CocycleHunter: cohomology-based circular gene set enrichment and genetic phase estimation in single-cell RNA-seq data.

Kelly Maggs*[1,2,3], Markus Youssef*[4,5], Cyril Pulver[5], Jovan Isma[8,11], Tâm J. Nguyên[5,6], Wouter Karthaus[5,7], Kathryn Hess[4,5,6], and G. Paolo Dotto.[8,9,10]

[1]Max Planck Institute for Cell Biology and Genetics, Dresden
[2]Max Planck Institute, Center for Systems Biology, Dresden
[3]Faculty of Mathematics, Technische Universität Dresden, Dresden
[4]UPHESS, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
[5]Faculty of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
[6]Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland.
[7]Swiss Institute for Experimental Cancer Research (ISREC)
[8]Cutaneous Biology Research Center, Department of Dermatology, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, USA
[9]Service d'Oto-rhino-laryngologie et chirurgie cervical faciale, Centre Hospitalier Universitaire Vaudois (CHUV), Université de Lausanne (UNIL), Lausanne, Switzerland
[10]International Cancer Prevention Institute, Epalinges, Switzerland
[11]Department of Immunobiology, University of Lausanne (UNIL), Switzerland.

January 9, 2025

Abstract

In this paper we introduce cocycle hunter (cHunter), a tool for identifying and exploiting diverse circular structure in single-cell RNAseq data. Our method is based on persistent cohomology in dimension one, which we use to define a system of statistically enriching gene sets for circular structure. We also develop a powerful, cohomology-based technique for estimating the phase of genes exhibiting cyclic expression patterns. We validate our method on synthetic RNA transcription models, as well as real datasets studying the cell cycle, tissue regeneration and senescence in diverse experimental conditions. Our work points towards cocycles as pervasive mathematical structures of the transcriptional program.

*Shared first-authorship. Correspondence: maggs@mpi-cbg.de.

1 Introduction

The concept of a transcriptomic phase space in biology goes back to the 1950s and Waddington's epigenetic landscape [Wad57]. The overarching vision in this paradigm is to associate the geometric and topological features of this sampled phase space with their underlying dynamical biological processes. For example, clustering may correspond to cell type; the density within clusters to heterogeneity of gene expression within a cell type; and a linear trajectory to a pathway of cell differentiation (Fig. 1). *Circular* structures in the epigenetic genetic landscape arise from diverse dynamic biological phenomena. For example, there are periodic biological processes such as the cell cycle [SFJ+20, ROD+22, LWHC20, LLT+24] and circadian rhythm [TGN23, PHY+21] that recur multiple times. Other processes may only occur once, such as negative feedback loops that drive a response to a stimulus back to homeostasis. Such processes are characterized by an initial wave of gene expression triggered by an external stimulus, followed by a transcriptional response that brings the system back to its resting state (Fig. 1C).

The epigenetic landscape is not measurable in practice; however, a single-cell experiment serves as a finite point cloud estimate of a snapshot in time [WZXW11, vDSN+18]. Reconstructing time-dependent expression patterns from such an estimate is a difficult proposition, and in theory requires attributing global geometric features to the underlying dynamics that generated them. This paper presents a topological approach to addressing two fundamental challenges in the *ab initio* exploitation of circular structure in single cell RNA seq data for estimating gene expression dynamics. The first challenge is *identifying* a circular structure in single-cell RNA expression data. Given that finding informative geometry in the full, high-dimensional data is not always possible, one needs a process of first detecting subspaces with meaningful circular structure. The second is working backwards from the phase space to determine the dynamic ordering of cell and gene phases driving an underlying circular process (Fig. 1D).

Circular analysis using cHunter To address the aforementioned challenges, our tool *cHunter* has two main components based on the identification and exploitation of circular structure via *cohomology*.

