Navigating the manifold of single-cell gene coexpression to discover interpretable gene programs

Sinan Ozbay[*] Aditya Parekh[*] Rohit Singh[*†‡]

Abstract

During development and differentiation, transcriptional regulation in the cell often occurs at the level of gene programs (i.e., sets of co-varying genes), rather than isolated genes. It is therefore crucial to identify differential program expression over time, or across case-vs-control samples. However, this has remained difficult: gene programs are inferred by analyzing gene coexpression, and mathematical operations on the latter are nontrivial. Gene coexpression is quantified as a symmetric positive-definite matrix, on which even basic quantities such as arithmetic differences are neither mathematically sound nor biologically interpretable. Here we exploit the structure of the Riemannian manifold of gene coexpression matrices to propose a novel abstraction of gene coexpression that is mathematically well-founded while being computationally tractable and statistically rigorous. Importantly, it also captures biological similarity better than standard coexpression. This conceptual advance enables us to introduce Sceodesic, an algorithm that invokes the log-Euclidean metric from differential geometry to quantify coexpression patterns specific to each cell state, and organizes them into a study-wide panel of interpretable gene programs. Applied to nine single-cell RNA-seq datasets, Sceodesic outperforms existing methods in early detection of cell fate commitment by leveraging differential expression of gene programs, and is also effective in discovering disease-linked programs in multi-sample studies. By respecting the manifold of gene coexpression matrices, Sceodesic resolves a longstanding challenge in relating biological variability to statistical analyses of single-cell RNA-seq data and enables the discovery of gene programs driving differentiation and disease.

Software availability: `https://singhlab.net/Sceodesic`

1 Introduction

The city of Istanbul is southeast of Boston, yet a flight from the latter to the former begins by heading northeast. This counter-intuitive path arises because the (Euclidean) straight-line route as seen on a Mercator map is sub-optimal; the true shortest path follows a (Riemannian) great circle. In this work, we show how a similar principle is key to effectively exploiting coexpression.

In cells, genes often operate in coordinated *gene programs*, evidenced by sets of co-varying genes in single-cell RNA-seq (scRNA-seq) data. Thus, two superficially similar gene expression profiles may mask significant differences in underlying gene program expression. Yet, methods that focus on analyzing changes in gene program activity are relatively few when compared to those studying individual genes. One major obstacle has been the lack of a mathematically rigorous yet biologically faithful and computationally efficient abstraction for gene

[*]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705
[†]Department of Cell Biology, Duke University, Durham, NC 27705
[‡]Corresponding author: rohit.singh@duke.edu

coexpression. For an individual gene, the arithmetic difference in expression levels is a biologically meaningful measure, while also being amenable to a range of statistical and machine learning techniques reliant on Euclidean distance metrics [1–3]. In contrast, gene coexpression is represented by matrices that fall within the specialized category of positive semi-definite (PSD) matrices. Naive arithmetic difference of matrices can compromise PSD integrity [4], raising computational challenges and muddying biological interpretability. Our method addresses these gaps, paving the way for more robust and nuanced analyses of gene programs.

While our method proves valuable in various scRNA-seq contexts, our motivating biological application is fate prediction. During development or disease, multipotent progenitor cells can differentiate into diverse terminal types [5, 6]. Pinpointing when such a cell has committed to a particular fate is crucial, both fundamentally and translationally. In cancer, this knowledge can predict a tumor's potential malignancy from early biopsies. Computationally, the task here can be formulated as identifying early markers of a given lineage or forecasting the most likely terminal cell type of a given transcriptional state. A seminal study by Weinreb et al. used lineage tracing and scRNA-seq to assess cell fate predictability, but found that using expression of individual genes was of limited effectiveness [7]. We hypothesize that using gene programs, rather than isolated genes, would make fate commitments more discernible (Fig. 1A). Due to gene expression's inherent stochasticity and scRNA-seq noise, isolated gene changes can be ambiguous. Yet, if changes occur in a coordinated gene program, the aggregated signal becomes easier to distinguish from noise.

Unfortunately, existing gene program discovery methods fall short in fate analysis. First, supervised methods that search among known gene sets will struggle to uncover de novo gene programs [8]. Second, the method must be able to discover gene programs active solely in select cell types or transient states. Traditional approaches like principal component analysis and non-negative matrix factorization, along with deep-learning methods like scVI [9], focus on global program discovery—- identifying gene programs based on global covariance. Such approaches overlook programs salient to only subsets of cell types (Fig. 1B), and risk underperformance in longitudinal or case-control datasets, where technical artifacts confound biological variability. A variety of batch-effect corrections have been developed to address the latter [10–12], but this raises concerns that technical adjustments to the data obscure sample-specific biological signals.

We propose that basing gene program discovery on local expression variability about a cell state could unveil even rare programs. Fortunately, current scRNA-seq assays now routinely profile 10^5–10^6 cells, providing ample coverage for assessing such cell state-specific stochasticity. Yet this introduces a third essential criterion: a coherent framework relating biological shifts across cell states to changes in gene program expression. For other biological tasks, methods have been proposed to uncover cell-state-specific gene programs and relate them across the entire dataset, However, whether pinpointing gene regulation in rare cell types [13] or tracing trajectories across studies [14], these methods struggle to ensure that the covariance estimates remain PSD, and are unable to account for the non-Euclidean nature of distances between PSD matrices. Consequently, they often rely on ad hoc mathematical techniques that challenge both evaluation and broader applicability.

We introduce Sceodesic, an algorithm designed to extract cell-state-specific and interpretable gene expression programs from scRNA-seq data. Sceodesic melds a novel blend of differential geometry, spectral analysis, and sparse estimation to pinpoint gene expression programs that are not only specific to cell states but also robust against variations in case-control, longitudinal, or batch conditions. We begin by partitioning the input set of cells into hundreds or thousands of "cell cohorts", each sufficiently populous to calculate an expression covariance matrix. Importantly, the sheer number of these cell cohorts ensures that even transient cell states are likely to be captured.

Sceodesic exploits the Riemannian geometry of coexpression to address the challenge of integrating cell state-specific covariance matrices into a single study-wide interpretable framework. Our key conceptual breakthrough is underpinned by the observation that the logarithmic map applied to the Riemannian manifold of positive semi-definite

