-receptor communication in peripheral tolerance skin model. a. Schematic of cluster-independent cell-cell communication pipeline. b. Skin cells from no antigen (AG), antigen (AG), and antigen with checkpoint inhibitor (AG_CPI) conditions visualized with PHATE. c. Skin cells colored by previously annotated cell types. d. Skin cells colored by PD-L1 and PD-1. e. Ligand-receptor pair embedding visualized with PHATE. f. Cell embedding colored by enrichment for ligands and receptors in LR pair module 19 (top), list of ligands and receptors, and enrichment per condition (bottom). g. Pathway embedding visualized with PHATE.

this dataset showed distinct cell types, we calculated cell type association scores across the gene space for each cell type. The gene embedding showed distinct regions of the gene space enriched for distinct cell types (Extended Data Figure 9a). Further, the top genes most closely associated with each cell type were strongly and specifically enriched (Extended Data Figure 9b).

Then, ignoring the cell type labels to identify communication events within and between cell types, we run our communication pipeline, then visualize and cluster LR pairs into LR modules. This reveals a diverse range of ligand and receptor profiles (Figure 5e, Extended Data Table 3, Extended Data Figure 10). In one LR module, the average ligand expression was higher in a subset of myeloid cells derived from the AG

and AG_CPI conditions, and the receptor expression was higher in a subset of T cells derived from the AG and AG_CPI conditions. This expression pattern matched the known PD-L1/PD-1 pattern specific to antigen expression, and this module included *Cd274* and *Pdcd1* as well as many other checkpoint inhibitory and chemokine signaling pathways, including *Pdcdlg2-Pdcd1*, *Cxcl16-Cxcr6*, and *Lgals9-Havcr2* (Figure 5f). Other gene modules showed interesting patterns, such as fibroblasts communicating to all other cell types (module 13) and multiple cell types communicating to antigen-specific CD8+ T cells, which includes axes related to T cell activation and exhaustion such as *Cd80-Cd28*, *Cd86-Cd28*, and *Pvr-Tigit* (Extended Data Figure 10). Finally, we learn a pathway embedding and identify the colocalization of several related pathways, including those related to PD-L1/2 and T cell cytokine and chemokine secretion (Figure 5g)). GSPA was therefore able to reveal both pair-pair and pathway-pathway relationships, demonstrating the ability to perform multiscale gene signaling analysis.

2.8 Spatially-colocalized gene modules correspond to key communication hubs

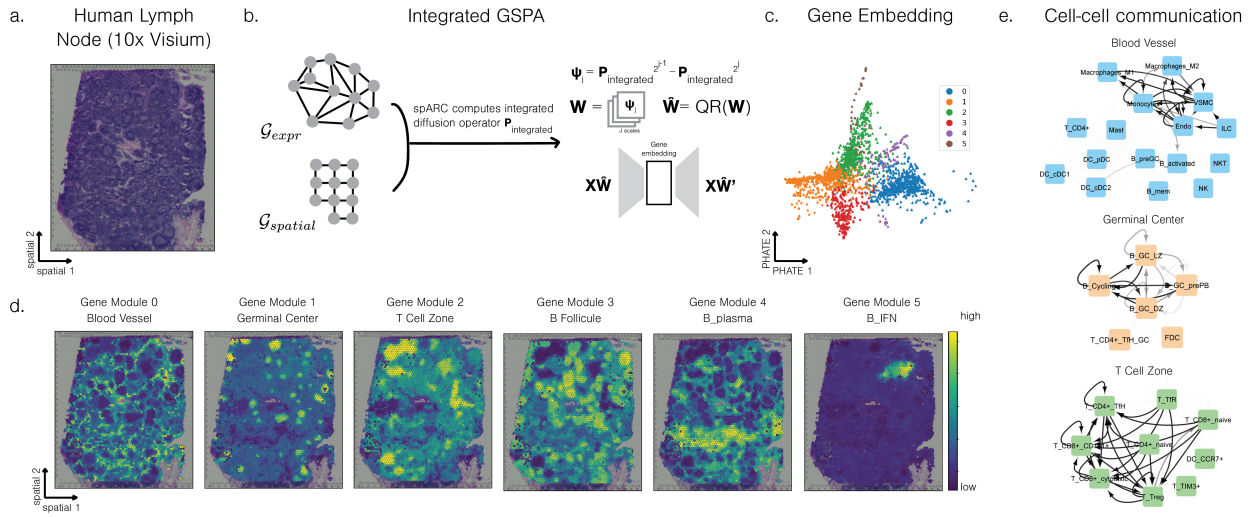


Figure 6. Gene embeddings capture communication pathways from 10x Visium human lymph node. a. H&E stain of human lymph node tissue. b. Schematic of integrated GSPA using spARC. c. PHATE visualization of gene embedding, colored by gene module assignment. d. Enrichment of gene modules spatially. e. Cell-cell communication networks derived from gene-gene interactions with OmniPathDB.

Gene signal pattern analysis can also derive feature embeddings for data modalities beyond scRNA-seq. Here, we demonstrate the capability of GSPA+QR on a spatial RNA-sequencing dataset of a healthy human lymph node from 10x Visium (Figure 6a) [1]. Spatially-resolved transcriptomics is a powerful tool for adding additional organizational context to gene expression analysis. Biological systems such as the lymph node contain many highly active microenvironments and diverse cell types corresponding to spatially distinct substructures and unique organization. However, like scRNA-seq data, most spatial technologies are accompanied by a large amount of technical noise and artifact [32]. This obfuscates the ability to assess gene patterns and relationships.

To this end, we leverage spARC [32] to compute an integrated diffusion operator $\mathbf{P}_{integrated}$ from both gene expression measurements at each spot $\mathcal{G}_{expr}$ and the spot spatial coordinate grid graph $\mathcal{G}_{spatial}$. We then calculate an integrated wavelet dictionary from this operator (Figure 6b). The resulting gene embeddings are thus informed by both the expression and spatial relationships between spots (Figure 6c). We cluster these gene embeddings and visualize the mean expression of each gene module on the spatial coordinates, highlighting the spatial localization of each signature (Figure 6d).

Notably, while there is a large variety of cell types, this data is not at cellular resolution; each spot contains on average 10 cells. Therefore, to aid in characterizing cell types enriched across the tissue, we

leveraged an integrated atlas of human secondary lymphoid organs composed of 73,260 cells from [28] and used the cell2location [28] mapping to determine the cell types enriched in each gene module. This reveals enrichment of monocytes, macrophages, mast cells, endothelial cells, FDCs, and VSMCs in gene module 0, which spatially corresponds to the blood vessel; GC-committed B cells, T follicular helper cells and follicular dendritic cells in gene module 1, enriched in the germinal centers of the tissue; T cells in gene module 2; non-GC-specific B cells in gene module 3, enriched in the B follicule; B plasma cells in gene module 4; and B_IFN cells in gene module 5 (Extended Data Figure 11).

As each spot contains cells from different cell types with a high degree of cell-cell interaction and organization, we hypothesized that we could identify gene signaling corresponding to communication events. We thus leveraged the OmniPathDB [57], a knowledge graph containing curated intercellular and intracellular gene-gene signaling interactions. For each gene module, we built a k-NN graph from the gene embeddings, and we pruned the graphs to include only edges that exist in the OmniPathDB network. As OmniPathDB edges are directional, this also allows us to impute directionality to the gene-gene graph. Then, we sought to identify the cell types involved in the gene-gene signaling events. We repeat the following for each gene module graph: for each directed edge ($gene_s$, $gene_t$), for all pairs of cell states ($celltype_a$, $celltype_b$), if $gene_s$ is differentially expressed in $celltype_a$, and $gene_t$ is differentially expressed in $celltype_b$, we add a directed edge from $celltype_a$ to $celltype_b$. For gene modules with multiple cell types, this revealed complex communication patterns (intercellular edges in blue, intracellular edges in red). Endothelial cells, VSMCs, macrophages, and monocytes showed strong signaling in gene module 0. B_cycling cells and other GC-committed B cells were communicating in gene module 1. Finally, T cells showed highly active communication in gene module 2 (Figure 6e). In sum, GSPA enabled spatial analysis of dynamic and active cell-cell communication characterizing the human lymph node.

2.9 Multiscale data features map patient manifolds and predict outcomes.

Importantly, the learned gene representations are highly informative and can be used for tasks beyond gene space mapping and characterization. Here, we show that gene features can be used to map manifolds of patient samples for highly interpretable and top-performing prediction. 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. However, because we know which cells are from which patient, 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 patient 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 (Figure 7a).

We test this approach in a dataset of 48 melanoma samples pre- and post-checkpoint immunotherapy [45]. Despite the successes of checkpoint immunotherapy in melanoma, many patients fail to show a response. The authors show that the immune cells in the tumor microenvironment play a key predictive role in distinguishing responders and non-responders. Therefore, we first sought to map each patient sample, corresponding to a scRNA-seq dataset of sorted and sequenced CD45+ cells. Then, we evaluate the classification of response based on this representation.

For comparison, we learn patient sample embeddings in four other ways. First, we build the combined cell-cell graph $\mathcal{G}_{cell}$ and consider the patients as indicator signals on the cells derived from each patient sample. We then use GSPA to represent patient signals. Second, we compute the mean expression across all cells through aggregation for each patient. Third, as the cells were clustered and annotated in the original work, we compute the proportion of each cell type and define patient features based on these proportions, similar to [9]. Finally, we do the same for each cell subtype within CD8+ T cells, which were previously shown to be highly predictive of response.

Visualization of the patient embeddings shows that the GSPA+QR gene embeddings and GSPA patient signals separate responders and non-responders, whereas the mean expression and proportion of cell states and substates do not do this as distinctly (Figure 7b). Additionally, classifying response based on the GSPA+QR embeddings showed the highest performance, followed by the mean expression embeddings (Figure 7c). Interestingly, though the gene-based embeddings outperformed the cell-based embeddings, GSPA on patient signals outperformed the cluster proportion representations, suggesting that learning multiscale features beyond clusters can improve response prediction.