1. **Circular gene set enrichment:** We first adapt traditional gene set enrichment analysis [STM+05] to persistent-cohomology based *circular* gene set enrichment. Notably, this analysis reveals that *many* circular structures are present over different resolutions and gene sets. Our approach is to rank their importance in a principled way, for which we develop a pipeline to quantify whether a user-provided gene set contains a statistically significant circular structure.

2. **Inferring circular pseudo-dynamics:** While topology identifies circular structures, the *geometry* of how the given topological feature is embedded in transcriptomic phase space allows us to estimate dynamic properties of expression:

   (a) **Estimating circular cell phases:** the identified cohomology class induces a canonical *harmonic* circle-valued map that is used for estimating cell phases whose differential is a model of dynamics from topological structure. Fol-

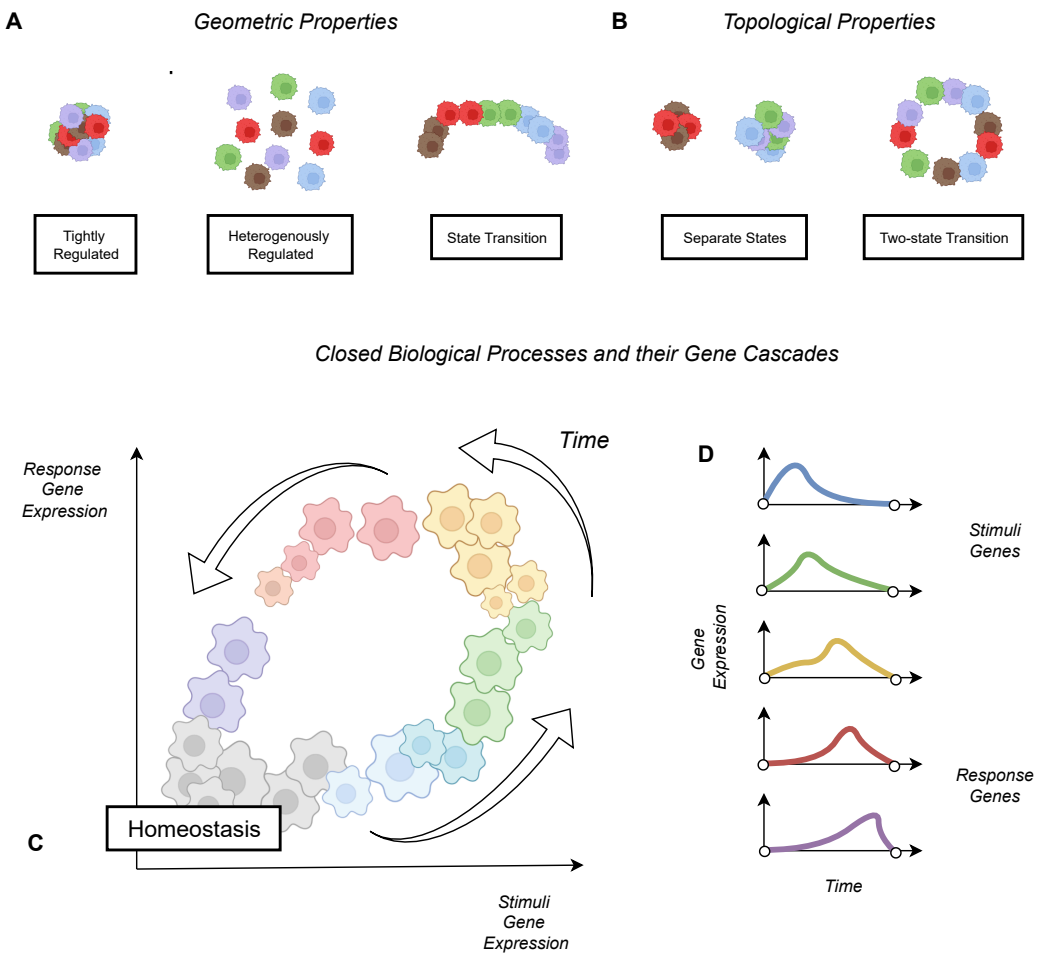

Figure 1: Biological interpretations of transcripto-geometric phenomena.

lowing [dSMVJ11], we use a simplicial cochain model to estimate this map in point-cloud data.