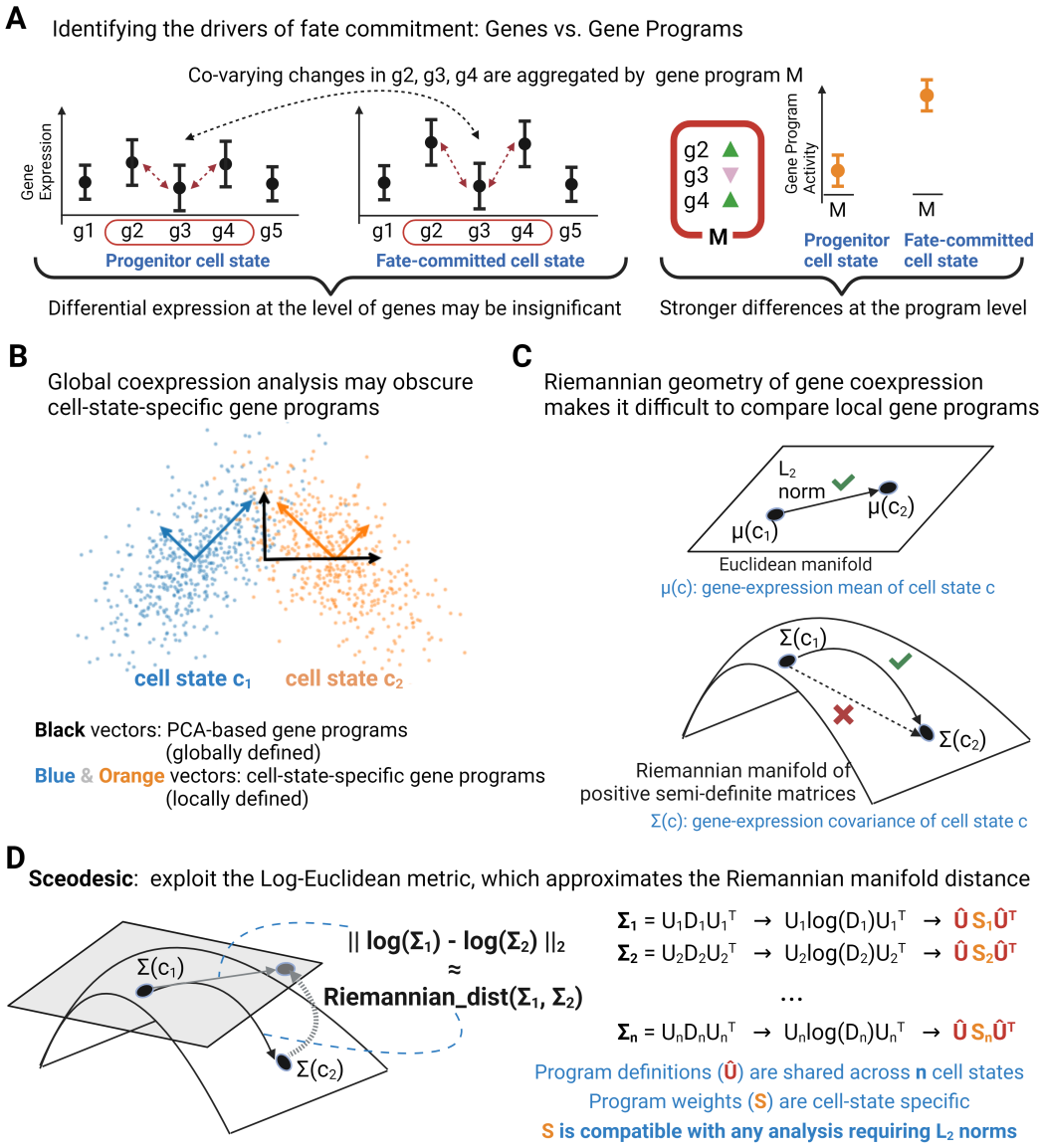

Figure 1: Leveraging Riemannian geometry for discovering scRNA-seq gene programs. **A**) Transcriptional drivers of cell differentiation and fate commitment may be more visible at the scale of gene programs (sets of co-varying genes) rather than individual genes since the former aggregate information over many genes. **B**) Global techniques like PCA are unable to access state-specific variations. With large scRNA-seq datasets now available, subsets of the data (cell cohorts) can be analyzed to uncover cell state-specific gene programs. **C**) Gene covariance obeys Riemannian, not Euclidean, geometry making it a challenge to compare and score gene programs with naive Euclidean distances. **D**) Sceodesic addresses this by approximating the Riemannian manifold distance with the Log-Euclidean metric, defined on the tangent bundle of the manifold. Using sparse eigenanalysis, we aggregate locally-defined programs and construct programs whose definitions are consistent across cell states and state-specific weights are amenable to Euclidean distances.

matrices affords a way to preserve the semantics of gene covariance while employing Euclidean distance metrics (Fig. 1C,D) [4]. Paired with population value decomposition [15], a framework for shared eigenbasis inference, as well as innovations in well-conditioned covariance estimation, the end product is principled, yet easy-to-use (no GPU required) and fully compatible with existing scRNA-seq analyses. Applying Sceodesic to fate prediction, we find that its programs significantly outperform globally-defined expression programs and individual genes in predicting

the likely trajectory of partially differentiated cells. In re-analyzing fate-mapped trajectories, shifts in Sceodesic programs become evident considerably sooner than changes related to fate commitment in individual genes.

Sceodesic is engineered to seamlessly augment a wide range of standard scRNA-seq analyses. It does so by appending to each cell's representation a feature vector that encapsulates cell-state-specific gene programs, and notably, relies solely on covariance rather than mean expression levels. We demonstrate some of the enhancements that we anticipate standard analyses will gain when integrated with Sceodesic. While we focus on a specific scRNA-seq application domain, the conceptual and methodological innovations we present are also relevant in other problems where gene coexpression arises as an entity of interest. By introducing deep mathematical principles into a framework with practical utility, Sceodesic is poised to drive fundamental and translational insights.

2 Background

2.1 Distance metrics for gene coexpression via positive semi-definite matrices

Consider cell states, C_1, C_2 and C_3, with C_1 differentiating into either C_2 or C_3. Distances between their transcriptional profiles are informative to predict whether C_2 or C_3 is more likely. If only a few independently-varying genes differentiate the states, Euclidean distances between gene expression values are suitable; however this would not be the case when gene programs are active and many genes are co-varying. For the latter case, suppose each cell state is represented by a "cell cohort," a group of similar cells exemplifying the stochastic variability *within* the state. We propose to represent each cell state by the matrix of gene expression covariance within the cohort: Σ_1, Σ_2, or $\Sigma_3 \in \mathbb{R}^{n\times n}$. Importantly, the n^2 entries in a covariance matrix for n genes are constrained by each other. Specifically, each matrix Σ must be symmetric and positive semi-definite (PSD), i.e., its singular values must be non-negative. The Euclidean-distance intuition can be extended to the cell cohort case by using the Frobenius norm of the difference between two matrices, given as $\|\Sigma^{(1)}-\Sigma^{(2)}\|_F^2=\sum_{i=1}^n\sum_{j=1}^n(\Sigma_{ij}^{(1)}-\Sigma_{ij}^{(2)})^2$. While this is computationally convenient, two key concerns arise. First, it is unclear if the Frobenius norm is the most biologically meaningful metric. More importantly, this operation does not preserve the PSD nature of the matrices, complicating analyses such as identifying shared gene programs across different cohorts.

2.2 Relevant concepts from Riemannian geometry

Riemannian geometry extends Euclidean notions like distances and angles to smoothly varying, non-Euclidean spaces. This is crucial as the subspace of PSD matrices in $\mathbb{R}^{n\times n}$ is a Riemannian manifold (often denoted $\mathcal{S}_+^n$). Briefly, an m-dimensional Riemannian manifold $\mathcal{M}$ in $\mathbb{R}^m$ is a space that locally approximates Euclidean space near each point, an intuition made precise through the introduction of a tangent space $T_p\mathcal{M}$ at each point p on the manifold. The tangent space $T_p\mathcal{M}$ is a vector space formed by the velocity vectors (at p) of curves on $\mathcal{M}$ that pass through p. It provides an Euclidean approximation to the manifold at that point, encapsulating local geometric properties. Many operations on $\mathcal{M}$ are defined in relation to $T_p\mathcal{M}$ or, more broadly, the tangent bundle $T\mathcal{M}$, which comprises such tangent spaces at every point on the manifold.

Two mappings facilitate the transfer of geometric intuition between the manifold and its tangent spaces: the logarithmic map transfers a neighborhood of p on $\mathcal{M}$ to a neighborhood in $T_p\mathcal{M}$, while the exponential map does the reverse (see also Sec. A.1). Through these maps, Euclidean operations on the tangent spaces can be transformed into corresponding non-Euclidean operations on $\mathcal{M}$. In particular, straight lines in the tangent space map to *geodesics* on the manifold. Geodesics represent the locally shortest paths between two points on $\mathcal{M}$. For example, on a sphere like the Earth, such paths trace along the great circle connecting two locations.

We argue that the geodesic distance provides the most appropriate measure for gauging the similarity between PSD matrices, and **the crux of our conceptual advance is exploiting the structure of the logarithmic map** for $\mathcal{S}_+^n$.

For the point $p \in \mathcal{S}_+^n$ corresponding to the PSD matrix Σ_p, the logarithmic map at p, $\log_p(\Sigma_x) : \mathcal{S}_+^n \to T_p\mathcal{S}_+^n$, is simply the *matrix logarithm*: $\log(\Sigma_x)$. We note that matrix logarithms are very amenable to spectral analyses: given the eigen-decomposition $\Sigma = UDU^T$ for $\Sigma \succ 0$, the matrix logarithm has the same eigenvectors but its eigenvalues are logarithms of the original's: $\log(\Sigma) = U\log(D)U^T$. This close connection is a linchpin of our approach.

3 Methods

The Sceodesic algorithm. The goal of the Sceodesic algorithm is to estimate biologically meaningful gene programs and recapitulate gene expression data as levels of gene program expression. The input to Sceodesic is a cell-by-gene count matrix $X \in \mathbb{R}^{N \times G}$. Sceodesic outputs a set of h programs $\hat{U} \in \mathbb{R}^{h \times h}$ (i.e., each program is a sparse h-dimensional vector) and a matrix of cell-by-program expression, $S_{sceo} \in \mathbb{R}^{N \times h}$. Here, h is a user-specified parameter designating the number of variable genes to be considered. The computation consists of four main steps:

1. Estimate cell cohort membership. We delineate groups of cells by k-means clustering on log- and count-normalized gene expression. Cells with similar levels of similar transcriptional profiles will be grouped together in the same cohort.

2. For each cell cohort, compute a covariance matrix. We newly introduce *condition-volume normalization*, which overcomes the problem of a singular sample covariance matrix by regularizing them into a well-conditioned, well-scaled positive-definite matrix close to it. The eigenvectors of this matrix constitute cohort-specific gene programs.

3. Using population value decomposition, we share gene program definitions across cell cohorts to learn a shared eigenbasis that is applicable globally. We use sparse estimation to learn a set of h interpretable features that best recapitulate local, cohort-specific features across the entire dataset. These features are Sceodesic's estimated gene programs $\hat{U}$.

4. Reconstruct local cohort-specific coexpression matrices in terms of the shared program definitions while respecting the manifold of PSD matrices. Crucially, the reconstruction error minimized is the log-Euclidean distance between the condition-volume normalized covariance matrix of each cell cohort and the reconstruction. This ensures that biological information is preserved in the reconstruction step.

Cell cohort construction. Sceodesic constructs cell cohorts by partitioning a scRNA-seq dataset into n cell cohorts c_i, $i = 1,...n$ under the premise that cells alike in terms of their transcriptional state will be grouped under the same cell cohort. In principle any single-cell clustering method can be used, but we chose k-means because we found it to deliver reasonably balanced clusters, which is important because well-estimated coexpression matrices of cell cohorts are the 'base estimator' of Sceodesic. Well-balanced cell cohorts avoid extremely small clusters which would need to be heavily regularized to reside on the positive-definite manifold, distorting the spectrum of the raw covariance matrix.