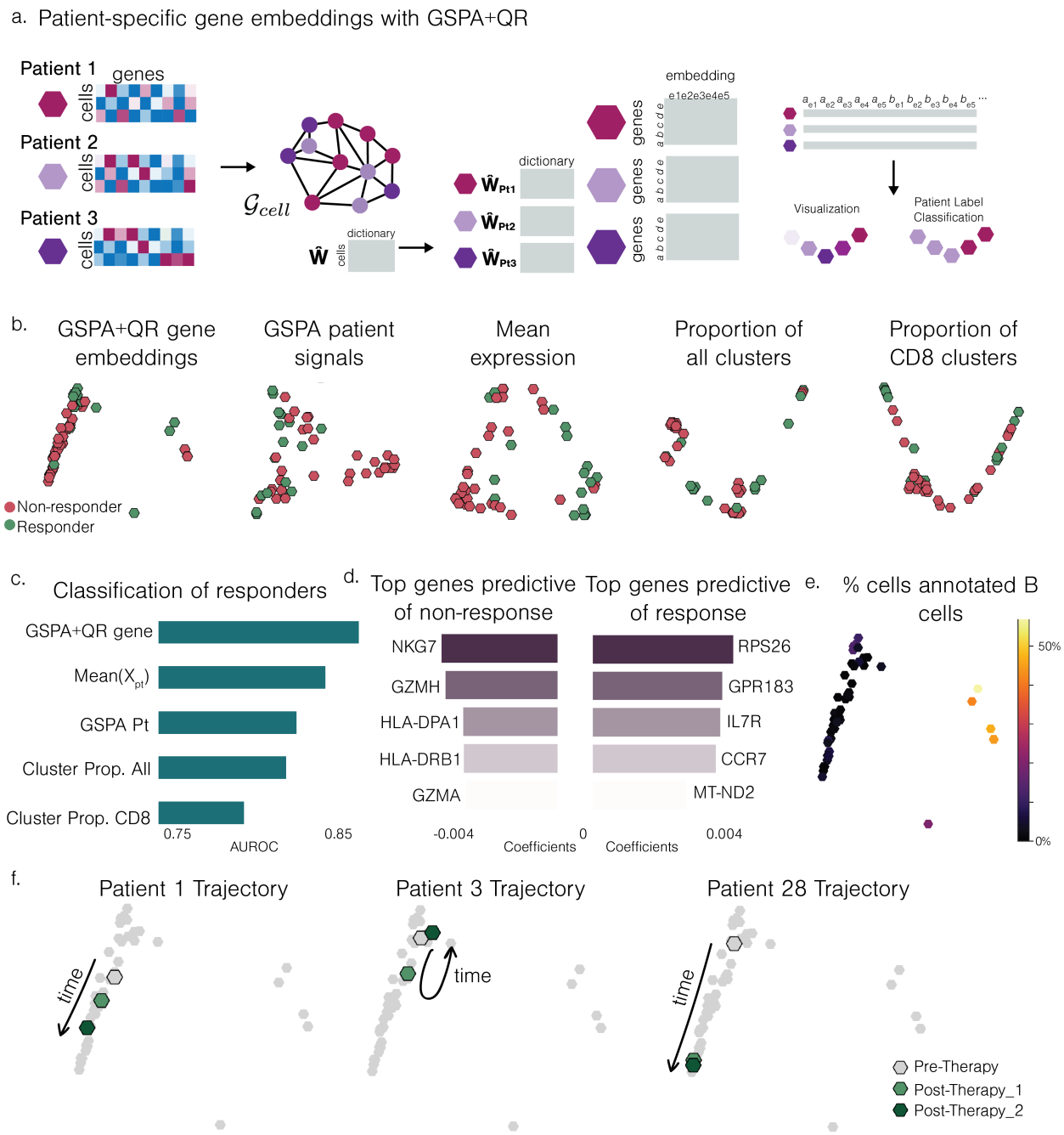


Figure 7. Patient manifold reveals response trajectories and biomarkers related to response prediction. a. Schematic of the derivation of patient-specific gene embeddings. b. PHATE visualization of patient embeddings based on GSPA+QR gene embeddings and comparisons. c. AUROC evaluation of response classification (logistic regression). d. Top genes predictive of response and non-response based on highest and lowest logistic regression coefficients. e. Trajectories of three patients with multiple samples obtained.

The authors showed in the original work that *Tcf7* was a key predictor of positive clinical outcome. We hypothesized that there were additional relevant genes to positive and negative clinical outcome. Because the patient embeddings are comprised of gene features, the coefficients of the logistic regression classification reflect the importance of different genes for predicting response and non-response (Extended Data Table 4).

Genes with the lowest coefficients for response include genes linked to T cell terminal differentiation (e.g. *Gzma*, *Cd38*). Genes with the highest coefficients for response include genes linked to T cell progenitor states such as memory, activation, and survival (e.g. *Il7r*, *Tcf7*) (Figure 7d). This reflects our understanding of the progenitor T cell state mediating the response and being the primary target of immunotherapy, as well as the inability of terminally differentiated cells to reacquire significant function [30, 48, 67].

While the original work achieved strong classification based on T cell signatures, building a patient manifold also provides information beyond T cell signatures. Five patient samples were visually separated from the remaining samples. Comparing the proportion of G1, the cluster annotated as majorly B cells, reveals that these patients' cell samples are over 40% G1, far more than the average B cell proportion of the other patient samples (Figure 7e). Furthermore, patient manifolds allow us to analyze the overall shape of the patient space, which is useful for predicting or characterizing patient trajectories. In this dataset, some patients had samples taken pre- and post-immunotherapy. The samples from patient 1 reflect a trajectory towards non-response over the disease progression, whereas the samples from patient 3 remained in the responder region of the manifold. Meanwhile, the samples for patient 28 showed a striking change from response to non-response after therapy, with little difference between the first and second post-therapy samples (Figure 7f). Together, this highlights the utility of interpretable and predictive patient representations for the analysis of clinical outcomes.

3 Discussion

Though there has been much interest in characterizing gene-gene relationships, mapping the gene space has not been sufficiently investigated and motivated. In this work, we defined relevant comparisons from graph representation learning and graph signal processing, and we built experiments with three simulated datasets to evaluate gene embeddings, establishing the groundwork for future research in this space. We demonstrated the utility of framing genes as signals on the cell-cell graph, and we described our method of decomposing gene signals using diffusion wavelets, then reducing the dimensionality using an autoencoder framework. We also highlighted the superior ability of this technique to preserve gene-gene relationships, capture structure in the visualization, and reveal salient gene modules.

Additionally, we presented two new gene ranking approaches facilitated by this analysis. For datasets with clear cluster-like structure, cell type association rankings allow us to identify genes associated with each cell type. Alternatively, differential localization prioritizes genes that are "localized", or specific to key regions on the cellular manifold, without any assumptions about its shape. Both of these rankings have broadly useful applications to single-cell analysis, which primarily identifies important genes through statistical differences in mean expression between clusters. Usefully, by bridging localization and gene embeddings, we can identify the top genes most relevant to each gene signature separately. While there exist feature selection approaches for single-cell analysis [46, 60], it is not guaranteed that the top genes will display a diversity in gene trends or enrichment in different cell states. In our method, users can ensure diversity by choosing the top localized genes from each gene module.

We demonstrate the clear and broad utility of gene embeddings through the characterization of different datasets in several vignettes. In a scRNA-seq dataset of CD8+ T cells in acute and chronic infection at three timepoints, we reveal key gene modules that correspond to enrichment at different points of the T cell differentiation process. We also analyze gene-gene relationships through network analysis of a single-cell CRISPR screen, identifying gene-gene coexpression pairs that were knocked out in the context of *Tbx21* and *Klf2*. We then analyzed ligand-receptor and pathway-pathway relationships without cell type annotation in a peripheral tolerance model [14], recovering condition-specific communication between subsets of two cell types. To establish the utility of this approach beyond scRNA-seq data, we presented GSPA for spatial transcriptomic data and identified spatially enriched gene modules and microenvironmental signaling events within each module. Finally, we leveraged GSPA to map single-cell datasets from patients with melanoma, using the gene space to learn an interpretable representation and prediction of patient response.

This work opens up new avenues for further exploration. Beyond gene clustering, future work can explore alternative ways of characterizing gene-gene relationships from the low-dimensional embedding, such as gene trajectories [44] or gene archetypes. Such analysis may reveal continuous dynamics and related gene signatures that cannot be resolved by cell trajectory or archetype-based approaches. Also, one could consider patterns of gene signals beyond localization; gene signals that are broadly pervasive and distributed throughout the

cellular state space may be useful for identifying key marker genes for the system or housekeeping genes for future experiments. Calculating gene distance to other synthetic signals, such as cell pseudotime or a condition-specific signal could also reveal new insights into the particular biological system being explored. The gene signal pattern analysis architecture is also flexible and can be combined with supervised losses for jointly-trained gene embeddings to predict treatment response or cell state.

Importantly, this work also represents a key contribution to graph feature or signal representation learning, and future work can leverage ideas presented here for a variety of graphs and signals. For example, localization scores can be used to characterize signals that are highly specific, which indicates specialization or commitment to a particular group of nodes. We thus expect numerous biological and non-biological data types to benefit from GSPA to extract rich feature representations from large-scale high-dimensional data.

4 Data availability

The accession codes to the data used for figure 4 will be provided before publication. scRNA-seq datasets for figure 5 are available from the Gene Expression Omnibus (GEO) database under accession number GSE228586. Visium data for figure 6 can be downloaded from the 10x website (via a function in the scanpy package). scRNA-seq data for figure 7 are accessible through accession number GSE120575.

5 Code availability

The source code is available at https://github.com/KrishnaswamyLab/Gene-Signal-Pattern-Analysis.

References

1. V1 Human Lymph Node, Spatial Gene Expression Dataset by Space Ranger 1.1.0, 10x Genomics, (2023, August 29).
2. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z.K., Wouters, J., Aerts, S.: SCENIC: single-cell regulatory network inference and clustering. Nat. Methods **14**(11), 1083–1086 (Nov 2017)
3. Amir, E.A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., Pe'er, D.: viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. **31**(6), 545–552 (Jun 2013)
4. Armingol, E., Officer, A., Harismendy, O., Lewis, N.E.: Deciphering cell-cell interactions and communication from gene expression. Nat. Rev. Genet. **22**(2), 71–88 (Feb 2021)
5. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. **37**(1), 38–44 (Dec 2018)
6. Bellazzi, R., Codegoni, A., Gualandi, S., Nicora, G., Vercesi, E.: The gene mover's distance: Single-cell similarity via optimal transport (2021)
7. Burkhardt, D.B., Stanley, 3rd, J.S., Tong, A., Perdigoto, A.L., Gigante, S.A., Herold, K.C., Wolf, G., Giraldez, A.J., van Dijk, D., Krishnaswamy, S.: Quantifying the effect of experimental perturbations at single-cell resolution. Nat. Biotechnol. **39**(5), 619–629 (May 2021)
8. Cazares, T.A., Rizvi, F.W., Iyer, B., Chen, X., Kotliar, M., Bejjani, A.T., Wayman, J.A., Donmez, O., Wronowski, B., Parameswaran, S., Kottyan, L.C., Barski, A., Weirauch, M.T., Prasath, V.B.S., Miraldi, E.R.: maxATAC: Genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks. PLoS Comput. Biol. **19**(1), e1010863 (Jan 2023)
9. Chen, W.S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., Krishnaswamy, S.: Uncovering axes of variation among single-cell cancer specimens. Nature Methods **17**(3), 302–310 (Mar 2020). https://doi.org/10.1038/s41592-019-0689-z, `https://doi.org/10.1038/s41592-019-0689-z`
10. Chung, H.K., McDonald, B., Kaech, S.M.: The architectural design of CD8+ T cell responses in acute and chronic infection: Parallel structures with divergent fates. J. Exp. Med. **218**(4) (Apr 2021)
11. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (Jul 2006)
12. Coifman, R.R., Maggioni, M.: Diffusion wavelets. Applied and Computational Harmonic Analysis **21**(1), 53–94 (Jul 2006). https://doi.org/10.1016/j.acha.2006.04.004, `https://doi.org/10.1016/j.acha.2006.04.004`
13. Connolly, K.A., Kuchroo, M., Venkat, A., Khatun, A., Wang, J., William, I., Hornick, N.I., Fitzgerald, B.L., Damo, M., Kasmani, M.Y., Cui, C., Fagerberg, E., Monroy, I., Hutchins, A., Cheung, J.F., Foster, G.G., Mariuzza, D.L., Nader, M., Zhao, H., Cui, W., Krishnaswamy, S., Joshi, N.S.: A reservoir of stem-like CD8+ T cells in the tumor-draining lymph node preserves the ongoing antitumor immune response. Sci. Immunol. **6**(64), eabg7836 (Oct 2021)
14. Damo, M., Hornick, N.I., Venkat, A., William, I., Clulo, K., Venkatesan, S., He, J., Fagerberg, E., Loza, J.L., Kwok, D., Tal, A., Buck, J., Cui, C., Singh, J., Damsky, W.E., Leventhal, J.S., Krishnaswamy, S., Joshi, N.S.: PD-1 maintains CD8 t cell tolerance towards cutaneous neoantigens. Nature **619**(7968), 151–159 (Jun 2023). https://doi.org/10.1038/s41586-023-06217-y, `https://doi.org/10.1038/s41586-023-06217-y`
15. Daniel, B., Yost, K.E., Hsiung, S., Sandor, K., Xia, Y., Qi, Y., Hiam-Galvez, K.J., Black, M., J Raposo, C., Shi, Q., Meier, S.L., Belk, J.A., Giles, J.R., Wherry, E.J., Chang, H.Y., Egawa, T., Satpathy, A.T.: Divergent clonal differentiation trajectories of T cell exhaustion. Nat. Immunol. **23**(11), 1614–1627 (Nov 2022)
16. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., Pe'er, D.: Recovering gene interactions from single-cell data using data diffusion. Cell **174**(3), 716–729.e27 (Jul 2018)
17. Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., Zhi, D.: Gene2vec: distributed representation of genes based on co-expression. BMC Genomics **20**(Suppl 1), 82 (Feb 2019)
18. Farahbod, M., Pavlidis, P.: Untangling the effects of cellular composition on coexpression analysis. Genome Res. **30**(6), 849–859 (Jun 2020)
19. F.R.S., K.P.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720
20. Giles, J.R., Ngiow, S.F., Manne, S., Baxter, A.E., Khan, O., Wang, P., Staupe, R., Abdel-Hakeem, M.S., Huang, H., Mathew, D., Painter, M.M., Wu, J.E., Huang, Y.J., Goel, R.R., Yan, P.K., Karakousis, G.C., Xu, X., Mitchell, T.C., Huang, A.C., Wherry, E.J.: Shared and distinct biological circuits in effector, memory and exhausted CD8+ T cells revealed by temporal single-cell transcriptomics and epigenetics. Nat. Immunol. **23**(11), 1600–1613 (Nov 2022)
21. Grayson, J.M., Zajac, A.J., Altman, J.D., Ahmed, R.: Cutting edge: increased expression of bcl-2 in antigen-specific memory CD8+ T cells. J. Immunol. **164**(8), 3950–3954 (Apr 2000)