(b) **Estimating gene phases:** we adapt cyclicity analysis [BS16] using differential forms to produce a linear projection that best captures the cohomology class. In certain cases, the axes of this plane estimate the peak phases of genes relative to the circular dynamics.

Our hypothesis is that the robustness of topological invariants permits the broad discovery of circular structures. Thus, we have focused on demonstrating that our topological method can identify such processes in a range of diverse biological settings and generate meaningful biological hypotheses. We show that our method is able to estimate interactions between genes in synthetic ODE models of feedback between RNA transcription and protein translation (Results 2.3). We also study three biological datasets. Firstly, we study the cell cycle in human cycling dermal fibroblast (Results 2.4), where we validate that our model produces sensible gene and cell phase estimates based on external phase estimation. Continuing with human dermal fibroblasts, we study a new dataset to illuminate the relations amongcell cycle, androgen receptor silencing, and senescence (Results 2.5). Finally, we study the role of androgen response

in tissue regeneration in the prostate from publicly available mouse model data (Results 2.6). Together, these examples highlight the broad applicability of our conceptual framework.

Related Work Many tools exist for modeling periodic biological phenomena, which also center on mapping the data into S^1. These are in theory generally applicable, but with examples mainly focused on the cell cycle [LWHC20, LLT+24, ROD+22, SFJ+20] and circadian rhythm [AFL22, TGN23, PHY+21]. In contrast to these, our approach is first to statistically identify a global topological feature and work backward to recover likely gene cascades that generate it. We believe this step serves as an important control in determining the statistical likelihood that there exists a circular model of good-fit in the first place.

Our tool is built on top of two existing classes of methods. First, the use of a harmonic cocycle in the circular coordinates algorithm goes back to [dSMVJ11]. We make extensive use of the later variants in [Per20, PP23] which deal with sparseness and density. Secondly, our algorithm is a direct generalization of cyclicity analysis for paths introduced in [BS16].

TDA has been used more generally to analyse single cell RNA seq data. These include analysing mapper representations [RMR+20, CR20], multi-species analysis [SDB+23] and cluster-diffusion models [KDS+23]. Combinatorial Hodge theory specifically is used in conjunction with TDA methods for single-cell analysis in [IH23, GYC19].

2 Results

2.1 *cHunter*: A pipeline for identifying statistically significant closed biological processes at many resolutions

Gene set enrichment is a highly successful statistical tool in genomic data analysis [STM+05]. We define *circular* gene set enrichment to be the process of deciding whether the projection of a single-cell experiment onto a given gene set contains a statistically significant circular structure. We hypothesized that there exist many closed biological processes observable geometrically in different subsets and at different resolutions. We developed circular gene set enrichment as a principled statistical method for ranking and sorting such geometric features.

Gene set projections The challenge of identifying a circular feature in high-dimensional single-cell expression space is that it requires *all* genes to be involved. Our approach was to restrict our search to *gene set projections* onto sub-spaces spanned by biologically relevant genes. This enabled (1) easier identification of closed sub-processes and (2) a preliminary biological interpretation based on the annotation of the gene sets. Our recommended pipeline is to use hallmark gene sets from the Molecular Signatures Database [LBT+15]. However, any user-provided gene set may be used.