Sceodesic computes condition-number normalized covariance matrices. Next, Sceodesic generates positive-definite covariance matrices for each cell cohort. First, the data matrix X is restricted to columns representing only h variable genes (determined by a gene having high dispersion across a large proportion of clusters, Sec. A.2), denoted $X_{hvg} \in \mathbb{R}^{N \times h}$. This ensures the number of samples needed per cohort to avoid a degenerate covariance matrix is not prohibitive. For each cell cohort c_i, the sample gene expression covariance matrix $\Sigma_i \in \mathbb{R}^{h \times h}$ is computed. However, for some cohorts, this matrix may have zero eigenvalues or be ill-conditioned, rendering matrix logarithms unstable. We address this by regularizing Σ_i, first computing its spectral decomposition $\Sigma_i = U_i D_i U_i^T$, then, given a condition number M as the target, we "round up" the smallest eigenvalues to enforce the target. Let the largest eigenvalue of D_i be σ_{max}. We define the diagonal matrix $\tilde{D}_i$ as:

$$[\tilde{D}_i]_{jj} = \max\left([D_i]_{jj}, \frac{\sigma_{max}}{M}\right) \text{ for } j \in \{1,...,h\}$$

This operation is well-justified mathematically. Tanaka and Nakata showed that $U_i \tilde{D}_i U_i^T$ is the semi-definite matrix with condition number at least M that is closest (in a Frobenius-norm sense) to Σ_i [16]. Lastly, we re-scale this matrix by scaling diagonal elements of $\tilde{D}_i$ by the smallest eigenvalue $\tilde{\sigma}_{min}$, so that $1 \leq \tilde{\sigma}_j \leq M \;\; j = 1...h$. This offers two advantages: a) when mapping to the tangent space by taking logarithms, the smallest value(s) become zero, enhancing interpretability and sparsity, b) we effectively bound the determinant to be $\leq M^h$. In linear algebra, the determinant has a 'volume' interpretation, thus bounding it makes the magnitude of overall covariance comparable across cohorts, analogous to count normalization for gene expression.

Sceodesic learns study-wide gene programs from local gene programs via population-value decomposition The eigenvectors $U_i \in \mathbb{R}^{h \times h}$ constitute local gene programs for the cell cohort c_i. Sceodesic next identifies frequently-occurring and rare programs across cohorts, producing a study-wide set of programs that could be a shared dictionary globally. To do so, we adapt population value decomposition (PVD), a method first introduced in image processing [15]. PVD estimates common features between images by estimating an optimal set of vectors that serve as a common basis for reconstructing *features* specific to individual images, rather than the images *themselves*. In our context, the key insight of PVD is that the set of eigenvectors across all cohorts can itself be viewed as observations in $\mathbb{R}^h$. Performing PCA on this set will reveal frequent and rare programs, as high- and low-ranked PCs, respectively. We select only the eigenvectors of each cell cohort sufficient to explain 90% of the variability in the cohort, in order to limit noisy features. Denoting this set U_i^v, we concatenate them across cohorts,

$$U = [U_1^v|,...,|U_n^v]$$

and perform a sparse principal component analysis on the rows of U^T as if they were themselves samples (although each row is really a gene program). This yields a series of sparse vectors $P_j \in \mathbb{R}^h$, for $j = 1,...,h$, that best reconstruct the features. The programs' sparse estimation increases their interpretability, especially in context of gene set enrichment analysis. We denote $\hat{U}$ as the matrix whose column vectors are P_j, underscoring that these programs are analagous to eigenvectors.

Sceodesic reconstructs gene coexpression while respecting the PSD manifold. Finally, having estimated study-wide gene programs $\hat{U}$, we reconstruct cell cohort-specific covariance matrices using these programs. This amounts to finding the set of weights on the estimated programs that minimize the distance to a cohort's covariance matrix. However, per the previous discussion, this distance computed between two matrices must respect the PSD manifold's curvature to remain mathematically meaningful and preserve biological information. Therefore, we crucially use the log-Euclidean metric as the basis for our objective function in the reconstruction task. We note that this metric is a lower bound on the manifold geodesic distance, i.e., $\delta_{\mathcal{M}}^{Riemannian}(\Sigma_1,\Sigma_2) \geq ||\log(\Sigma_1) - \log(\Sigma_2)||_F$ [4]. The reconstruction problem for cohort c_i is:

$$\min_{S_i} ||\log(\tilde{\Sigma}_i) - \log(\hat{U} S_i \hat{U}^T)||_F \text{ where } \tilde{\Sigma}_i = U_i \tilde{D}_i U_i^T,$$

subject to the constraint that S_i is diagonal. Essentially, we are solving for closeness in the tangent space. Since $\log(\hat{U} S_i \hat{U}^T) = \hat{U} \log(S_i) \hat{U}^T$, we work with $\log(S_i)$, denoting it S_i^{log}. We describe the solution in Sec. A.2, where we derive the final form for the diagonal entries of S_i^{log}:

$$[\hat{S}_i^{log}]_{jj} = [\hat{U}^T \log(\tilde{\Sigma}_i) \hat{U}]_{jj} \text{ for } j = 1...h$$

The feature weights of $\hat{S}_i^{log}$ recapitulate gene program activity of cohort c_i on the tangent bundle of the $\mathcal{S}_+^n$ manifold, allowing for the weights to be consistently manipulated using Euclidean metrics. At the same time, though the weights correspond to shared study-wide dictionary of gene programs, they are still specific to c_i's cell state.

The final product of Sceodesic is thus twofold: $\hat{U}$, a set of h distinct gene programs each in $\mathbb{R}^h$, and second $S_{sceo} \in \mathbb{R}^{N \times h}$, a matrix of cell-by-program expression. All cells in the cohort c_i have the feature vector $\hat{S}_i^{log} \in \mathbb{R}^h$. These per-cohort features can be concatenated with per-cell features from PCA or NMF to provide a representation capturing both global (PCA or NMF) and local (Sceodesic) variability.

Guidance on hyperparameter choices. The two key hyperparameters in Sceodesic are the number of cell cohorts n, and the number of variable genes h for which coexpression will be estimated. These parameters directly determine whether the base estimators, the sample covariance matrices of the cohorts, tend to be singular. Increasing n provides greater cell-state granularity but decreases the average number of cells per cohort and thus increases the chances that some sample covariance matrices are singular. Similarly, a higher h increases the hurdle to a non-singular matrix. We recommend setting h to about 300, and n to be N/h, so that the typical cohort is just big enough to likely produce a non-singular covariance matrix. For a dataset of 30,000 cells, this implies $n = 100$, enough to cover transient cell states in most studies. These settings are also the default in the code. There are also minor hyperparameters: the target condition number M, the sparsity parameter λ in the sparse PCA step of the population value decomposition, and the amount of explained variance v_{pvd} that determines how many cohort-specific eigenvectors should be included in the PVD. These can be left to their default values (see guidance and robustness analysis in Sec. A.2).

Runtimes & memory requirement: On a dataset of 100,000 cells, Sceodesic takes approximately 20 minutes on an Intel Xeon 3GHz CPU, with runtimes increasing roughly linearly with number of cells. If interpretability of the gene programs is less crucial, these runtimes can be improved up to 3.5x by replacing sparse PCA with regular PCA during PVD. Memory requirements scaled linearly with the size of the dataset, with only 25 GB required to construct program embeddings on a dataset of approximately 1 million cells. See Sec. A.2 for details.

4 Results

4.1 Validating the biological accuracy of the log-Euclidean metric

The central hypothesis underlying Sceodesic is that mathematically consistent distances between gene coexpression of two groups of cells are a reliable proxy for biological similarity. To test this, we investigated three research questions:

1. Does the second moment of gene expression, i.e., covariance, harbor biological information that an appropriate metric can extract?
2. Is the information captured in it distinct from that in the first moment, i.e., mean expression?
3. Among metrics on the second moment, does the Riemannian-based metric more accurately encapsulate biological similarity than the Euclidean metric?

An affirmative answer to the third question would validate our Sceodesic formulation. Moreover, we believe this is the first rigorous comparison of the information in the first and second moments of gene expression, even though earlier work has used Euclidean distances between coexpression matrices without such scrutiny. To answer these questions, we analyzed four large scRNA-seq datasets from the Chan-Zuckerberg CELLxGENE portal [17], and assessed the ability of gene expression-based distance metrics to recapitulate expert cell-type annotations. For each dataset, we applied k-means clustering to compute cell cohorts and marked each cohort with its dominant cell-type label; we excluded cohorts lacking a dominant cell type. We then computed distances between all pairs of cohorts and evaluated each distance metric's ability to predict label similarity. Intuitively, a good metric will