22. Grover, A., Leskovec, J.: 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)
23. Hoekzema, R.S., Marsh, L., Sumray, O., Carroll, T.M., Lu, X., Byrne, H.M., Harrington, H.A.: Multiscale methods for signal selection in single-cell data. Entropy (Basel) **24**(8), 1116 (Aug 2022)
24. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.H., Myung, P., Plikus, M.V., Nie, Q.: Inference and analysis of cell-cell communication using CellChat. Nat. Commun. **12**(1), 1088 (Feb 2021)
25. Joshi, N.S., Cui, W., Chandele, A., Lee, H.K., Urso, D.R., Hagman, J., Gapin, L., Kaech, S.M.: Inflammation directs memory precursor and short-lived effector CD8(+) T cell fates via the graded expression of t-bet transcription factor. Immunity **27**(2), 281–295 (Aug 2007)
26. Kipf, T.N., Welling, M.: Variational graph auto-encoders. CoRR **abs/1611.07308** (2016), `http://arxiv.org/abs/1611.07308`
27. Kiselev, V.Y., Andrews, T.S., Hemberg, M.: Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. **20**(5), 273–282 (May 2019)
28. Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M.S., Park, J.S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., Bayraktar, O.A.: Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. **40**(5), 661–671 (May 2022)
29. Knapp, T.R.: Canonical correlation analysis: A general parametric significance-testing system. Psychological Bulletin **85**(2), 410–416 (Mar 1978). https://doi.org/10.1037/0033-2909.85.2.410, `https://doi.org/10.1037/0033-2909.85.2.410`
30. Krishna, S., Lowery, F.J., Copeland, A.R., Bahadiroglu, E., Mukherjee, R., Jia, L., Anibal, J.T., Sachs, A., Adebola, S.O., Gurusamy, D., Yu, Z., Hill, V., Gartner, J.J., Li, Y.F., Parkhurst, M., Paria, B., Kvistborg, P., Kelly, M.C., Goff, S.L., Altan-Bonnet, G., Robbins, P.F., Rosenberg, S.A.: Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. Science **370**(6522), 1328–1334 (Dec 2020)
31. Kuchroo, M., Huang, J., Wong, P., Grenier, J.C., Shung, D., Tong, A., Lucas, C., Klein, J., Burkhardt, D.B., Gigante, S., Godavarthi, A., Rieck, B., Israelow, B., Simonov, M., Mao, T., Oh, J.E., Silva, J., Takahashi, T., Odio, C.D., Casanovas-Massana, A., Fournier, J., Yale IMPACT Team, Farhadian, S., Dela Cruz, C.S., Ko, A.I., Hirn, M.J., Wilson, F.P., Hussin, J.G., Wolf, G., Iwasaki, A., Krishnaswamy, S.: Multiscale PHATE identifies multimodal signatures of COVID-19. Nat. Biotechnol. **40**(5), 681–691 (May 2022)
32. Kuchroo, M., Miyagishima, D.F., Steach, H.R., Godavarthi, A., Takeo, Y., Duy, P.Q., Barak, T., Erson-Omay, E.Z., Youlten, S., Mishra-Gorur, K., Moliterno, J., McGuone, D., Günel, M., Krishnaswamy, S.: spARC recovers human glioma spatial signaling networks with graph filtering (Aug 2022)
33. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B.d., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J.d., Saliba, A.E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A.: Eleven grand challenges in single-cell data science. Genome Biol. **21**(1), 31 (Feb 2020)
34. Langfelder, P., Horvath, S.: Eigengene networks for studying the relationships between co-expression modules. BMC Syst. Biol. **1**(1), 54 (Nov 2007)
35. Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**(1), 559 (Dec 2008)
36. Leone, S., Venkat, A., Huguet, G., Tong, A., Wolf, G., Krishnaswamy, S.: Graph fourier MMD for signals on graphs. SAMPTA (2023)
37. Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. **15**(6), e8746 (Jun 2019)
38. McGinnis, C.S., Murrow, L.M., Gartner, Z.J.: DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Systems **8**(4), 329–337.e4 (Apr 2019). https://doi.org/10.1016/j.cels.2019.03.003, `https://doi.org/10.1016/j.cels.2019.03.003`
39. McNab, F., Mayer-Barber, K., Sher, A., Wack, A., O'Garra, A.: Type I interferons in infectious disease. Nat. Rev. Immunol. **15**(2), 87–103 (Feb 2015)
40. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., van den Elzen, A., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S.: Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. **37**(12), 1482–1492 (Dec 2019)
41. Moon, K.R., Stanley, III, J.S., Burkhardt, D., van Dijk, D., Wolf, G., Krishnaswamy, S.: Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr. Opin. Syst. Biol. **7**, 36–46 (Feb 2018)
42. Ortega, A., Frossard, P., Kovačević, J., Moura, J.M.F., Vandergheynst, P.: Graph signal processing: Overview, challenges and applications (2017)

43. Perlmutter, M., Tong, A., Gao, F., Wolf, G., Hirn, M.: Understanding graph neural networks with generalized geometric scattering transforms. SIAM Journal on Mathematics of Data Science **5**(4), 873–898 (2023). https://doi.org/10.1137/21M1465056, https://doi.org/10.1137/21M1465056
44. Qu, R., Cheng, X., Sefik, E., Platt, S., Garritano, J., Odell, I., Flavell, R.A., Myung, P., Kluger, Y.: Gene trajectory inference for single-cell data by optimal transport metrics. bioRxiv (Jul 2022)
45. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., Freeman, S.S., Reuben, A., Hoover, P.J., Villani, A.C., Ivanova, E., Portell, A., Lizotte, P.H., Aref, A.R., Eliane, J.P., Hammond, M.R., Vitzthum, H., Blackmon, S.M., Li, B., Gopalakrishnan, V., Reddy, S.M., Cooper, Z.A., Paweletz, C.P., Barbie, D.A., Stemmer-Rachamimov, A., Flaherty, K.T., Wargo, J.A., Boland, G.M., Sullivan, R.J., Getz, G., Hacohen, N.: Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell **175**(4), 998–1013.e20 (Nov 2018)
46. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature Biotechnology **33**(5), 495–502 (Apr 2015). https://doi.org/10.1038/nbt.3192, https://doi.org/10.1038/nbt.3192
47. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (Nov 2003)
48. Siddiqui, I., Schaeuble, K., Chennupati, V., Fuertes Marraco, S.A., Calderon-Copete, S., Pais Ferreira, D., Carmona, S.J., Scarpellino, L., Gfeller, D., Pradervand, S., Luther, S.A., Speiser, D.E., Held, W.: Intratumoral Tcf1+PD-1+CD8+ T cells with stem-like properties promote tumor control in response to vaccination and checkpoint blockade immunotherapy. Immunity **50**(1), 195–211.e10 (Jan 2019)
49. Sidwell, T., Kallies, A.: Bach2 is required for B cell and T cell memory differentiation. Nat. Immunol. **17**(7), 744–745 (Jul 2016)
50. Song, D., Li, K., Ge, X., Li, J.J.: ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping (Jul 2023). https://doi.org/10.1101/2023.07.21.550107, https://doi.org/10.1101/2023.07.21.550107
51. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. Cell **177**(7), 1888–1902.e21 (Jun 2019). https://doi.org/10.1016/j.cell.2019.05.031, https://doi.org/10.1016/j.cell.2019.05.031
52. Sumida, T.S., Dulberg, S., Schupp, J.C., Lincoln, M.R., Stillwell, H.A., Axisa, P.P., Comi, M., Unterman, A., Kaminski, N., Madi, A., Kuchroo, V.K., Hafler, D.A.: Type I interferon transcriptional network regulates expression of coinhibitory receptors in human T cells. Nat. Immunol. **23**(4), 632–642 (Apr 2022)
53. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., von Mering, C.: STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research **43**(D1), D447–D452 (Oct 2014). https://doi.org/10.1093/nar/gku1003, https://doi.org/10.1093/nar/gku1003
54. Tong, A., Huguet, G., Shung, D., Natik, A., Kuchroo, M., Lajoie, G., Wolf, G., Krishnaswamy, S.: 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). https://doi.org/10.1109/ICASSP43922.2022.9746556
55. Tong, A.Y., Huguet, G., Natik, A., Macdonald, K., Kuchroo, M., Coifman, R., Wolf, G., Krishnaswamy, S.: Diffusion earth mover's distance and distribution embeddings. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10336–10346. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/tong21a.html
56. Traag, V.A., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Scientific Reports **9**(1) (Mar 2019). https://doi.org/10.1038/s41598-019-41695-z, https://doi.org/10.1038/s41598-019-41695-z
57. Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T., Saez-Rodriguez, J.: Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol. Syst. Biol. **17**(3), e9923 (Mar 2021)
58. van Unen, V., Höllt, T., Pezzotti, N., Li, N., Reinders, M.J.T., Eisemann, E., Koning, F., Vilanova, A., Lelieveldt, B.P.F.: Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nat. Commun. **8**(1) (Dec 2017)
59. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative. J Mach Learn Res **10**(66-71), 13 (2009)
60. Vandenbon, A., Diez, D.: A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. Nat. Commun. **11**(1), 4318 (Aug 2020)
61. Venkat, A., Bhaskar, D., Krishnaswamy, S.: Multiscale geometric and topological analyses for characterizing and predicting immune responses from single cell data. Trends Immunol. **44**(7), 551–563 (Jul 2023)

62. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biology **19**(1) (Feb 2018). https://doi.org/10.1186/s13059-017-1382-0, https://doi.org/10.1186/s13059-017-1382-0
63. Wu, T., Ji, Y., Moseman, E.A., Xu, H.C., Manglani, M., Kirby, M., Anderson, S.M., Handon, R., Kenyon, E., Elkahloun, A., Wu, W., Lang, P.A., Gattinoni, L., McGavern, D.B., Schwartzberg, P.L.: The TCF1-Bcl6 axis counteracts type I interferon to repress exhaustion and maintain T cell stemness. Sci. Immunol. **1**(6), eaai8593–eaai8593 (Dec 2016)
64. Yang, C.Y., Best, J.A., Knell, J., Yang, E., Sheridan, A.D., Jesionek, A.K., Li, H.S., Rivera, R.R., Lind, K.C., D'Cruz, L.M., Watowich, S.S., Murre, C., Goldrath, A.W.: The transcriptional regulators id2 and id3 control the formation of distinct memory CD8+ T cell subsets. Nat. Immunol. **12**(12), 1221–1229 (Nov 2011)
65. Zapatero, M.R., Tong, A., Sufi, J., Vlckova, P., Rodriguez, F.C., Nattress, C., Qin, X., Hochhauser, D., Krishnaswamy, S., Tape, C.J.: Trellis single-cell screening reveals stromal regulation of patient-derived organoid drug responses (Oct 2022)
66. Zappia, L., Phipson, B., Oshlack, A.: Splatter: simulation of single-cell RNA sequencing data. Genome Biol. **18**(1) (Dec 2017)
67. Zehn, D., Thimme, R., Lugli, E., de Almeida, G.P., Oxenius, A.: 'stem-like' precursors are the fount to sustain persistent CD8+ T cell responses. Nat. Immunol. **23**(6), 836–847 (Jun 2022)
68. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**(1), 14049 (Jan 2017)

6 Methods

6.1 Related Work

Prior literature on examining gene-gene relationships Effective strategies for identifying gene-gene relationships have been employed in bulk RNA sequencing data, including building gene coexpression graphs through gene-gene correlation [34, 35] or mapping from large databases [17]. However, gene-gene relationships resolved in bulk data are acutely sensitive to changes in sample cellular composition [18]. Single-cell transcriptomics avoids the compositional issues faced in bulk data, but constructing networks from these measurements directly is hindered by technical factors. Networks constructed without correcting for high levels of noise in gene expression measurements can be unreliable, as two genes that are related may not be expressed in the same cell. To overcome this problem, some studies first impute or denoise gene expression before calculating coexpression [8, 16], which recovers relationships within a local neighborhood of the graph but does not account for both local and global distances between gene signals. Others have mapped genes to a high-dimensional space for improved prediction, which does not resolve low-dimensional embeddings for genes [6]. Finally, others filter predicted expression networks using additional evidence of interaction, such as the presence of a predefined transcription factor binding site among members of a co-regulated network [2].

Prior literature for patient sample comparison Patient comparison of single-cell samples has historically compared cellular abundance [45]. Some approaches have built upon this idea to represent each patient as a vector of relative cluster proportions, which can then be directly compared and visualized [9, 31]. More recently, there have been methods that go beyond cluster abundances by modeling each patient dataset as a distribution on a common graph, enabling comparison via optimal transport between distributions [55, 65]. Patient embeddings can demonstrate how variation between patients corresponds to response to treatment. This becomes relevant for the development of predictive models of clinical outcomes, which can then be interpreted for the identification of predictive features [61].

6.2 Gene signal pattern analysis (GSPA) detailed overview

GSPA can be described by the following steps explained in detail below:

1. Constructing a cell-cell similarity graph from single-cell data
2. Building a dictionary of graph diffusion wavelets from the cell-cell similarity graph
3. Projecting gene signals onto the wavelet dictionary
4. Learning a low-dimensional representation of gene signal projections with an autoencoder

Constructing a cell-cell similarity graph from single-cell data To build the graph $\mathcal{G}_{cell} = (V_{cell}, E_{cell})$, we compute the Euclidean distances between all pairs of cells and apply an α-decaying kernel to calculate affinities [7, 40]. The α-decaying kernel is defined as

$$K_{k,\alpha}(x, y) = \frac{1}{2} \exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon_k(x)}\right)^{\alpha}\right) + \frac{1}{2} \exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon_k(y)}\right)^{\alpha}\right), \tag{4}$$

where x and y are cells $\in V_{cell}$, viewed as points in $\mathbb{R}^m$ corresponding to columns of $\mathbf{X}$, $\varepsilon_k(x), \varepsilon_k(y)$ are the distance from x, y to their k-th nearest neighbors, respectively, and α is a parameter that controls the decay rate (i.e., heaviness of the tails) of the kernel. This construction generalizes the popular Gaussian kernel typically used in manifold learning while addressing some of its key limitations, as explained in [41].

This defines a fully connected and weighted graph between cells such that the connection between cells x and y is given by $K(x, y)$. To increase computational efficiency, we sparsify the graph by setting very small edge weights (i.e., $\leq$1e-4) to 0. Additionally, $\mathbf{A}$ is defined as the adjacency / affinity matrix of $\mathcal{G}_{cell}$, and $\mathbf{D}$ is the diagonal degree matrix of $\mathcal{G}_{cell}$.

Building dictionary of graph diffusion wavelets for gene representation In classical signal processing for, e.g., the analysis of temporal data, a wavelet dictionary is defined by taking a function ψ and a set of transformations of this function by time and scale. To adapt these methods to graphs, where there is no concept of time or linear space, we center wavelets at vertices on the graph and change scales via diffusion. Given the cellular graph $\mathcal{G}_{cell}$, we define $\mathbf{P} = \mathbf{A}\mathbf{D}^{-1}$ as the diffusion operator. Each wavelet of scale j centered at vertex v can thus be calculated by computing the matrix $\boldsymbol{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j}$ for $1 \leq j \leq J$ (and $\boldsymbol{\Psi}_0 = \mathbf{I} - \mathbf{P}$) and then extracting the v-th row via $\delta_v^T \boldsymbol{\Psi}_j$, where δ_v is the Kronecker delta centered at the v-th vertex. Then $\{\boldsymbol{\Psi}_j^T \delta_v\}_{v \in V_{cell}, j \in 0,1 \ldots J}$ defines our wavelet dictionary $\mathbf{W}$ (where we use $(\delta_v^T \boldsymbol{\Psi}_j)^T = \boldsymbol{\Psi}_j^T \delta_v$ because our projection $\mathbf{X} \to \mathbf{X}\mathbf{W}$ is performed via multiplication on the right). $\mathbf{W}$ is an $n \times Jn$ matrix (every wavelet takes values on the whole vertex set, and we have a wavelet for every vertex at each scale).

Because the diffusion operator $\mathbf{P}$ is smoothing, we make the assumption that the numerical rank decreases as we take powers of the operator [12]. The faster the spectrum of $\mathbf{P}$ decays, the smaller the numerical rank of the powers of $\mathbf{P}$, the more these can be compressed, and the faster the construction of $\boldsymbol{\Psi}_j$ for large j. Therefore, to decrease the size of the wavelet dictionary, remove redundant wavelets, and increase their interpretability, we can perform QR factorization to reveal the low numerical rank of $\boldsymbol{\Psi}_j$ for large j. This results in a compressed wavelet dictionary $\widehat{\mathbf{W}} = \{\widetilde{\boldsymbol{\Psi}}_j^T \delta_v\}_{v \in V, j \in 0,1 \ldots J}$, where for each j, $\widetilde{\boldsymbol{\Psi}}_j$ is a set of linear combinations of wavelets at j that account for the most variance. For large j, $\boldsymbol{\Psi}_j \delta_x \approx \boldsymbol{\Psi}_j \delta_y$ if vertex x is near vertex y on the graph. Thus, we reduce the number of columns in $\boldsymbol{\Psi}_j$ by taking a linear combination to form $\widetilde{\boldsymbol{\Psi}}_j$ such that the total error in projecting $\boldsymbol{\Psi}_j$ onto $\widetilde{\boldsymbol{\Psi}}_j$ is less than some ϵ fraction of the norm of the whole $\boldsymbol{\Psi}_j$. That is,

$$\frac{1}{||\boldsymbol{\Psi}_j||^2} ||\boldsymbol{\Psi}_j - \widetilde{\boldsymbol{\Psi}}_j (\widetilde{\boldsymbol{\Psi}}_j^T \widetilde{\boldsymbol{\Psi}}_j)^{-1} \widetilde{\boldsymbol{\Psi}}_j^T \boldsymbol{\Psi}_j||^2 < \epsilon \tag{5}$$

We test gene signal embeddings both with (GSPA+QR) and without (GSPA) QR factorization.

Projecting gene signals onto the wavelet dictionary Given a gene signal $\mathbf{X}_i$ and wavelet dictionary $\widehat{\mathbf{W}}$ (or $\mathbf{W}$), we project $\mathbf{X}_i$ onto the dictionary, which, for all gene signals, corresponds to $\mathbf{X}\widehat{\mathbf{W}}$ or $\mathbf{X}\mathbf{W}$. Alternatively, the solution can be framed as a set of m inner products provided by $\widehat{\mathbf{W}}\mathbf{X}_i^T$ or $\mathbf{W}\mathbf{X}_i^T$ as a transformation of $\mathbf{X}_i$, or $\widehat{\mathbf{W}}^T\mathbf{X}^T$ or $\mathbf{W}^T\mathbf{X}^T$ for all gene signals. This transformation reveals each gene signal's spatial and frequency information over the corresponding cell-cell graph $\mathcal{G}_{cell}$.

Learning a low-dimensional representation of gene signal projections with an autoencoder $\mathbf{X}\widehat{\mathbf{W}}$ is unnecessarily high-dimensional, where each feature corresponds to the gene signal projection of a particular cell at a particular diffusion scale. To reduce redundancy and improve computational tractability for downstream analysis, we reduce the dimensionality with autoencoder $D \circ E$, where

$$\mathbf{X}\widehat{\mathbf{W}}' \;\approx\; D(E(\mathbf{X}\widehat{\mathbf{W}})), \tag{6}$$

i.e. 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.

$$GSPA(\mathbf{X}) = E(\mathbf{X}\widehat{\mathbf{W}})$$

where $\widehat{\mathbf{W}}$ is the compressed wavelet dictionary, E is the encoder discussed above, and GSPA to taken to be the map Θ discussed in the problem setup.

Wavelet Projection and the Wasserstein distance Earth Movers 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 [54] (see also [55]), it was shown that

the Wasserstein distance (with a truncated geodesic distance as ground distance) could be approximated by the Unbalanced Diffusion Earth Movers Distance (UDEMD) defined below. Here, we will show that the metric induced by our wavelets is continuous with respect to this UDEMD, i.e., $\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \lesssim \mathrm{UDEMD}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$.

In [54], the UDEMD [54] between two signals (genes) $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}$ is defined as

$$\mathrm{UDEMD}_{\beta,J}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) := \sum_{j=0}^{J} \left\| T_{\beta,k}\left(\mathbf{X}_{i_1}\right) - T_{\beta,k}\left(\mathbf{X}_{i_2}\right)\right\|_1, \tag{7}$$

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

$$T_{\beta,j}\left(\mathbf{X}_i\right) := 2^{-(J-j)\beta}\left(\mu_i^{\left(2^j\right)} - \mu_i^{\left(2^{j-1}\right)}\right), \quad \mu_i^{(t)} := \mathbf{P}^t\mathbf{X}_i, \tag{8}$$

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

The following result is a more detailed version of Theorem 1 first introduced in Section 2.3.