Persistent cohomology We used one-dimensional persistent cohomology (Methods 4.4) to identify all possible circular features (up to continuous deformation) at all distance resolutions within a gene set projection. The points in the persistence diagram

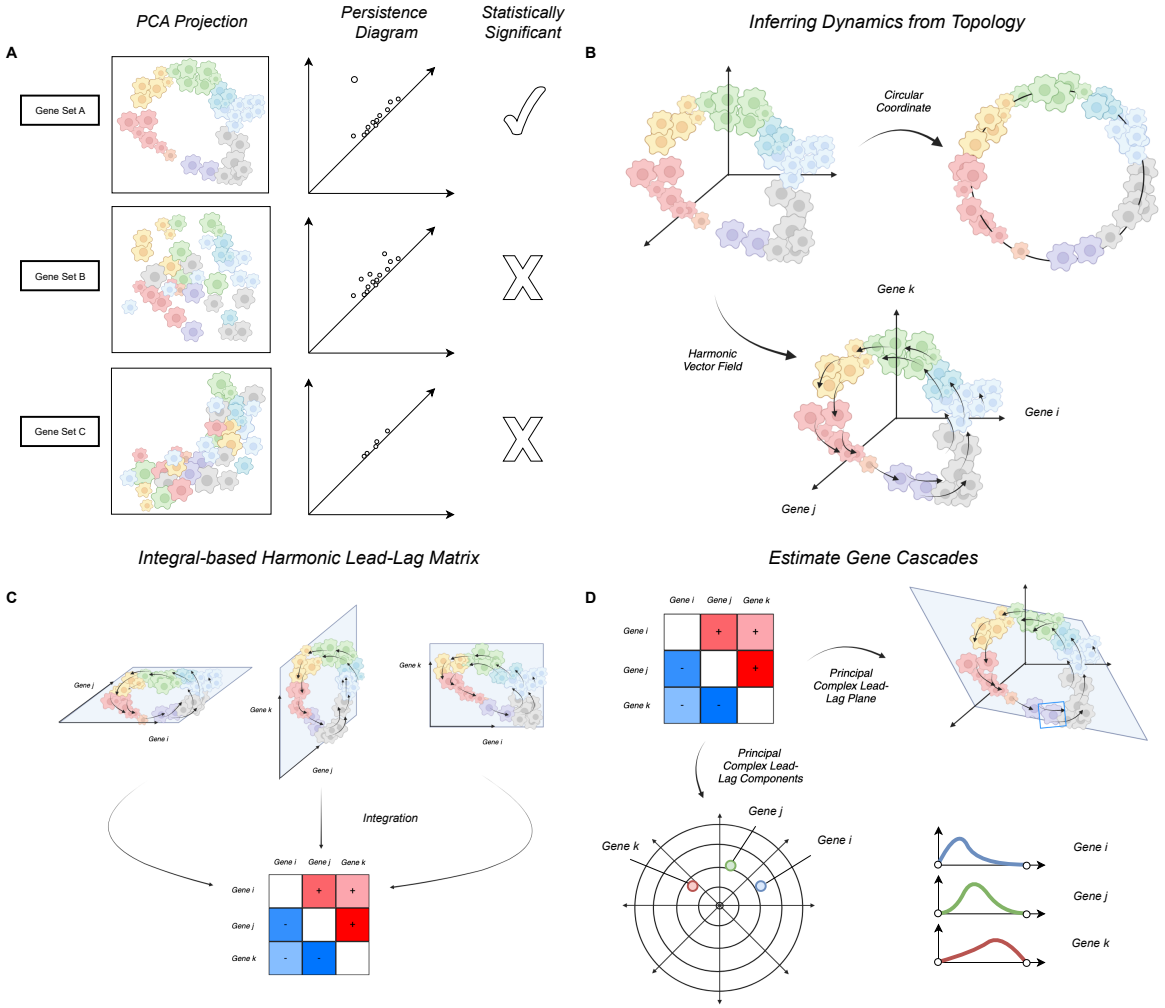

Figure 2: The basic *cHunter* pipeline for circular gene set enrichment and harmonic cyclicity analysis. **(A)** Circular gene set enrichment analysis consists of testing low-dimensional PCA projections different gene sets by evaluating a ring score on their persistent cohomology diagram in dimension 1. A permutation test determines statistical significance. **(B)** Once a significant 1-cocycle is identified, its harmonic representative is used to extract a circular coordinate in the usual way. This corresponds to a discrete vector field on the Rips complex at a given parameter. **(C)** The complex and cochain are projected onto each pairwise ij-gene plane and integrated against the 1-form $w_{ij} = x_i dx_j - x_j dx_i$ to calculate the lead-lag matrix. **(D)** Eigenvector analysis of the lead-lag matrix determines the plane that maximizes circular information, the complex components of which generate estimations of gene cascade phases associated to the closed process.