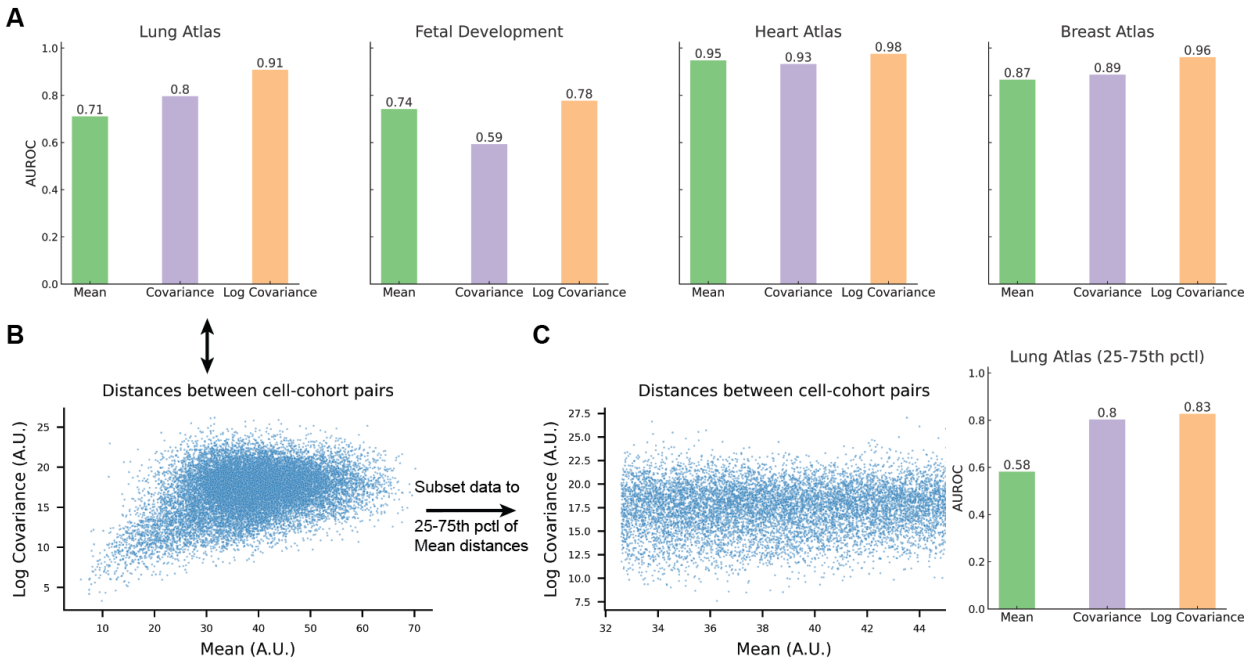

Figure 2: Comparison of distance metrics between cell cohorts: **A)** Comparison of Euclidean distances between cohort means (d_μ), Euclidean distance between covariance matrices (d_Σ), and log-Euclidean distances between covariance matrices ($d_{\log(\Sigma)}$) as 1-D classifiers of cohort-cell type similarity as measured by AUROC. **B)** Points signify cohort pairs with position (d_μ, $d_{\log(\Sigma)}$), showing a stronger relationship between the mean and log-covariance at small distances. **C)** Identical analysis restricted to middle quartiles on the lung dataset show a weaker relationship of d_μ and $d_{\log(\Sigma)}$ on medium distances, with classifier outperformance persisting. (A.U. = Arbitrary Units)

assign a small distance between cell cohorts with the same labels. We used a 1-D discriminant as the classifier (i.e., 'true if distance $<$ threshold, false otherwise'), computing the area under the receiver operating characteristic (AUROC) for each metric. Three distance metrics were compared: a) d_μ, the Euclidean distance between mean gene expression of each cell cohort, b) d_Σ, the Euclidean metric on the second moment, i.e., the Frobenius norm of the difference between the gene expression covariance matrices, and c) $d_{\log(\Sigma)}$, the Riemannian metric on second moment: the Frobenius norm of the difference between the logarithms of the gene expression covariance matrices. We emphasize that the two second-moment metrics were computed identically via condition-number control (see Sec. 3) and that the program-discovery power of Sceodesic was not applied here (see Sec. A.3 for details)

We found that the Riemannian metric $d_{\log(\Sigma)}$ consistently outperformed the first moment d_μ or the Euclidean second-moment metric d_Σ (Fig. 2A). This held true on datasets from diverse organ systems and origins: Human Lung Cell Atlas (2.3 million cells [18]), the 1 million cell survey of embryonic development from Cao et al. [19], the Human Breast Cell Atlas (800,000 cells [20]), and the DCM/ACM Heart Cell Atlas (881,000 cells [21]).

The log-Euclidean metric outperformed the Euclidean metric when applied to the covariance matrices by 0.04 to 0.20 AUROC across all cases. Remarkably, the Riemannian metric also outperformed the first moment metric d_μ, sometimes by substantial amounts (Fig. 2A). These findings demonstrated that a mathematically consistent treatment of gene coexpression contains biological information, affirming our first research question. However, as the second question above asks, it could be that the second moment simply captures the same biological information as the first. In particular, gene expression is often modeled as having a Negative Binomial or Poisson distribution, in which the means and variance are closely related. Thus, the second moment might simply be picking up on the first. We reasoned that this was unlikely since $d_{\log(\Sigma)}$ outperformed d_μ most of the time, but we also tested this directly.

A pair of experiments confirmed that the Riemannian metric contains independent, useful information. First, scatterplots comparing the two metrics, d_μ and $d_{\log(\Sigma)}$, on the same pairs of cohorts showed that they were intrinsically different (Fig. 2B). For cohort pairs in the lowest quartile of d_μ distances, the two metrics appeared to be in meaningful agreement (Spearman correlation 0.46). However, this agreement broke down for medium and larger distances. When we restrict the analysis to pairs of cohorts for which d_μ fell within the 25-75th percentiles, the two metrics showed little agreement (Fig. 2C, Spearman correlation 0.05), suggesting that they were capturing different information in this regime. Moreover, we compared the biological fidelity of the metrics in this domain and found that the log-Euclidean metric continued to fare better at capturing biological similarity between cohorts than does the distance between means, as measured by AUROC. Thus, we also affirmatively answered the second question— whether the information captured in the log-Euclidean metric is distinct from that in the Euclidean metric.

These observations illustrate the value of considering the Riemannian manifold of gene coexpression matrices in delineating biological similarity more accurately. Our finding that the log-Euclidean metric could accurately characterize biological state, or differences in state, establishes a solid mathematical foundation for Sceodesic.

4.2 Sceodesic reveals novel insights into fate commitment

Sceodesic offers a fresh perspective for the early detection of *fate commitment*. We distinguish this task from fate mapping, which is performed by methods like CellRank [22] and Palantir [23] that annotate sets of cells as veering towards one terminal lineage over another as differentiation unfolds. These tools do so by using RNA velocity (CellRank) and/or pseudotime (Palantir) to first establish the dynamics of differentiation and delineate cell-state transitions. They then identify terminal and progenitor states, and allocate to intermediary states a probability distribution over the terminal states. Fate prediction is the next logical step: revealing the early, definitive transcriptional "drivers" of commitment to a specific lineage by either correlating transcriptional profiles with fate probabilities or looking at differential expression of genes or programs between end states. This knowledge is critical to understanding and modulating the fate of a differentiating or tumor cell. While CellRank does offer the ability to estimate individual driver genes whose expression correlates with a lineage, we hypothesized that such drivers may be more discernible at the gene program level than at individual genes, given the potential for technical noise and stochasticity in gene expression to obscure the insights. Furthermore, the range of biological phenomena that can be expressed as gene programs, as compared to individual genes, should be greater. We reasoned that Sceodesic is well-suited for this task as it not only enables the discovery of programs relevant to specific cell states, but also provides a mathematically-consistent framework for a structured analysis of differential expression at the gene program level.

We applied Sceodesic on two datasets from studies of cell differentiation. We used CellRank to estimate differentiation dynamics and cell-specific fate probabilities. To identify individual driver genes for a specific lineage, CellRank uses a generalized additive model (GAM) that relates gene expression to lineage-specific fate probabilities [24]; the default GAM configuration uses cubic B-splines and the Gaussian link function. Exploiting Sceodesic's mathematically-consistent formulation, we used the same statistical formulation of a GAM to identify driver programs. Using CellRank-estimated fate probabilities and the GAM, we estimated the driver gene, Sceodesic program, or principal component (PC) for each lineage. We sought to assess if the Sceodesic programs kicked into action earlier and more strongly than the driver genes or PCs. Our comparison also evaluated if the global programs identified by PCA could match the sensitivity of Sceodesic's locally-informed programs in uncovering fate-specific differentiation. If Sceodesic indeed were to outperform, it would highlight the advantage of analyzing differential expression of locally-defined gene programs and reconstruction of gene-expression data using the log-Euclidean metric.

On RNA velocity studies of hematopoeisis (5780 cells, [22]) and murine pancreatic development (2531 cells, [25]) we showed that estimating Sceodesic gene programs and searching for driver programs among them allowed us to