Theorem 1. *For $0 < \beta < 1/2$, the diffusion wavelet transform $\mathbf{W}$ (with maximal scale J) is Lipschitz continuous with respect to $UDEMD_{\beta,J}$, 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*

$$\|\mathbf{X}_{i_1}\mathbf{W} - \mathbf{X}_{i_2}\mathbf{W}\|_2 \le C \cdot \ UDEMD_{\beta,J}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}). \tag{1}$$

Proof. Let $\mathbf{X}_{i_1} \neq \mathbf{X}_{i_2}$. (The inequality (1) holds trivially in the case where $\mathbf{X}_{i_1} = \mathbf{X}_{i_2}$.) We may compute

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

where C is a constant depending on J and β. It follows from Proposition 2.2 of [43] that

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

where C is a constant depending only on the ratio between the maximal vertex degree and minimal vertex degree. ([43] 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

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

which in turn implies

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

The lower bound in Proposition 2.2 of [43] 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.

6.3 Computation of cell type association

For datasets where cells organize into distinct clusters, or cell types, gene signal pattern analysis can rank genes based on their specificity to each cell type. We do this by calculating the multiscale representation of an indicator signal $\mathbf{1}_{celltype}(v)$ and computing the affinity between this and each gene signal representation. For each cell type, genes are then ranked, where those that are most cell type-specific are closest to the cell type indicator representation.

Given normalized cell type indicator signal $\mathbf{1}_{celltype}(v) = [x \in celltype]$ for v in V_{cell}, and wavelet representation $\widehat{\mathbf{W}}$, the cell type association score, $c(i)$ for each gene signal $\mathbf{X}_i$ is defined as:

$$c(i) := -\|\mathbf{X}_i\widehat{\mathbf{W}} - \mathbf{1}_{celltype}\widehat{\mathbf{W}}\|_2^2$$

6.4 Computation of differential localization

For many datasets, cells do not organize into discrete cell types, and it can be nontrivial to assign cluster labels and define differentially expressed genes [27]. In fact, defining more flexible paradigms for discovering complex differential gene expression patterns was named a grand challenge in single-cell analysis [33]. There is therefore utility in identifying genes specific to particular populations without prior clustering or annotation. To this end, we naturally extend GSPA to a framework we call *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 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 u and wavelet representation $\widehat{\mathbf{W}}$, is defined as:

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

6.5 Comparison to other gene mapping strategies

Here, we describe in detail each comparison for our experiments with gene signal pattern analysis:

- Raw measurements approach embeds $\mathbf{X}$.
- $\mathrm{GAE}_{\text{no-att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$.
- $\mathrm{GAE}_{\text{att}}(\mathcal{G}_{gene})$ embeds $\mathcal{G}_{gene}$.
- Node2Vec($\mathcal{G}_{gene}$) embeds $\mathcal{G}_{gene}$.
- MAGIC($\mathbf{X}$) embeds X after denoising with $\mathcal{G}_{cell}$.
- DiffusionEMD($\mathbf{X}$, $\mathcal{G}_{cell}$) embeds $\mathbf{X}$ via optimal transport on $\mathcal{G}_{cell}$.
- 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 $\mathcal{G}_{cell}$.

We summarize and diagram these comparisons in Extended Data Figure 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.

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.

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 (Section 2.1), 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 $\mathbf{M}$ powered to t, and left-multiplies $\mathbf{M}^t$ to $\mathbf{X}^T$ as a low-pass filter. For comparison, we left-multiply $\mathbf{X}$ to $\mathbf{M}^t$, which practically denoises gene signals and performs comparatively to MAGIC (data not shown). We then employ an autoencoder $D \circ E$, where $\mathbf{XM}^t \approx D(E(\mathbf{XM}^t))$ and $E(\mathbf{XM}^t)$ is the embedding.

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

Diffusion Earth Mover's Distance (DiffusionEMD) [55] computes optimal transport based on multiscale diffusion kernels. This construction is related to UDEMD [54] described above. 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) [36] 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 $T = 1$. Finally, GeneTrajectory [44] (which did not have code available for comparison) computes optimal transport via $\text{GeneTrajectory}^{(p)}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \min_{F \in \Pi_{\mathbf{X}_{i_1}, \mathbf{X}_{i_2}}} \left\langle F, C^{(p)} \right\rangle^{1/p}$, where the cost matrix C is defined as $C^{(p)}_{x,y} = d_{G_{cell}}(x,y)^p$ and $d_{G_{cell}}$ for $x, y \in V_{cell}$ is the shortest path distance. $\Pi_{\mathbf{X}_{i_1}, \mathbf{X}_{i_2}} = \{F, F_{x,y} \geq 0, \sum_v F_{x,y} = \mathbf{X}_{i_1}(x)$ for all $x, \sum_x F_{x,y} = \mathbf{X}_{i_2}(y)$ for all $y\}$ denotes the set of transport plan F which pushes from the source distribution $\mathbf{X}_{i_1}$ to the target distribution $\mathbf{X}_{i_2}$. To improve speed and simplify the problem for the particular use case of identifying gene trajectories, GeneTrajectory additionally coarse-grains the cell graph into "meta-cells" via clustering and sparsifies the gene-gene graph by zeroing out entries above a certain threshold.

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.

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 [23]. 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), Eigenscore(i) :=

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 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 Eigenscore($\mathbf{X}$) $\approx$ $D(E(\text{Eigenscore}(\mathbf{X})))$ and $E(\text{Eigenscore}(\mathbf{X}))$ is the embedding.

6.6 Diffusion wavelets and comparison to MAGIC

Diffusion wavelets, as compared to diffusion maps [11] and related approaches, such as MAGIC, perform multiscale analysis of graphs and functions on graphs. By representing the cell-cell graph at multiple scales, diffusion wavelets are able to decompose low-frequency and medium-frequency components of a signal, in addition to removing noise, whereas MAGIC acts as a low-pass filter and only maintains the low-frequency components (Extended Data Figure 2).

Additionally, diffusion wavelets represent the local and global geometry of the cell-cell graph, which enables the representation of the distance between signals that are far apart. We demonstrate this with an experiment using a linear simulated trajectory with noise (also used in Figure 2 and Figure 3) (Extended Data Figure 3a). This trajectory defines our cell-cell graph, and we simulate signals on this graph as diracs defined on each cell in the trajectory. Each signal naturally has a label associated with it - the pseudotime value of the cell it was defined on (Extended Data Figure 3b). Notably, as each signal is not overlapping, approaches that do not use the cell-cell graph would not be able to capture meaningful distances between signals.

To test the signal embeddings, we regress the pseudotime label from the latent space (Extended Data Figure 3c). We hypothesized that methods that locally smooth the signal would perform well when signals are close together, as local geometry sufficiently captures pseudotime, and poorly when the signals are far apart. Therefore, we evaluated the prediction with Spearman correlation and increased separation between signals. As expected, MAGIC worsens in the performance as signals are increasingly spread apart, whereas GSPA and GSPA+QR do not show a trend related to the separation between signals (Extended Data Figure 3d). We also perform unsupervised evaluation of the embeddings at spacing 2^6 via the correlation between the Fiedler vector of the signal graph and the pseudotime label. GSPA and GSPA+QR show high correlation versus MAGIC (Extended Data Figure 3e), and also show a qualitative association with the pseudotime label when visualized versus MAGIC (Extended Data Figure 3f).

6.7 Training Details

Default GSPA hyperparameter selection and training details The cell-cell graph was built with PHATE using default parameters (k=5 and α decay=40). The number of scales for the wavelet dictionary J was defined as the *log* of the number of cells N (default). The power was set by default to 2 to mimic the dyadic scales in [12]. For GSPA+QR, the epsilon parameter was set to 1e-3. The autoencoder was designed with 2 layers with bias in the encoder and decoder, with a relu activation function between layers. The latent dimension was set to 128. The models were trained 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.

Comparison hyperparameter and training details For method comparisons in Figure 2 and Figure 3, we ran each method three times, including reconstructing the graph with new seeds, for robust comparison. 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 (and 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 [23], 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.

6.8 Datasets and Pre-processing

Simulated datasets with Splatter Three datasets were simulated using Splatter [66] 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. After simulation, genes expressed in less than 50 cells were removed, and the matrix was L1 normalized for library size and square-root transformed. 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.

Three-timepoint scRNA-seq dataset (previously unpublished) 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 single-cell RNA sequencing at day 4, day 8, and day 40. We then processed the data using Cell Ranger [68] 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 [56], 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.

Transcription factor perturbation scRNA-seq dataset (previously unpublished) The sgRNA library was cloned into an MSCV retroviral backbone containing an expression cassette for human CD2 as a selectable marker. The library targets 39 genes the majority of which are transcription or epigenetic factors. sgRNA sequences were chosen via CHOPCHOP. Negative control sgRNAs were spiked in to make up 15 percent of the final library. Retrovirus production was performed using platinum-E cells. P14 Cas9 transgenic CD8+ T cells were isolated and activated with anti-CD3/CD28 activation beads for 1 day prior to infection with sgRNA library retrovirus using retronectin and protamine sulfate. The cells were then grown in human IL-2 (5ng/mL) for 2 days prior to magnetic selection to enrich for transduced cells based on the human CD2 selection marker. 100,000 cells were then transferred into day 1 LCMV-Armstrong infected mice. At day 8 of infection, mice were sacrificed and P14 T cells were sorted from the spleen. Equal numbers of P14s were pooled from 7 mice and then loaded into 6 wells of a ChromiumX instrument to perform paired 5' scRNA and CRISPR sequencing. The data was then processed using Cell Ranger [68] prior to downstream analyses. We removed cells with less than 200 genes detected and cells with less than 500 counts and more than 20000 counts. Cells with mitochondrial percentage greater than 5% were removed. To remove doublets, the data was log-transformed, scaled, and centered, and doublets were identified with DoubletFinder [38] with parameterSweep_v3 and expected doublet percentage of 7.5%. We then clustered cells using SNN-based Louvain clustering [51] and removed contaminating populations enriched for non-CD8+ T cell markers. Finally, we integrated the replicates with CCA [29] using the top 2000 highly variable genes and removing TCR-related, mitochondrial, proliferation, and immunoglobulin genes. This resulted in 23,206 cells and 1,795 genes across all perturbations.

Peripheral tolerance scRNA-seq dataset (Damo et al.) We obtained scRNA-seq data from [14] and pre-processed it as in the scRNAseq Methods section. There were 21,178 cells and 21,515 genes after pre-processing. This corresponds to 8,167 cells from Ag ON samples; 3,944 cells from Ag ON/CPI samples; and 9,067 cells from Ag OFF (i.e. no AG) samples. Cells were visualized with PHATE and key 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 40, and α was set to 10 to match the analysis in the original work.

10x Human Lymph Node Spatial Transcriptomics Dataset 10x Genomics data was obtained from the 10x website [1] and downloaded via the `scanpy` package [62]. 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.

Immunotherapy response in melanoma patients scRNA-seq dataset (Sade-Feldman et al.) We obtained pre-processed scRNA-seq data with annotated cell types and other relevant metadata (e.g. sample labels, patient response) from [45] 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.