(Figure 2A, middle) of these projections each correspond to a circular structure that is born and dies at resolutions given by the x and y coordinates, respectively. Points further from the diagonal correspond to circular features that *persist* across a longer range of resolutions.

Statistical significance Our end goal was an automatic framework for deciding whether a given gene set projection contains a dominant circular structure. Since persistence diagrams are computed across gene set projections of varying dimension, it was necessary to devise a statistical framework for inter-gene set comparison. We defined a summary statistic over persistence diagrams called the *ring score* (Methods 4.3), which roughly measures whether there is a uniquely important circular structure. We employed a permutation test wherein random gene set projections were utilized to estimate the p-value of an observed ring score, and thus the significance of a circular feature (Methods 4.3).

Ring score We provide a summary statistics, the *ring score* that measures the degree to which a given point cloud exhibits a circular structure. It provides a robust way for sub-sampling as a method for making circular enrichment computationally feasible across a large collection of gene sets.

2.2 Harmonic cyclicity analysis for estimating gene cascades

The identification of salient circular features is *topological* in nature, relying on persistent cohomology as the main representation of the data. In the second phase of the pipeline, we introduce *geometric* tools (integration, harmonics, density) to estimate the *phase* of cells in the closed process as well as the ordering of the associated gene expression cascade.

Circular coordinates and phase estimation Once a salient persistent cohomology class is identified, we use the cohomology-based circular coordinates algorithm [dSMVJ11] assign each cell a *phase* by mapping it into the circle (Figure 2B). A main challenge is that the algorithm is ill-suited to the highly non-uniform densities (Figure 13) common to single-cell data. This is overcome by the combination of two additional methods (Methods 4.4): (1) using the density-weighted harmonic from [PP23] and (2) using the sparse circular coordinate methods from [Per20].

Integral based lead-lag matrices The circular coordinates correspond to a harmonic 1-form on the space at given resolution, which we take as a model for canonical pseudo-dynamics rotating around a circular feature (Figure 2B). This permits the calculation of an integration-based lead-lag matrix (Figure 2C, Methods 4.4) beyond the original context in [BS16] of lead-lag for paths where the time parameter is known. The ij-th entry of the lead-lag matrix indicates whether gene i is active before (positive entry) or after (negative entry) gene j relative to the circular coordinate.

Estimation of gene cascades The calculation of the lead-lag matrix allows the standard linear-algebraic approach of [BS16] to be applied to estimate gene cascades.

Lead-lag projection of the data onto the principal complex lead-lag eigen-plane represents the two-dimensional linear projection that maximizes the total rotational energy of the pseudo-dynamics as represented by the harmonic. The amplitude and phase of complex components in the principal eigenvector estimate the contribution and peak phase of a gene relative to the harmonic (Figure 2D, Methods 4.4).

2.3 Synthetic experiments and benchmarking

2.3.1 ODE model of gene expression

We set out to test whether our model was capable of estimating gene-to-gene interactions in synthetic datasets. To do so, we introduced a modified version of the linear ODE model of gene transcription and translation in [CHC98].

Classical model The original linear model in [CHC98] is governed by two ODEs:

$$\frac{d\mathbf{r}}{dt} = \mathbf{A}\mathbf{p} - \mathbf{B}\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = \mathbf{C}\mathbf{r} - \mathbf{D}\mathbf{p} \tag{1}$$

where

1. $\mathbf{r}$ and $\mathbf{p}$ are the vectors of mRNA and protein levels of the n genes at a given time.
2. The $n \times n$ interaction matrix $\mathbf{A} = [a_{i,j}]_{i,j}$ describes the linear effect of protein j on the RNA production of gene i.
3. The matrices $\mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ are diagonal matrices corresponding to degradation rates of RNA, and translation and degradation rates of proteins respectively.