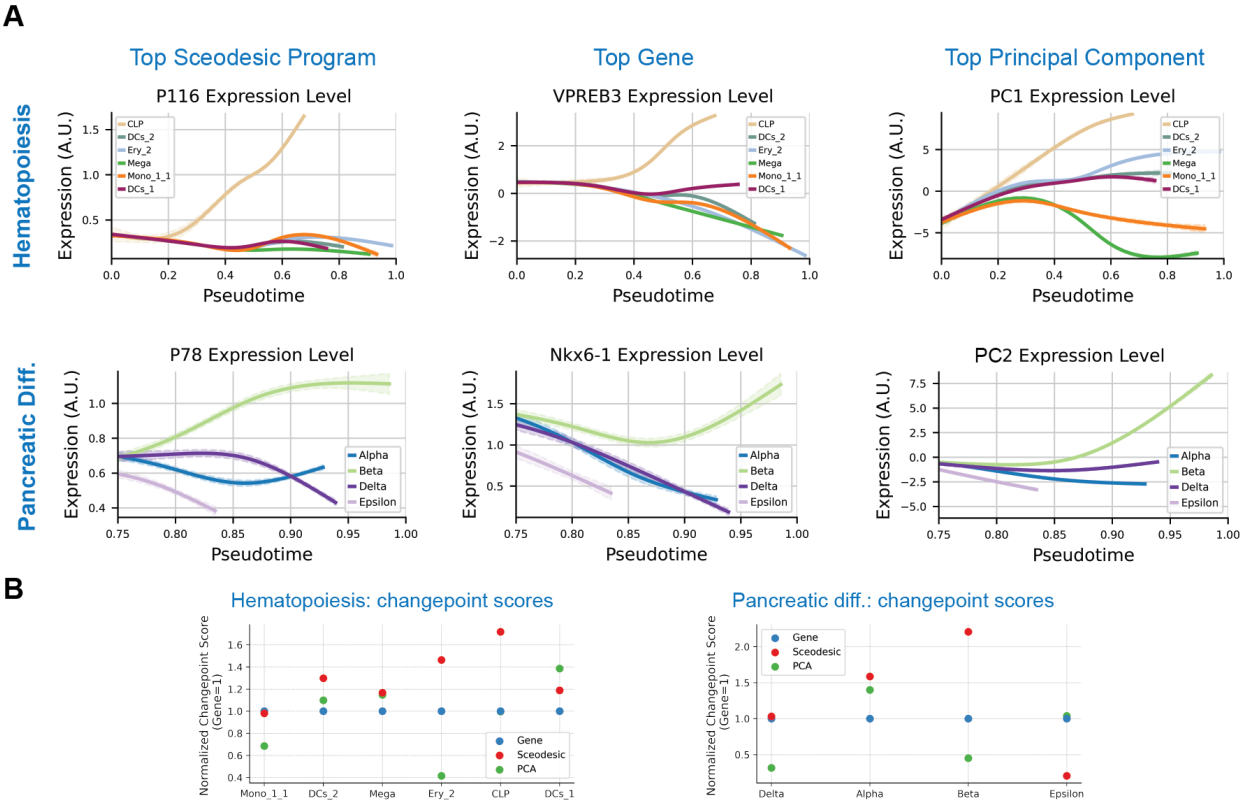

Figure 3: Comparison of distance metrics between cell cohorts: **A)** Pseudotime-course expression levels of driver programs, genes, and principal components. Each line represents expression for a single lineage. For hematopoiesis, the lineage of interest is the Common Lymphoid Progenitor (CLP). For pancreatic differentiation, it is Beta cells. **B)** Changepoint scores denote drivers' average ability to differentiate lineages of interest early in pseudotime. Because lineages vary in their differentiation speed, we show here scores within lineages normalized to Gene= 1. (A.U. = Arbitrary Units).

uncover earlier and stronger fate commitment than was possible with driver genes or PCs (see Sec. A.3 for methodological details). For example, Sceodesic identified the program that drives and predicts the CLP (common lymphoid progenitor) earlier than driver genes or driver principle components (Fig. 3A). This gene program was not only earlier and stronger in signal than both the driver genes and driver programs estimated via PCA, it were also more interpretable. We applied Enrichr to identify the cell types associated with the gene sets in the CLP-associated Sceodesic program. In addition to being relatively sparse (only a few dozen genes maximum per program), the top three associations (all p-vals <0.0002) were to lymphoid cells in different tissues: this is *precisely* the cell which CLP is the progenitor of. Additionally, the key genes in this program include IGKC, which is known to be relevant to the CLP lineage. In contrast, the top association of the driver PC does not have as clear of a connection to the CLP lineage. Taken together, the early differential expression of driver programs and the results from Enrichr suggest that Sceodesic recapitulates latent biological information that exhibits a high correlation to lineage specific fate probabilities.

Quantitative assessment of driver gene/program identification: To quantitatively compare Sceodesic programs' ability to capture early fate commitment versus genes or PCs, we considered how differential expression of a putative driver evolved over time. To score a driver, we computed the integral of the distance between the GAM corresponding to the lineage of interest and those of other lineages over time. This produced an average score of a driver's differential expression through time (see Sec. A.3 for details).

On both datasets tested, Sceodesic had the highest changepoint scores across a majority of lineages, capturing

fate commitment better than genes or PCs (Fig. 3B). Importantly, PCA-estimated driver programs frequently underperformed, suggesting that the global variation it captures may be fairly insensitive to gene coexpression relevant for biological phenomena like fate commitment. Sceodesic is not uniformly stronger than genes or PCs. For example, in the pancreas differentiation dataset, Sceodesic performs worse than genes in the Epsilon lineage, in which the driver gene identified was ghrelin. We believe this is because Epsilon cells are intimately related to the single gene ghrelin. In general, we expect both mechanisms driven by single genes and programs will occur. Thus Sceodesic provides a powerful and complementary way of discovering fate commitment phenomena that might otherwise be transient or weakly evidenced by single genes.

4.3 Sceodesic reveals disease-specific gene programs despite individual-level variability

We anticipate a powerful use of Sceodesic will be in studies where samples originate from different batches, individuals, or case-versus-control setups. In particular, consider the following setting that has remained difficult to analyze: there are two covariates (e.g., sample and disease-condition) and while one of them (disease-condition) is of primary interest, it is confounded by the other (sample); in a longitudinal study, one of these covariates may be time. Currently, this setting is usually handled by batch correction at the level of samples. But samples are either from healthy or diseased individuals, and correcting cross-sample variability may obscure disease-condition variability. We reasoned that Sceodesic could reveal disease-specific gene programs in such situations. The core intuition is two fold: a) Sceodesic programs are estimated in each cell cohort locally and b) if there are batch effects that are solely in terms of the first moment of gene expression, those will not affect the second moment-based Sceodesic.

We analyzed three datasets, covering diabetes (985,167 cells [26]), the flu (32,207 cells [27]), and Parkinson's disease (57,166 cells [28]); in each, samples were assayed from healthy individuals and patients. We stratified cell cohort construction by sample, and used Scanpy to performed a one-tailed t-test for the Sceodesic program most differentially expressed between diseased and healthy individuals. For comparison, we also performed the same differential expression analysis at the level of a) uncorrected gene expression (with count- and log-normalization), b) Scanorama's batch-corrected gene expression, and c) Harmony's batch-corrected PC scores. For each, we obtained the 100 highest-scoring genes, noting that some of these may be down-regulated, and used Enrichr [29] to perform a gene set enrichment study to identify enriched disease associations (diabetes and Covid) or Gene Ontology biological process (Parkinson's). All p-values were corrected using the Benjamini-Hochberg procedure (FDR α=0.05). In Table 1 we show the top six terms associated with the gene sets reported by each method, and highlight entries that are prima facie relevant to the disease. In some cases (e.g., Harmony on Parkinson's), the method reported no significant associations after multiple hypothesis correction.

Sceodesic consistently outperformed the other methods. In every case, its most significantly-associated term was relevant to the disease, with multiple relevant hits in each disease. For instance, Sceodesic was the only method to associate 'obesity' and 'Type II diabetes mellitus' with diabetes, and 'influenza' with the flu. It was also the only method to identify the importance of synaptic transmission in Parkinson's. This was not the case for other programs. Intriguingly, Sceodesic's diabetes program was also associated with Anorexia nervosa. We note that Sceodesic programs contain both positively and negatively-weighted genes, suggesting that Sceodesic could be indicating the relative *under-weighting* of Anorexia-related genes in diabetes.

5 Discussion

As single-cell datasets grow ever-larger and the complexity of biological measurements expands, there is an imperative to make maximal use of the data. But this is contingent on preserving mathematical consistency in how the data's variability is quantified. Using the log-Euclidean metric to operate on gene coexpression matrices represents a major step forward in providing such consistency. In turn, this conceptual advance together with a

Type II Diabetes (24 samples, 2 conditions)

No significant hits for Scanorama and Diff. Gene Expr.

Sceodesic	pval	Harmony (adj PC)	pval
Obesity	<0.001	Cancer	<0.001
Kidney cancer	<0.001	Carcinoma	0.002
Type II diabetes mellitus	<0.001	Chronic granulomatous disease	0.016
Melanoma	<0.001	Arthritis	0.016
Kawasaki disease	0.002	Colon cancer	0.016
Anorexia nervosa	0.003	**Glomerulonephritis**	0.019

Flu (9 samples, 2 conditions)

Sceodesic	pval	Harmony (adj PC)	pval	Scanorama (adj genes)	pval	Diff. Gene Expr.	pval
Lung disease	<0.001	Arthritis	<0.001	Arthritis	<0.001	Arthritis	<0.001
Arthritis	<0.001	Pneumonia	0.002	MWS	0.001	Peritonitis	<0.001
Influenza	<0.001	Histiocytosis	0.002	MEVA	0.001	**Lung disease**	<0.001
HIV	0.002	Leishmaniasis	0.003	**Lung disease**	0.002	Meningitis	<0.001
Meningitis	0.003	HLH	0.007	Meningitis	0.003	MWS	<0.001
TCP	0.005	Vasculitis	0.008	EPP	0.013	Histiocytosis	<0.001

Parkinson's (17 samples, 2 conditions)

No significant hits for Harmony and Scanorama

Sceodesic	pval	Diff. Gene Expr.	pval
Modulation Of Chemical Synaptic Transmission	0.002	**Response To Unfolded Protein**	<0.001
Detection Of Calcium Ion	0.010	Cellular Response To Heat	<0.001
Positive Regulation Of Synaptic Transmission	0.010	Chaperone-Mediated Protein Complex Assembly	<0.001
Cell Communication By Electrical Coupling Involved In Cardiac Conduction	0.010	**Chaperone Cofactor-Dependent Protein Refolding**	0.002
Regulation Of Cardiac Muscle Contraction By Regulation Of The Release Of Sequestered Calcium Ion	0.010	**Protein Stabilization**	0.002
Regulation Of Cardiac Muscle Contraction By Calcium Ion Signaling	0.017	Regulation Of Protein Modification By Small Protein Conjugation Or Removal	0.002