6.9 Computational Details

Diffusion wavelets versus MAGIC For Extended Data Figure 2, the smooth signal was defined based on visualization coordinates to capture the low-dimensional axis of variation. The oscillating signal was defined by the sine of 4 times the smooth signal to define a medium-frequency component. The noise was generated based on random samples from a uniform distribution between $[-1, 1)$. The aggregate signal was defined based on the sum of these three components. The denoised signal was computed with MAGIC and the wavelet dictionary was computed with GSPA and $J = 5$.

Dirac signals comparing GSPA and MAGIC For the dirac experiment (Extended Data Figure 3), we simulated and pre-processed a linear trajectory of cells as described above, and we defined signals as a dirac on each cell, where the signals naturally had a pseudotime label based on the cell it was defined on. We learned unsupervised embeddings for GSPA+QR, GSPA, and MAGIC as described above. Then, we increased the distance between signals by subsampling every other cell in the pseudotime-ordered trajectory, then every 2^2 cells, every 2^4 cells, and so on, and we learned embeddings for the signals increasingly spread apart. We then evaluated the embeddings by repeated K fold splits and ridge regression to predict the pseudotime labels, comparing methods based on Spearman correlation between prediction and true labels across 10 runs. Gene embeddings were visualized with PHATE.

Simulated coexpression experimental details For the coexpression experiment with a linear trajectory (Figure 2) and two and three branches (Extended Data Figure 4), we generated simulated data as described above, 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 the genes 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.

Finally, we visualized the GSPA+QR embedding with PHATE and used spectral clustering to identify gene modules. To map these gene modules back to the cells most enriched for the modules, we leveraged a

gene set enrichment approach from [46] and implemented in [62]. This approach provides a score defined on all cells as the average expression of a set of genes subtracted with the average expression of a reference set of genes. Using the 25% of the number of genes as the number of bins, we calculated a cell enrichment score for each gene module and visualized this enrichment score over pseudotime.

Simulated localization experimental details For the localization experiment with a linear trajectory (Figure 3) and two and three branches (Extended Data Figure 6), we generated simulated data as described above. However, instead of using the genes as signals, we designed signals with "ground truth" localization labels. 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. To mimic this, we defined window sizes as 20%, 40%, 60%, 80%, or 100% of the full trajectory based on the pseudotime (given by the simulation engine). The window can be defined anywhere on the trajectory, so is defined only by the size of the window. Then, we sample 500 cells from this window and let the signal equal 1 on these cells and 0 on all other cells. Generating 50 signals for each of 5 windows, this results in 250 signals. The localization label for each signal is defined by the size of the window. Importantly, such a metric is not equivalent to the number of cells the signal is defined on, as one can imagine two signals with the same number of non-zero values, where one is more spread out on the cellular graph; therefore, all signals are defined on exactly 500 cells.

Then, we define a uniformly-enriched gene as a signal equal to 1 on all cells. L2 normalizing all signals, we embed the uniform gene with other signals and compute the (multiscale) distance of all signals to the uniform distribution. We intuit that the uniformly expressed gene should be closest to signals that were sampled from the full trajectory (window size 100%) and furthest from the signals that were sampled from a small region of the trajectory (window size 20%). That is, there is an anti-correlation between the window size and the distance to uniform. We therefore computed the L2 distance of the signals to the uniform and evaluated methods based on how anti-correlated the distance is with the window size using Spearman correlation.

For GSPA+QR, GSPA, and MAGIC, this 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.

Computing cluster rank for localization versus cluster rank comparison For the comparison in Figure 4d, we performed Leiden clustering on the cells, which identified 9 cell clusters. Using `scanpy`, we ranked genes based on the Wilcoxon rank sum test to identify genes differentially enriched in each cluster. This results in a z-score underlying the computation of a p-value for each gene for each group. For each gene, we stored the maximum score across clusters, and we ranked genes based on this maximum score. This ranking reflects how enriched the gene is, which we compare against the computed localization score.

Computing gene module enrichment per condition Genes were visualized with PHATE and modules were identified via Leiden clustering (Figure 4b). Using the same gene module enrichment approach as in the simulated coexpression experiment, now using the highly variable genes as the reference set, we calculated a cell enrichment score for each gene module. As positive values of this score indicate an enrichment over the reference set, we counted cells with a score > 0 from each sample and normalized this count by the number of cells from each sample. This represents the gene module enrichment of cells per condition (Figure 4e).

Building module-specific gene coexpression networks To build module-specific gene coexpression networks (Figure 4f), 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 [47].

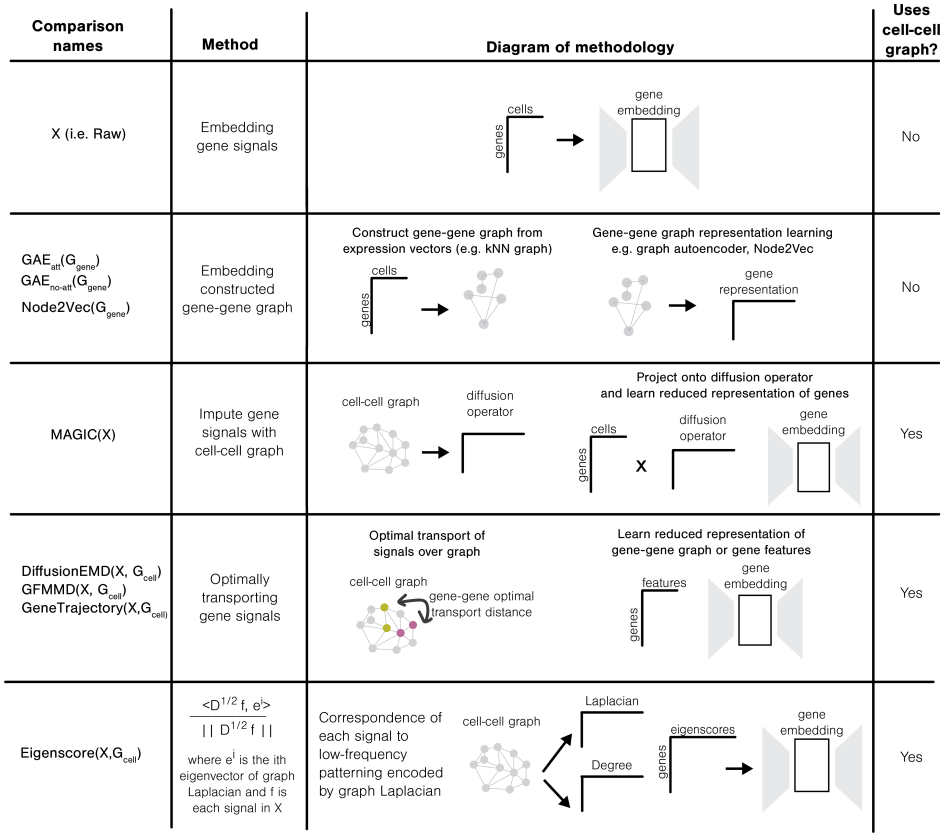
Building perturbation-specific gene coexpression networks First, we identified genes that were in the top 25% localized in both the negative control and the knockout (KO). Then, we built a k-NN graph with $k = 5$ for the negative control genes, and a k-NN graph with $k = 100$ for the KO genes from the 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.

Determining cells enriched with ligand and receptor in cell type-independent cell-cell communication analysis LR pairs were embedded with PHATE, and modules were identified via Leiden clustering (Figure 5e). Then, to map these gene modules back to the cells most enriched for the modules, we leveraged the gene set enrichment approach from [46] to identify cells enriched for the ligands and the receptors for each module (Extended Data Figure 10). We visualize positively enriched cells (score > 0) for the ligands and the receptors in module 19 and calculate the condition-specific score as above.

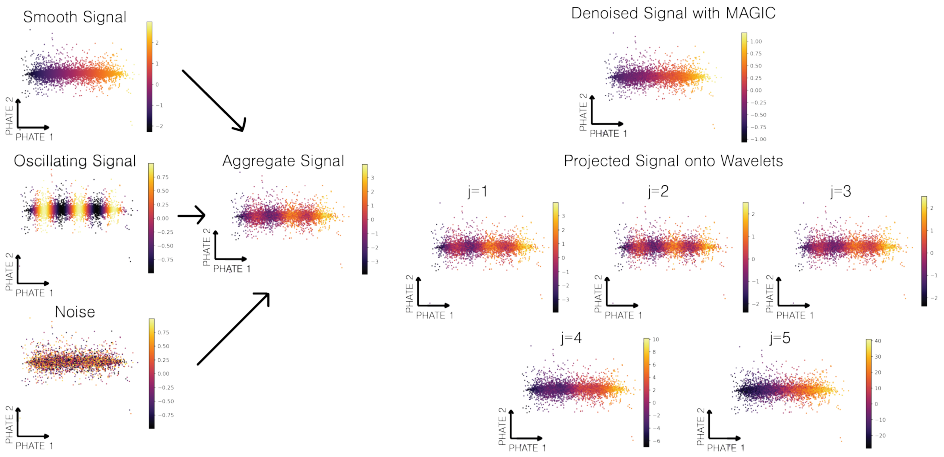
Building module-specific cell-state communication networks from spatial transcriptomic data We visualize gene embeddings with PHATE and identify gene modules via Leiden clustering (Figure 6c). For each gene module, we built a k-NN graph (k=5) and kept only edges that existed in OmniPathDB, adding directionality based on the OmniPathDB annotation. We then mapped the gene signaling network to the cell-type signaling networks. We repeat the following for each gene module graph: for each directed edge ($gene_s$, $gene_t$), for all pairs of cell states ($celltype_a$, $celltype_b$), if $gene_s$ is differentially expressed in $celltype_a$, and $gene_t$ is differentially expressed in $celltype_b$, we add a directed edge from $celltype_a$ to $celltype_b$. We finally use Cytoscape to visualize edges between different cell types (i.e. $celltype_a \neq celltype_b$) in blue, and between the same cell type (i.e. $celltype_a = celltype_b$) in red.

Learning patient embeddings and immunotherapy response We built a combined cell-cell graph across all patients, then split the wavelet dictionary into patient-specific dictionaries and projected the gene signals for each patient onto the patient-specific dictionary. We then performed PCA with 5 components and flattened these gene representations into a single vector of size 1 x $5m$ to represent the patient. For comparison, we performed GSPA using the patient indicator signals on the cell-cell graph, then ran PCA+AE as described above. We also computed the mean expression across all genes for each patient. Finally, we computed the proportion of all clusters (representing immune cell types) and all CD8 clusters (representing CD8 cell types). 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 (Extended Data Table 4). We visualize all patient embeddings with PHATE.