Transcriptional switch model RNA transcription is known to occur in bursts [CFX06, PE00]. More recent models of RNA transcription account for this by designating discrete transcriptional states with constant transcription rates [GPZC05, SMG+11, LMSZ+18, BLP+20]. To adapt the linear ODE model in (1) to this phenomemon, as well as to constrain the RNA counts to finite quantities, we introduced a threshold function

$$\sigma(-;\alpha,\tau) : \mathbb{R}^n \to \mathbb{R}^n \qquad x_i \mapsto \begin{cases} \alpha_i & \text{if } x_i \geq \tau \\ 0 & \text{else} \end{cases}.$$

The parameters α_i correspond to transcriptional rates for gene i given that the interaction is above certain threshold $\tau \in \mathbb{R}$. Our updated transcriptional switch model is then given by

$$\frac{d\mathbf{r}}{dt} = \sigma(\mathbf{A}\mathbf{p};\alpha,\tau) - \mathbf{B}\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = \mathbf{C}\mathbf{r} - \mathbf{D}\mathbf{p}. \tag{2}$$

Given initial conditions $\mathbf{r}_0$ and $\mathbf{p}_0$, solutions for the above system of ODEs were approximated using the `scipy.integrate` package over a finite sequence of discrete times

$$t_0 < t_1 < \ldots < t_m.$$

Example solutions to this model are given in Figure 3D and Figure 9D.