Table 1: Sceodesic identifies disease-specific gene programs despite multi-sample variability. In datasets where both sample-level and disease-level variability exists, Sceodesic outperforms data-integration approaches Harmony and Scanorama, as well as standard differential gene expression analysis. For Sceodesic's most differentially-expressed program, as well as corresponding gene sets from other methods, the top six associated disease (diabetes and flu) or GO biological process terms (Parkinson's disease) are shown. All p-values were adjusted using Benjamini-Hochberg correction, with cutoff FDR $\alpha < 0.05$. TCP: Thrombocytopenia; HLH: Hemophagocytic lymphohistiocytosis; EPP: Erythropoietic protoporphyria; MWS: Muckle-Wells syndrome; MEVA: Mevalonic aciduria

number of other innovations for characterizing local (condition volume normalization, PVD, local highly variable genes), enables Sceodesic to consolidate disparate and noisy cohort-level coexpression patterns into a study-wide set of biologically meaningful gene programs. This expands the set of complex biological phenomena that can be modeled with high fidelity, for example fate commitment (see Sec. A.4 for additional applications).

Existing work on gene coexpression, like Trajectorama [30], has shown the utility of analyzing differences in gene coexpression but also noted the pressing need to use metrics respecting the PSD manifold. Our novel contributions meet this need: For computational tractability, we posited the use of the log-Euclidean metric, and to ensure that it captured biological similarity well, we benchmarked it to existing metrics. Moreover, condition volume normalization ensures that estimation issues do not preclude use of the log-Euclidean metric. Sceodesic builds on prior fate commitment work, particularly CellRank, which leverages lineage fate probabilities to identify driver *genes*. We have extended this paradigm by analogously identifying driver *programs*. Fate commitment is about detecting changes.

As gene programs aggregate weaker signals from individual genes, they are a natural mechanism for detecting early fate commitment changes. Another related work is MetaCell [31], which partitions single-cell datasets into disjoint and homogeneous groups of profiles. While it does consider the properties of inter-metacell gene-gene covariance, it focuses on conventional Euclidean metrics for considering this. In contrast, we start with intra-cohort covariance and our use of the log-Euclidean metric enables consistent treatment of differences in inter-cohort variability.

There are a number of significant future extensions to this work. First, for large enough datasets, traditional analyses such as kNN-graphs could have the Euclidean distance replaced by inter-cohort log-Euclidean distances. Second, identification of meaningful programs is a crucial first step to inferring gene regulatory networks (GRNs) [32–35]. In particular, they fit nicely into a mechanistic model of GRNs where one can investigate which transcription factors are driving the programs estimated by Sceodesic. Third, recent work has focused on using known gene sets to estimate gene programs [8], and Sceodesic parameters could be calibrated to achieve a certain degree of agreement with such sets when they are known. Finally, to enforce greater heterogeneity among cells within the same cohort, it would be useful to consider extending the method to include soft clustering methods.

To sum, we have shown that the log-Euclidean metric is an efficient and principled way to treat gene coexpression, both on its own and in the service of extracting meaningful gene programs that can guide biological investigation. As datasets grow ever larger, we expect this metric's application, as in Sceodesic, will be indispensable for maximally extracting signal from these data.

Acknowledgements We thank Alexander Wu and Kapil Devkota for helpful discussions. The authors were supported by a Duke Univ grant to RS. Some of the figures were created using Biorender. We used `GNU-Parallel` to parallelize some of our analysis.

Declaration of Interests None

References

[1] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular systems biology*, vol. 15, no. 6, p. e8746, 2019.

[2] B. Hie, J. Peters, S. K. Nyquist, A. K. Shalek, B. Berger, and B. D. Bryson, "Computational methods for single-cell rna sequencing," *Annual Review of Biomedical Data Science*, vol. 3, pp. 339–364, 2020.

[3] R. Singh, B. L. Hie, A. Narayan, and B. Berger, "Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities," *Genome biology*, vol. 22, no. 1, pp. 1–24, 2021.

[4] R. Bhatia, *Positive definite matrices*. Princeton university press, 2009.

[5] L. Haghverdi and L. S. Ludwig, "Single-cell multi-omics and lineage tracing to dissect cell fate decision-making," *Stem Cell Reports*, vol. 18, no. 1, pp. 13–25, 2023.

[6] L. Li, S. Bowling, S. E. McGeary, Q. Yu, B. Lemke, K. Alcedo, Y. Jia, X. Liu, M. Ferreira, A. M. Klein, *et al.*, "A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells," *Cell*, 2023.

[7] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein, "Lineage tracing on transcriptional landscapes links state to fate during differentiation," *Science*, vol. 367, no. 6479, p. eaaw3381, 2020.

[8] R. Z. Kunes, T. Walle, M. Land, T. Nawy, and D. Pe'er, "Supervised discovery of interpretable gene programs from single-cell data," *Nature Biotechnology*, pp. 1–12, 2023.

[9] R. Lopez, A. Gayoso, and N. Yosef, "Enhancing scientific discoveries in molecular biology with deep generative models," *Molecular systems biology*, vol. 16, no. 9, p. e9198, 2020.

[10] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nat Methods*, vol. 16, pp. 1289–1296, Dec 2019.

[11] B. Hie, B. Bryson, and B. Berger, "Efficient integration of heterogeneous single-cell transcriptomes using scanorama," *Nature biotechnology*, vol. 37, no. 6, pp. 685–691, 2019.

[12] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park, "Bbknn: fast batch alignment of single cell transcriptomes," *Bioinformatics*, vol. 36, no. 3, pp. 964–965, 2020.

[13] A. P. Wu, J. Peng, B. Berger, and H. Cho, "Bayesian information sharing enhances detection of regulatory associations in rare cell types," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i349–i357, 2021.

[14] B. Hie, H. Cho, B. Bryson, and B. Berger, "Coexpression uncovers a unified single-cell transcriptomic landscape," *bioRxiv*, p. 719088, 2019.

[15] C. M. Crainiceanu, B. S. Caffo, S. Luo, V. M. Zipunnikov, and N. M. Punjabi, "Population value decomposition, a framework for the analysis of image populations," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 775–790, 2011.

[16] M. Tanaka and K. Nakata, "Positive definite matrix approximation with condition number constraint," *Optimization Letters*, vol. 8, no. 3, pp. 939–947, 2014.

[17] "CZ CELLxGENE Discover." `https://cellxgene.cziscience.com/`, 2023. Accessed: Nov 1, 2023.

[18] L. Sikkema, C. Ramírez-Suástegui, D. C. Strobl, T. E. Gillett, L. Zappia, E. Madissoon, N. S. Markov, L.-E. Zaragosi, Y. Ji, M. Ansari, *et al.*, "An integrated cell atlas of the lung in health and disease," *Nature Medicine*, pp. 1–15, 2023.

[19] J. Cao, D. R. O'Day, H. A. Pliner, P. D. Kingsley, M. Deng, R. M. Daza, M. A. Zager, K. A. Aldinger, R. Blecher-Gonen, F. Zhang, *et al.*, "A human cell atlas of fetal gene expression," *Science*, vol. 370, no. 6518, p. eaba7721, 2020.

[20] A. D. Reed, S. Pensa, A. Steif, J. Stenning, D. Kunz, P. He, A.-J. Twigger, K. Kania, R. Barrow, I. Goulding, *et al.*, "A human breast cell atlas mapping the homeostatic cellular shifts in the adult breast," *bioRxiv*, pp. 2023–04, 2023.

[21] D. Reichart, E. L. Lindberg, H. Maatz, A. M. Miranda, A. Viveiros, N. Shvetsov, A. Gärtner, E. R. Nadelmann, M. Lee, K. Kanemaru, *et al.*, "Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies," *Science*, vol. 377, no. 6606, p. eabo1984, 2022.

[22] M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, *et al.*, "Cellrank for directed single-cell fate mapping," *Nature methods*, vol. 19, no. 2, pp. 159–170, 2022.

[23] M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, and D. Pe'Er, "Characterization of cell fate probabilities in single-cell data with palantir," *Nature biotechnology*, vol. 37, no. 4, pp. 451–460, 2019.

[24] S. N. Wood, *Generalized additive models: an introduction with R*. CRC press, 2017.

[25] A. Bastidas-Ponce, S. Tritschler, L. Dony, K. Scheibner, M. Tarquis-Medina, C. Salinno, S. Schirge, I. Burtscher, A. Böttcher, F. J. Theis, *et al.*, "Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis," *Development*, vol. 146, no. 12, p. dev173849, 2019.

[26] Parse Biosciences, "How to analyze a 1 million cell data set using scanpy and harmony." `https://rb.gy/747zfi`. Accessed on October 11, 2023.

[27] J. S. Lee, S. Park, H. W. Jeong, J. Y. Ahn, S. J. Choi, H. Lee, B. Choi, S. K. Nam, M. Sa, J.-S. Kwon, *et al.*, "Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19," *Science immunology*, vol. 5, no. 49, p. eabd1554, 2020.