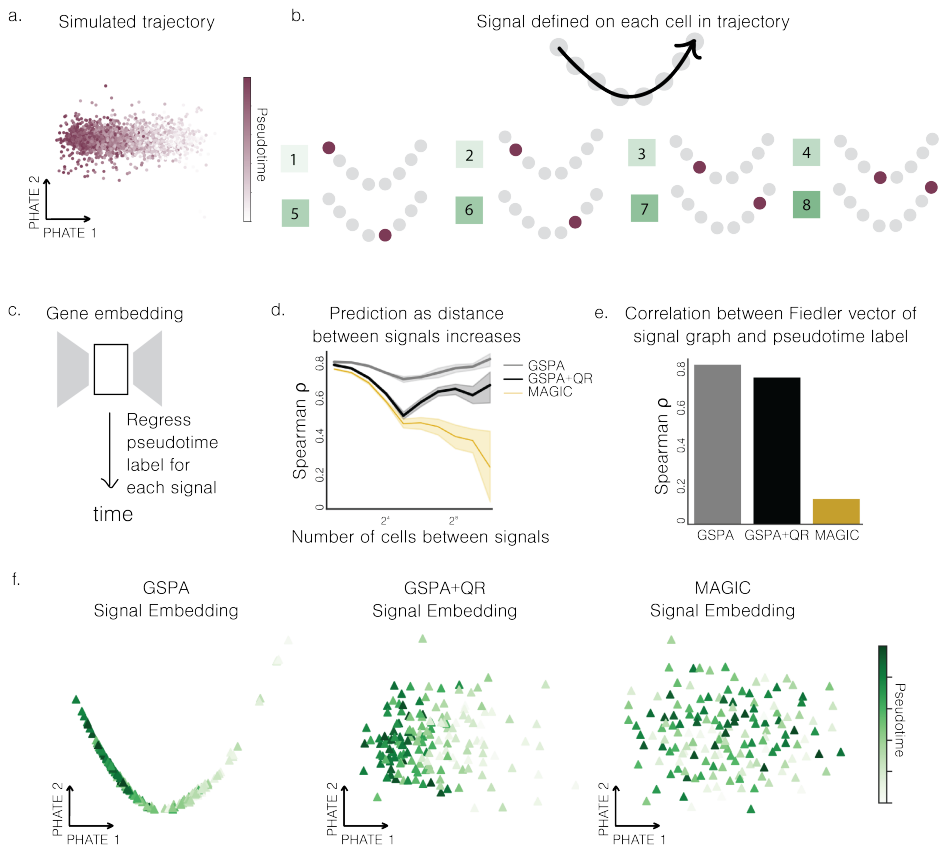
6.10 Extended Data Figures



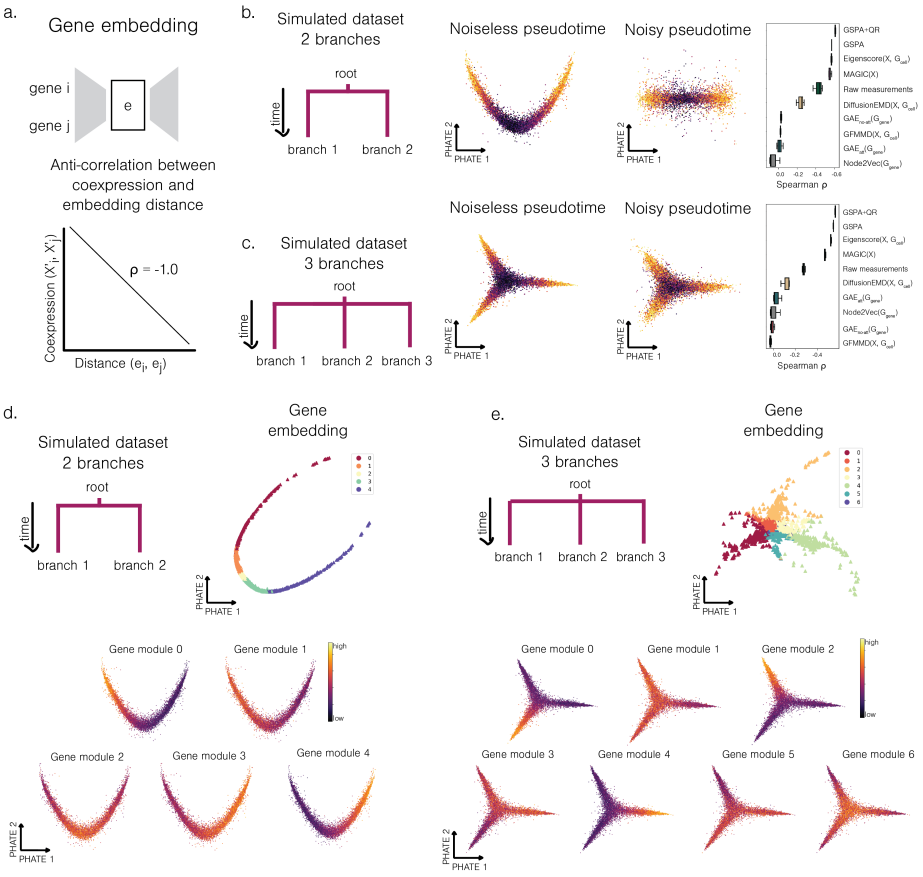
Extended Data Figure 1. Overview of Gene Signal Pattern Analysis Comparisons.



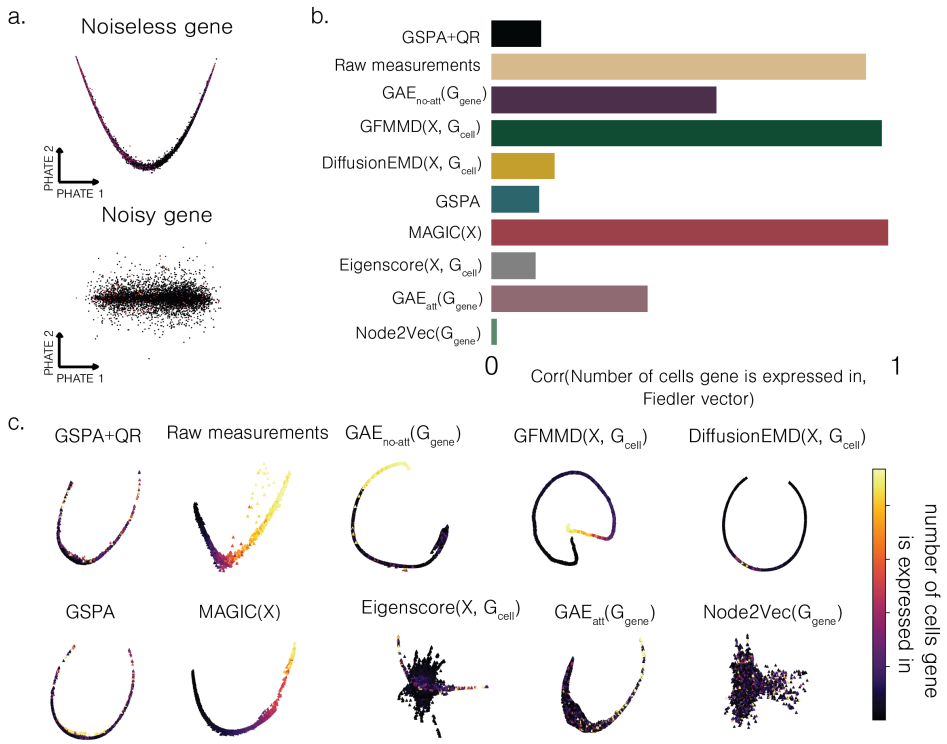
Extended Data Figure 2. Simulated signal with low-frequency, medium-frequency, and high-frequency components. MAGIC denoises the signal and preserves the low-frequency components, whereas projection onto the wavelets reveals both medium and high-frequency components.



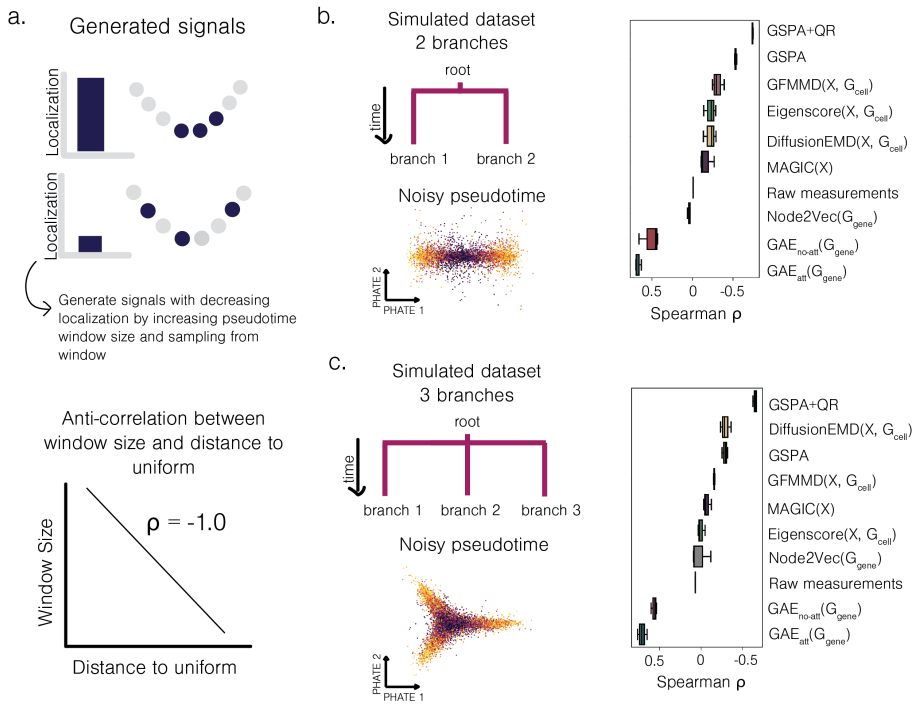
Extended Data Figure 3. GSPA captures local and global geometry of cell-cell graph, enabling comparison of far-apart gene signals. a. Simulated linear trajectory with noise visualized with PHATE. b. Schematic of generation of signals, defined on a single cell for each cell. c. Schematic of experimental setup. d. Spearman correlation evaluating performance on regression task as signals are further separated across 10 runs. e. Spearman correlation between Fiedler vector of signal graphs and pseudotime label. f. PHATE visualizations of signal embeddings colored by pseudotime labels.



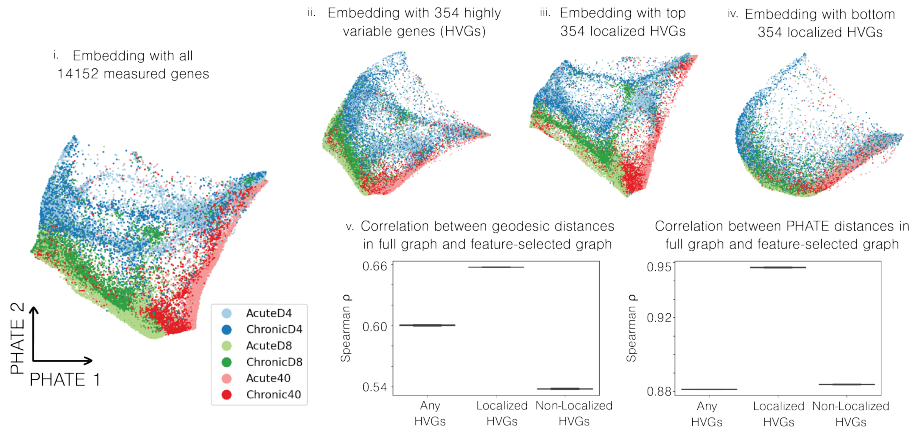
Extended Data Figure 4. GSPA preserves coexpression in two alternative single-cell simulations. a. Experimental setup. b. Simulated dataset with two branches schematic. PHATE embedding of cells from noiseless simulation and noisy simulation, colored by pseudotime. Spearman correlation evaluating performance 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.