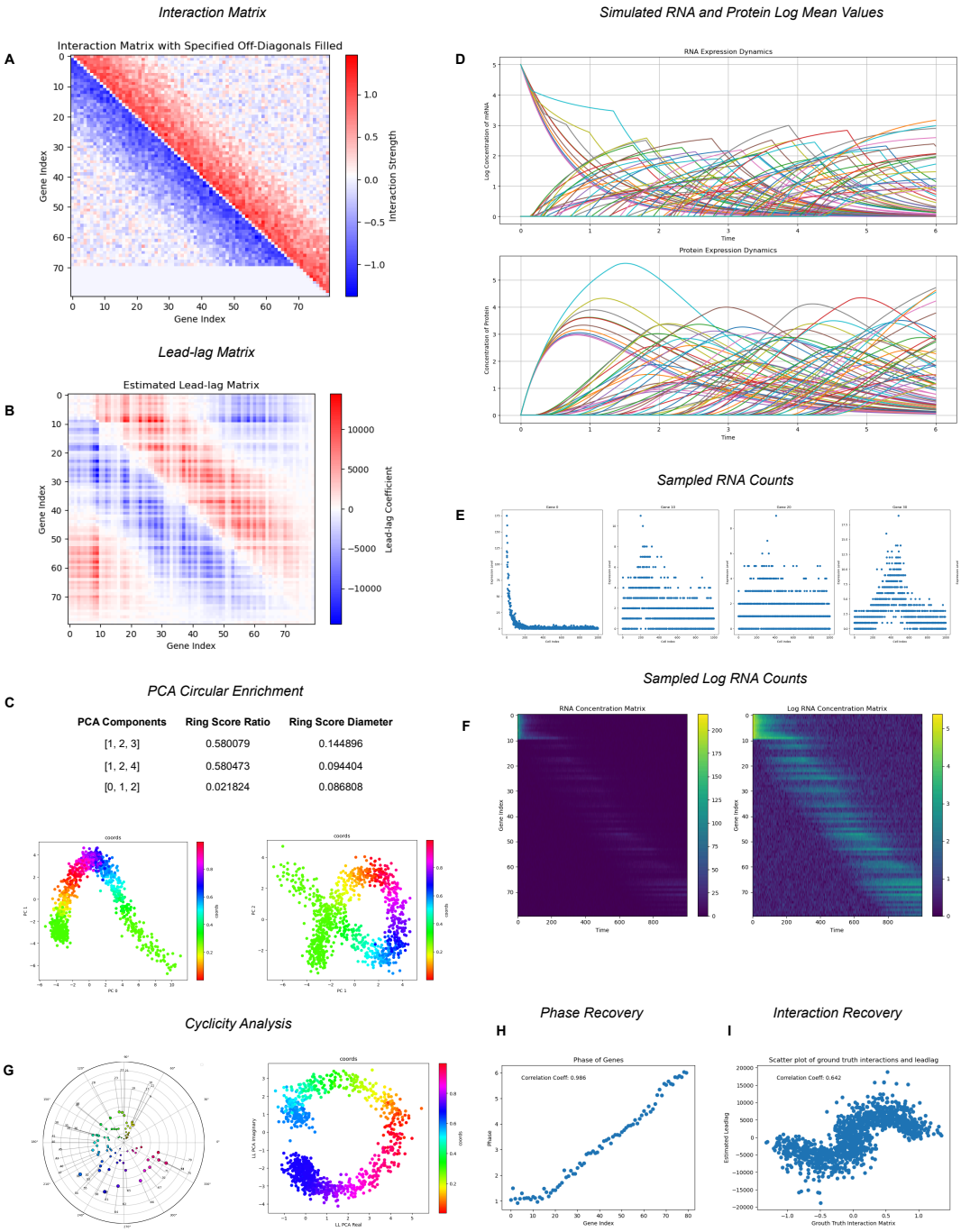

Figure 3: Analysis of a synthetic, transient mRNA transcription event. **(A)** The synthetic interaction matrix **A** of a transient mRNA transcription event. **(B)** The estimated lead-lag matrix. **(C)** PCA circular enrichment and selected projections onto principal components. **(D)** The simulated log means of RNA and protein. **(E,F)** The negative binomial sampled RNA counts and their log transforms. **(G)** Cyclicity analysis based phase estimation and lead-lag projection. **(H)** The estimated gene phases from cyclicity analysis and their correlation with the ground-truth gene ordering. **(I)** The values of the synthetic interaction matrix plotted against their values in the lead-lag matrix.

Sampling with negative binomial noise To realistically model noisy and stochastic single-cell data, we adopted the approach of [LLT+24] of negative binomial sampling, which has been observed to be a good mdoel for the distribution of RNA molecules [CFX06, PE00]. Our approximate solutions of mRNA levels according to Figure 8 were used as the mean in a negative binomial sampling with a constant dispersion parameter to generate a random count matrix that converges to our solved system in log mean (Appendix A.1). Examples of the sampling are given in Figure 3E-G and Figure 9E-G.

Negative feedback loops We first wanted to connect theoretical ODE models of RNA transcription to the emergence of global topological circular structures. Based on the chain-of-offsets model in Section 4.4, we hypothesized that when the protein-RNA interaction matrix $\mathbf{A}$ in the transcriptomic switch model 2 is a skew-symmetric rank 2

$$\mathbf{A} = \mathbf{u} \otimes \mathbf{v} - \mathbf{v} \otimes \mathbf{u} \qquad \mathbf{u}_k = \cos(2\pi k/g), \qquad \mathbf{v}_k = \sin(2\pi k/g)_k$$

there would be an evident circular feature in low-dimensional PCA. Our rationale was that such processes model a dominant negative feedback loop in the data whose dynamics is described by the principal complex eigenvector of $\mathbf{A}$ (See Section 4.4). Despite the modified transcriptional switch model and realistic noise, the lead-lag matrix (Figure 9B) provided a relatively robust model of the relationships between different genes (Figure 9C).

Transient processes We also wanted to demonstrate the viability of using circular enrichment in studying transient process that are not necessarily periodic. To recreate such a process in our model, we designed an interaction matrix where the earlier genes positively influenced later genes and vice versa (Figure 3). Circular gene set enrichment revealed significant circular structure outside of the first principal component (Figure 3G,H). The reconstructed phase estimates reliably recovered the ordering of genes (Figure 3I,J), and the lead-lag matrix provided a reasonable estimation of the underlying gene-protein interaction matrix (Figure 3C).