[28] A. J. Lee, C. Kim, S. Park, J. Joo, B. Choi, D. Yang, K. Jun, J. Eom, S.-J. Lee, S. J. Chung, *et al.*, "Characterization of altered molecular mechanisms in parkinson's disease through cell type–resolved multiomics analyses," *Science Advances*, vol. 9, no. 15, p. eabo2467, 2023.

[29] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[30] B. Hie, H. Cho, B. Bryson, and B. Berger, "Coexpression enables multi-study cellular trajectories of development and disease," 2019.

[31] Y. Baran, A. Bercovich, A. Sebe-Pedros, Y. Lubling, A. Giladi, E. Chomsky, Z. Meir, M. Hoichman, A. Lifshitz, and A. Tanay, "Metacell: analysis of single-cell rna-seq data using k-nn graph partitions," *Genome biology*, vol. 20, no. 1, pp. 1–19, 2019.

[32] J. Ding, C. Lin, and Z. Bar-Joseph, "Cell lineage inference from snp and scrna-seq data," *Nucleic Acids Research*, vol. 47, no. 10, pp. e56–e56, 2019.

[33] L. Wang, N. Trasanidis, T. Wu, G. Dong, M. Hu, D. E. Bauer, and L. Pinello, "Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics," *Nature Methods*, vol. 20, no. 9, pp. 1368–1378, 2023.

[34] R. Singh, A. P. Wu, A. Mudide, and B. Berger, "Unraveling causal gene regulation from the rna velocity graph using velorama," *bioRxiv*, pp. 2022–10, 2022.

[35] B. Van de Sande, C. Flerin, K. Davie, M. De Waegeneer, G. Hulselmans, S. Aibar, R. Seurinck, W. Saelens, R. Cannoodt, Q. Rouchon, *et al.*, "A scalable scenic workflow for single-cell gene regulatory network analysis," *Nature Protocols*, vol. 15, no. 7, pp. 2247–2276, 2020.

[36] Z. Lin, "Riemannian geometry of symmetric positive definite matrices via cholesky decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 40, no. 4, pp. 1353–1370, 2019.

[37] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM journal on matrix analysis and applications*, vol. 29, no. 1, pp. 328–347, 2007.

[38] "Open problems in single cell analysis." `https://openproblems.bio/`. Accessed: 2023-11-02.

[39] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular systems biology*, vol. 17, no. 1, p. e9620, 2021.

A Appendix

A.1 Background

Additional discussion of the $\mathcal{S}_+^n$ Riemannian manifold. Symmetric positive definite matrices $\mathcal{S}_+^n$ of fixed dimension n can be viewed as a smooth convex submanifold of the space $\mathbf{R}^{n(n+1)/2}$ with the inherited euclidean metric. But the direct application of this metric on $\mathcal{S}_+^n$ is problematic, especially because it induces finite distance between points in $\mathcal{S}_n^+$ and matrices with zero or negative eigenvalues [36]. Similarly, the Frechet mean obtained using the euclidean norm induces a swelling effect, where the determinant of the euclidean norm is strictly larger than the determinants of sample points [37]. This might introduce unwanted distortions in our calculation of distance between the correlation matrices.

Many alternative norms have been developed to mitigate the finite-distance and swelling effects inherent in the Euclidean norms [36, 37]. The simplest substitute is to develop a norm in $\mathcal{S}_+^n$ that induces isometry in the smooth map $f: GL(n,\mathbb{R}) \rightarrow \mathcal{S}_+^n, f(A) = A^T A$ [4]. The lower-bound approximation of the geodesic distance induced by this isometric norm is the logarithmic distance, given by $d_{\log}(A,B) = \|\log(A) - \log(B)\|_F$, where $A, B \in \mathcal{S}_+^n$ and $\log(A)$ represents the matrix logarithm of A. This $d_{\log}$ metric solves the finite-distance issue, as the distance between any point in $\mathcal{S}_+^n$ and matrices with zero or negative eigenvalues is infinity.

A.2 Method Details

Selection of Locally Variable Genes: To ensure that the selection of highly variable genes is not dominated by the cells constituting larger cell cohorts, we utilize the notion of 'locally variable genes' which allows for the selection of h distinct genes that are highly variable across and within cohorts, rather than only across the entire study. Consider n distinct cohorts c_1, c_2, ..., c_n and a parameter h for the desired number of genes to select. Then, within each *cohort* i, we use Seurat's HVG selection procedure (implemented in Scanpy) to obtain the top h genes, which we denote by the set:

$$G_i = \{g_1^i, g_2^i, ..., g_h^i\}$$

We then score each gene by the number of times it appears in a cohort-specific HVG list, i.e. for a gene g:

$$s(g) = \sum_{i=1}^{n} \mathbb{I}_{g \in G_i}$$

We determine the final list of 'locally variable genes' by selecting the h highest scoring genes.

Proof of the optimality of S_i^{log} :

We now show that the solution $\hat{S}_i$ corresponding to

$$[\hat{S}_i^{log}]_{jj} = [\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U}]_{jj} \quad \text{for } j = 1...h$$

is the optimal feasible solution to the covariance matrix reconstruction problem:

$$\min_{S_i} \|\log(\tilde{\Sigma}_i) - \log(\hat{U} S_i \hat{U}^T)\|_F \quad \text{where } \tilde{\Sigma}_i = U_i \tilde{D}_i U_i^T,$$

subject to the constraint that S_i is diagonal. The objective function of interest, which minimizes the distance from the reconstruction matrix $\hat{U} S_i \hat{U}^T$ to the condition-normalized sample covariance matrix $\tilde{\Sigma}_i$ under the log-Euclidean metric, is equivalent to:

$$\min_{S_i} \|\log(\tilde{\Sigma}_i) - \hat{U} \log(S_i) \hat{U}^T\|_F$$

Factoring on the left and right yields:

$$\min_{S_i} \|\hat{U}(\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U} - \log(S_i))\hat{U}^T)\|_F$$

But since the columns of $\hat{U}$ are orthonormal and do not change the norm of any matrix, the preceding minimization is equivalent to minimizing the simpler objective:

$$\min_{S_i} \|\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U} - \log(S_i)\|_F$$

Setting $\log(S_i) = S_i^{log}$, we have:

$$\min_{S_i^{log}} \|\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U} - S_i^{log}\|_F$$

But since S_i^{log} must be diagonal because S_i must be diagonal, the optimal feasible solution is just the diagonal of the first term $\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U}$, in other words:

$$[\hat{S}_i^{log}]_{jj} = [\hat{U}^T \log(\tilde{\Sigma}_i)\hat{U}]_{jj} \quad \text{for } j = 1...h$$

This concludes the proof. In the context of our method, we note that this solution is exact when $\lambda = 0$ i.e. when regular PCA is used in the population value decomposition.

The difference of two Hermitian matrices with same eigenvectors is determined by their eigenvalues For two Hermitian matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ with common eigenvectors $\boldsymbol{U}$, the Frobeinus norm of their difference (or $\|\boldsymbol{A} - \boldsymbol{B}\|_F$) equals the ℓ^2 norm of difference between their eigenvalues.

Proof: $\boldsymbol{A}$ and $\boldsymbol{B}$ can be both be diagonalized by a shared basis $\boldsymbol{U}$, as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D_A}\boldsymbol{U}^*$ and $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D_B}\boldsymbol{U}^*$ (where $\boldsymbol{U}$ is unitary and $\boldsymbol{D_A}, \boldsymbol{D_B}$ are diagonal matrices containing eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively). Now, the equivalent definition of $\|\cdot\|_F$ is

$$\|\boldsymbol{A}\|_F = (tr(\boldsymbol{A}^*\boldsymbol{A}))^{1/2} \tag{1}$$

Simplifying $\|\boldsymbol{A} - \boldsymbol{B}\|_F$ using (1), we get

$$\|\boldsymbol{A} - \boldsymbol{B}\|_F = \left(tr(\boldsymbol{U}^*(\boldsymbol{D_A} - \boldsymbol{D_B})\boldsymbol{U})^2\right)^{1/2} = \left(tr(\boldsymbol{D_A} - \boldsymbol{D_B})^2\right)^{1/2} = \|\lambda^{(\boldsymbol{A})} - \lambda^{(\boldsymbol{B})}\|_2 \tag{2}$$

Where, $\lambda^{(\boldsymbol{A})}$ and $\lambda^{(\boldsymbol{B})}$ are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ corresponding to the unitary matrix $\boldsymbol{U}$

Hyperparameters M, λ, v_{pvd} A condition value M between 50 and 1000 tends to work well, with lower values being more appropriate if cell cohorts have fewer cells on average. λ and v_{pvd} control how closely the sparse PCA fits to the local features. High values of λ lead to more sparsity in global program estimation and thus more interpretable programs that reconstruct local programs less faithfully. Thus, a v_{pvd} between 70% and 95% tends to work well and anything from 0.01 to 0.1 works well for λ, though the exact values that work for the parameter will vary based on h, among other dataset specific properties.