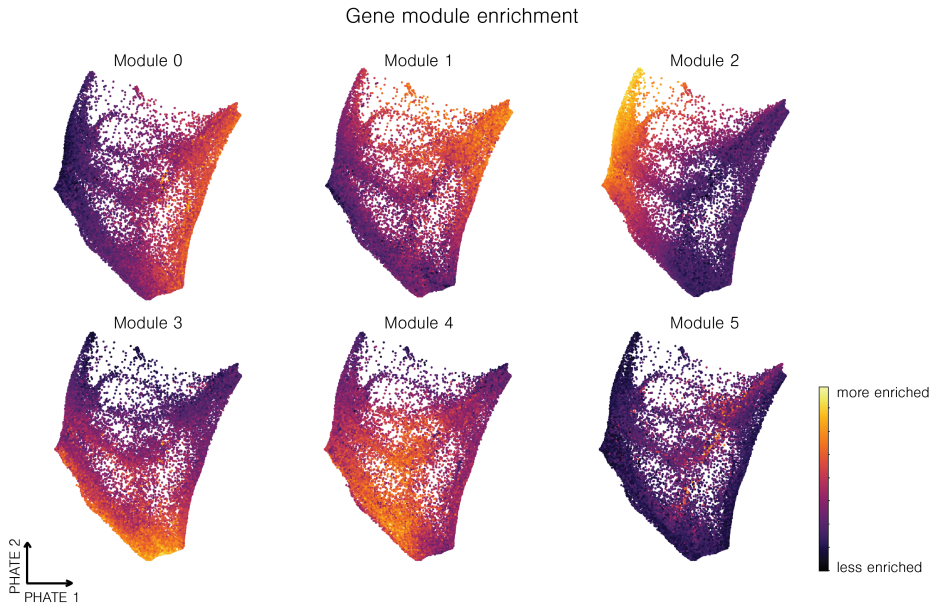
Extended Data Figure 5. Majority of methods primarily capture number of cells expressing gene as the primary axis of variation. a. Noiselss versus noisy linear trajectory. b. Spearman correlation between Fiedler vector of signal graphs and number of cells gene is expressed in. c. PHATE visualizations of gene embeddings colored by number of cells gene is expressed in.



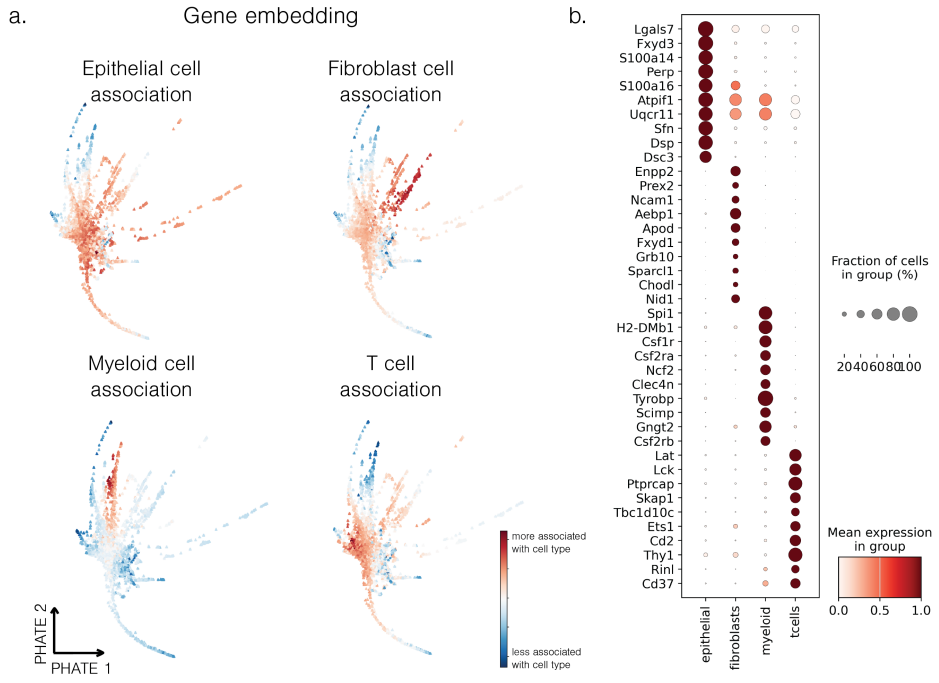
Extended Data Figure 6. GSPA captures localized genes in alternative single-cell simulations. a. Diagram of generated signals based on pseudotime window and anti-correlation between window size and localization. b. Two branch noisy simulated dataset, visualized with PHATE and colored by pseudotime. Spearman 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.



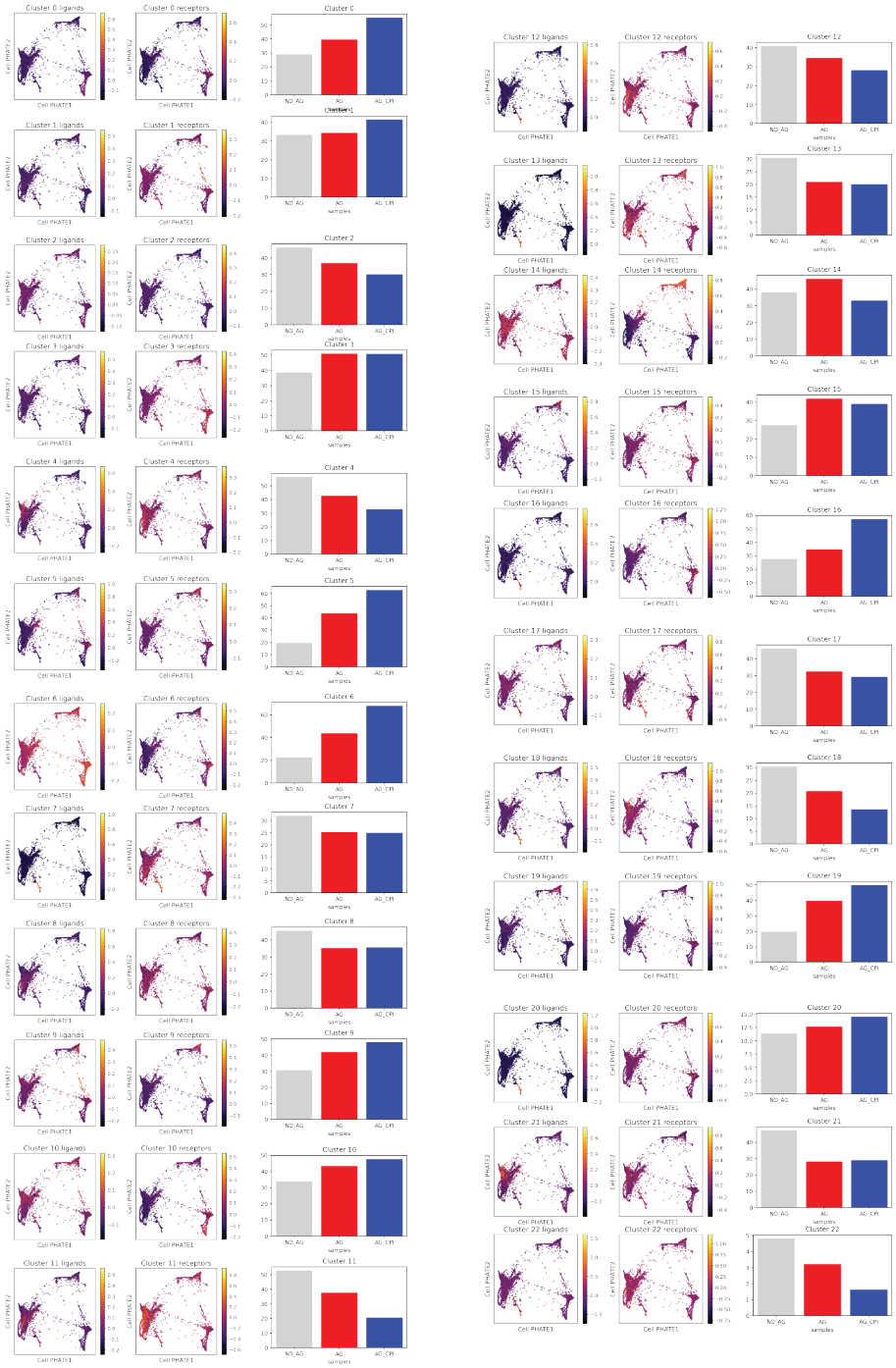
Extended Data Figure 7. GSPA captures biologically-relevant gene patterns from data. Cell embeddings visualized with PHATE, with (i) all genes (ii) a random 25% subset of the top 10% highly variable genes, (iii) top 25% localized genes of highly variable genes, and (iv) bottom 25% localized genes of highly variable genes. (v). Left: 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. Right: Spearman correlation between random subset of 100,000 pairwise PHATE distances (chosen in two runs) in full cell-cell graph and feature-selected cell-cell graphs.



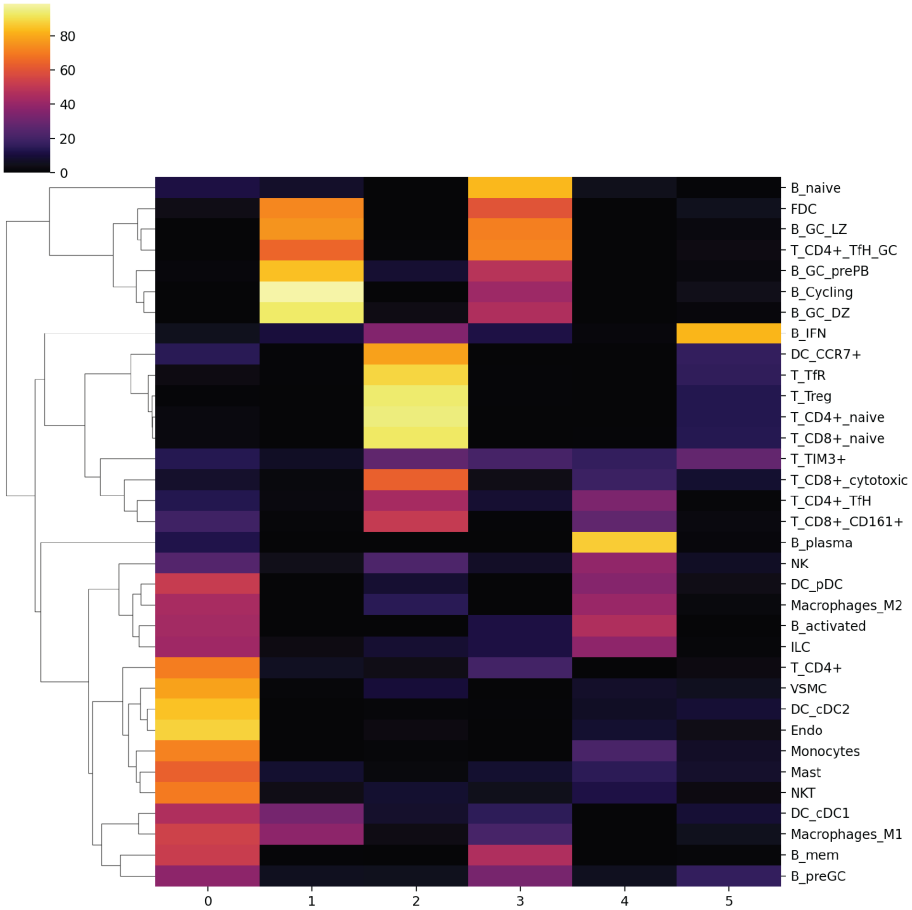
Extended Data Figure 8. Gene module enrichment score for each gene module.



Extended Data Figure 9. Cell type association scores for four cell types in peripheral tolerance model. a. Gene embedding colored by cell type association ranking. b. Dot plot with top 10 genes associated with each cell type.



Extended Data Figure 10. Ligand enrichment score, receptor enrichment score, and enrichment in each condition for all LR pair modules.



Extended Data Figure 11. Differentially expressed genes per cell state in each gene module.