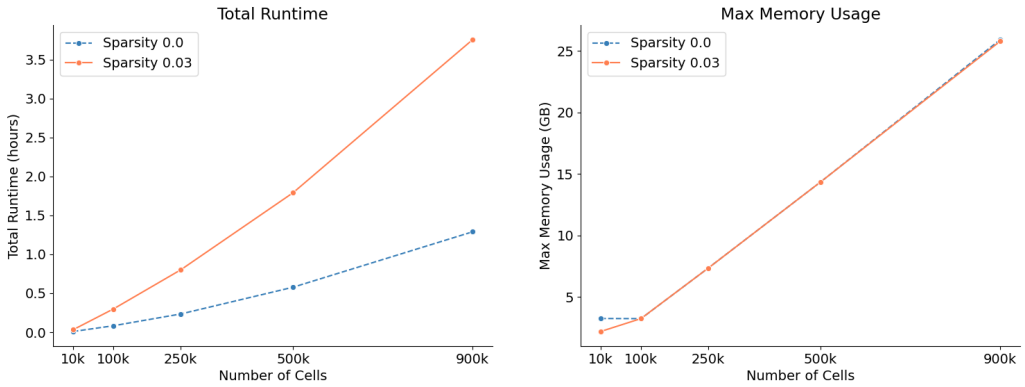

Figure A.1: Runtime and memory requirements

.

Robustness to hyperparameters. The optimal values of n and h are dataset dependent, and we have provided recommendations for choosing these values in the main text and immediatly above. In this section we demonstrate robustness of Sceodesic to the choice of M and λ. We evaluate the robustness of Sceodesic to the choice of M and λ by assessing the variability in cell type prediction performance. A dataset consisting of 100,955 FACS-isolated cells across 17 *C. elegans* samples was obtained from [38]. Cells were pre-annotated by cell type (169 total cell types) and were sequenced in nine experimental batches. The task was to predict the cell-type using gene expression data; cells from one experimental batch were used as a hold-out set for evaluating performance, and Sceodesic programs were estimated using the remaining cell-gene expression data, which were used as features in a $l2$-regularized logistic regression classifier. Classification performance was assessed using accuracy and weighted f1 score. As shown in Tables A.1 and A.2, classification performance is stable across a range of choices for λ and M. Our default values are $\lambda = 0.03$ and $M = 50$.

	0.03	**0.06**	**0.09**
50	0.730977	0.704014	0.711204
500	0.722588	0.702816	0.718993

Table A.1: Cell type prediction performance evaluated by classification accuracy.

	0.03	**0.06**	**0.09**
50	0.746619	0.720146	0.722764
500	0.729649	0.706524	0.728353

Table A.2: Cell type prediction performance evaluated by weighted f1 score.

Runtime and memory requirements. Runtime and memory requirements both scale efficiently with dataset size (Figure A.1). All benchmarking runs were performed on the same dataset of diabetes patients provided by Parse Biosciences consisting of 985,167 cells and 26,862 genes [26]. Runtime was monitored using the python `time` module. Memory usage was monitored using the UNIX command line tool `ps`, with snapshots taken every 60 seconds. Data was subsampled uniformly at 10,000, 100,000, 250,000, 500,000, and 900,000 cells using `sc.pp.subsample`. We ran Sceodesic with $\lambda = 0.0$ for Sceodesic program estimation without sparsity, and $\lambda = 0.03$ for sparse Sceodesic program estimation.

A.3 Results & Analysis

Computing the μ, Σ, and $\log(\Sigma)$ distance metrics. We started by computing n cell cohorts using k-means clustering (in PC space). For the 2.3 million cell lung dataset, n was set to 2000. For the three other smaller datasets, i.e. the breast atlas, the fetal development survey, and the heart atlas, a value of $n=1500$ was used. We used expert cell type annotations available with each dataset. We then assigned each cohort i a cell type q if at least a majority of cells in that cohort had a particular cell type. We did not consider cohorts with a mere plurality cell-type, setting the necessary threshold for cohorts cell type purity to $p=0.5$. Furthermore, to generate 'labels' of similarity or difference between pairs of cell cohorts with an assigned cell type, we used the following method: If two cell cohorts had the same cell-type, then the label of the pair was 0. If they differed, the label was 1. This labelling captures the notion of biological similarity versus difference in a binary setting.

For each cohort, d_μ was computed on the normalized gene count data. Then, d_Σ and $d_{\log(\Sigma)}$ were computed as follows: First data was subsetted to the first 500 highly variable genes determined via Scanpy with dispersion ranking for all datasets. Then, raw covariance matrices were computed for each cohort and condition volume normalization was applied with condition number 50 for the inputs of both d_Σ and $d_{\log(\Sigma)}$.

We computed all three metrics for all unique pairs of clusters with cell types. We then treated the distance generated by each distance metric for each cluster pair (c_i,c_j) as a predictor variable $X_{ij}\in\mathbb{R}$ and the cell pair label as the target variable $Y_{ij}\in\{0,1\}$. We then computed the AUROC for each distance, where the X's were treated as a 1-D classifier of the Y's. If a metric had a high AUROC, it indicated that the metric captured biological similarity well.

Additionally, as a matter of robustness, we found that empirically, the ordinal ranking of the performance of the distance metrics was not significantly affected if instead of the main setting described above we 1) Used the raw covariance matrix to compute the Euclidean distances between cell cohorts, 2) if smaller values of the highly variable gene count were used e.g. 300, 400.

Fate prediction For our analysis on fate prediction, we used CellRank. Using their API, we performed the following analysis for each dataset: First, we estimated for each cell a position in pseudotime via Palantir and a cell fate probability distribution. Then, we estimated putative driver genes, which have maximally correlated expression with the fate probabilities of a particular lineage. We then computed principal components, as well as Sceodesic programs. For the bone marrow dataset, we used 40 cohorts, 200 locally variable genes, and sparsity of 0.07. For the pancreas dataset, we used 20 cohorts, 200 locally variable genes, and sparsity of 0.06. All other parameters were at their default values. The difference in cohort number is due to the bone marrow dataset having a small multiple of cells that are present in the pancreas dataset. We then performed the same driver analysis as was done in genes, identifying putative drivers by finding the principle component or program that had maximum correlation with a particular lineage's fate probabilities. We then fit GAMs to these expression levels, one for each lineage with cells and their expression weighted by their fate probabilities as estimated by CellRank. These GAMs constitute the plots seen in the main text.

Quantitative analysis of fate prediction performance Given the above discussion of generalized additive models, we now detail how we quantitatively characterized the strength of differentiation that a putative driver exhibits for a particular target lineage. Given a putative driver d and given a target lineage $i\in[L]$, where L is the total number of lineages, we represent each lineage as a GAM f_i. Then, at a particular point in pseudotime $t\in[0,1]$, the degree of differentiation from other lineages, which we call the changepoint score, is quantified as:

$$q(t,i,d)=\min_{j\neq i}||f_i^d(t)-f_j^d(t)||,i,j\in[L],$$

in other words, the minimum distance from the GAM of the target lineage to the GAM representing a different lineage. Higher levels of distance denote more differentiation. In line with this, the average changepoint score is

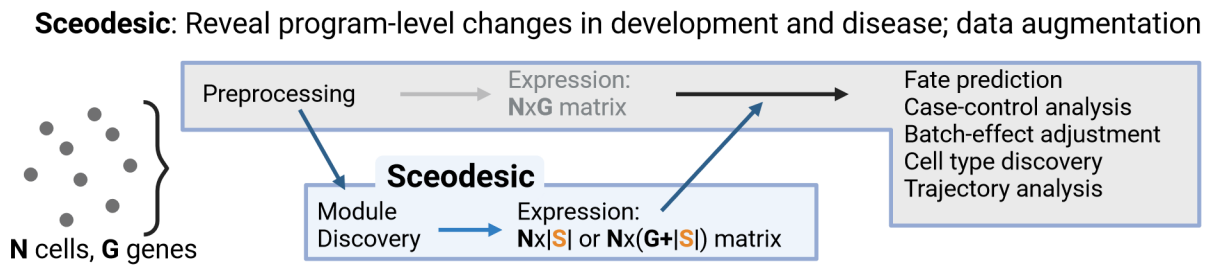

Figure A.2: Sceodesic fits seamlessly into existing single-cell workflows (e.g. scverse)

the integral over time of this score up to pseudotime $\rho < 1$ to spot changes prior to the final stretch of pseudotime, at which point fate commitment has likely already occurred:

$$\int_0^{\rho} q(t,i,d)dt$$

A standard choice for the bound of integration would be 0.75, which is what we used for the bone marrow dataset. However, in some cases differentiation among all lineages occurs very late in pseudotime, as was the case in the pancreas dataset. We therefore used 0.85 for that dataset, as almost no lineages were differentiating across any drivers by 0.75.

A.4 Broader applicability of Sceodesic

We anticipate Sceodesic to be useful in a variety of scRNA-seq analyses. It only requires transcript count information and produces a feature vector per cell; all cells in a cell cohort share the same feature values. These features can be used as-is, or concatenated with other features like principal components to augment existing off-the-shelf analyses (Figure A.2). Leveraging two datasets from a public benchmarking study focused on cell type prediction, we explored the potential enhancement Sceodesic could contribute to this task. A logistic regression on PCA was previously determined to be the optimal baseline [38]. When Sceodesic embeddings were employed directly as the input to logistic regression, the performance was comparable to single-cell annotation methods like scanVI [39]. Concatenating PCA-based and Sceodesic embeddings improved performance substantially compared to solely using Sceodesic embeddings, and slightly outperformed the current best approach. These diverse applications collectively accentuate the versatility and practical utility of Sceodesic in handling real-world challenges encountered in scRNA-seq data analysis.