Velocycle synthetic model We also benchmarked the estimation of circular coordinates against the synthetic model presented in [LLT+24]. In this model, mean unspliced RNA expression is modeled directly as first harmonics in the Fourier series:

$$\mathbf{r}_u(t) = \mathbf{a}\cos(2\pi t + \mathbf{b}) + \mathbf{c}$$

where $\mathbf{a} \in \mathbb{R}_+^g, \mathbf{b} \in [0, 2\pi)^g$ and $\mathbf{c} \in \mathbb{R}^g$ are the respective vectors of amplitudes, phases and offsets of each gene. As for the previous examples, the vector $\mathbf{r}(t)$ is used as the mean in the negative binomial distribution for sampling the data as per Equation (22).

Benchmarking circular coordinate estimation In addition to Velocycle, we also used the angle in the projection onto the first two principal components as a baseline for benchmarking. Sensitivity analysis over the same parameter set as in [LLT+24] showed that cHunter, velocycle and the PCA baseline were roughly comparable in estimating the ground truth circular coordinate, where accuracy was correlated with a more pronounced ring score (Figure 10A). We extracted additional information in a

case study with 3000 cells and 300 genes. There was a clear circular structure in the first two PCs (Figure 10B), which ultimately meant that all three methods roughly attained the same circular coordinate. Additionally, using the first two principal components served as a good estimate of gene phases (Figure 10E).

Remark 1. *The high performance of the PCA baseline in this benchmark is noteworthy. The mathematical theory in the methods section offers some explanation. When all genes are roughly described by the first harmonic, the data is close to a chain-of-offsets model (12). In this case, the data lies on an ellipse lying in the plane described by the rank 2 decomposition of the lead-lag matrix (16). This plane will also be recovered by PCA, and if the data is mean-centered, the angle in this plane will coordinatize the ellipse, explaining the high performance.*

2.4 Validating harmonic lead-lag analysis on cell cycle data

Firstly, we wanted to validate that our harmonic lead-lag method produced sensible results on the most commonly studied closed biological process: the cell cycle. To do so, we used a dataset of cycling human dermal fibroblast [CKL+22] (Appendix A.2), which has been used previously as a validation dataset on cell-cycle tools [LLT+24].

Identifying and indexing the cell cycle We projected the data onto the gene set spanned by known S and G2M markers (we use the markers in [HSK+23]). After identifying a persistent cohomology class, the lead-lag projection (Figure 4A) exhibits coherence between the estimated phase and the transitions from S to G2M phases (orientation and start point were assigned to fit these plots). Further, the log of the total number of molecules was consistent with passage through the S phase into mitosis. The high concentration of G0/G1 cells at an initial phase is a result of low transcriptomic activity relative to cell cycle gene sets containing solely S and G2M genes.

Harmonic lead-lag analysis estimates gene cascade The calculated lead-lag matrix was consistent with the high-level dynamics of S genes leading G2M genes (Figure 4D). This consistency was also present in the gene phase plots (Figure 4C) and when ordering gene expression against cell phase (Figure 4E,F). There was an observable gap of genes estimated to peak at angle 0, which again corresponds to the omission of G0/G1 genes from the cell cycle marker gene set in [HSK+23].

Incorporating splicing data The human fibroblasts dataset contains the additional information of spliced and unspliced molecules. We reran our method on the same cell cycle gene set, separating out spliced and unspliced molecules. As in the previous case, we identified a salient persistent cohomology class, where the lead-lag projection was consistent with Seurat cell-cycle scoring (Figure 11A). The lead-lag matrix showed that unspliced RNA levels led spliced RNA levels of corresponding genes as expected (Figure 11A). Further, the estimated phases of unspliced to spliced increased broadly across the gene set (Figure 11B,D).

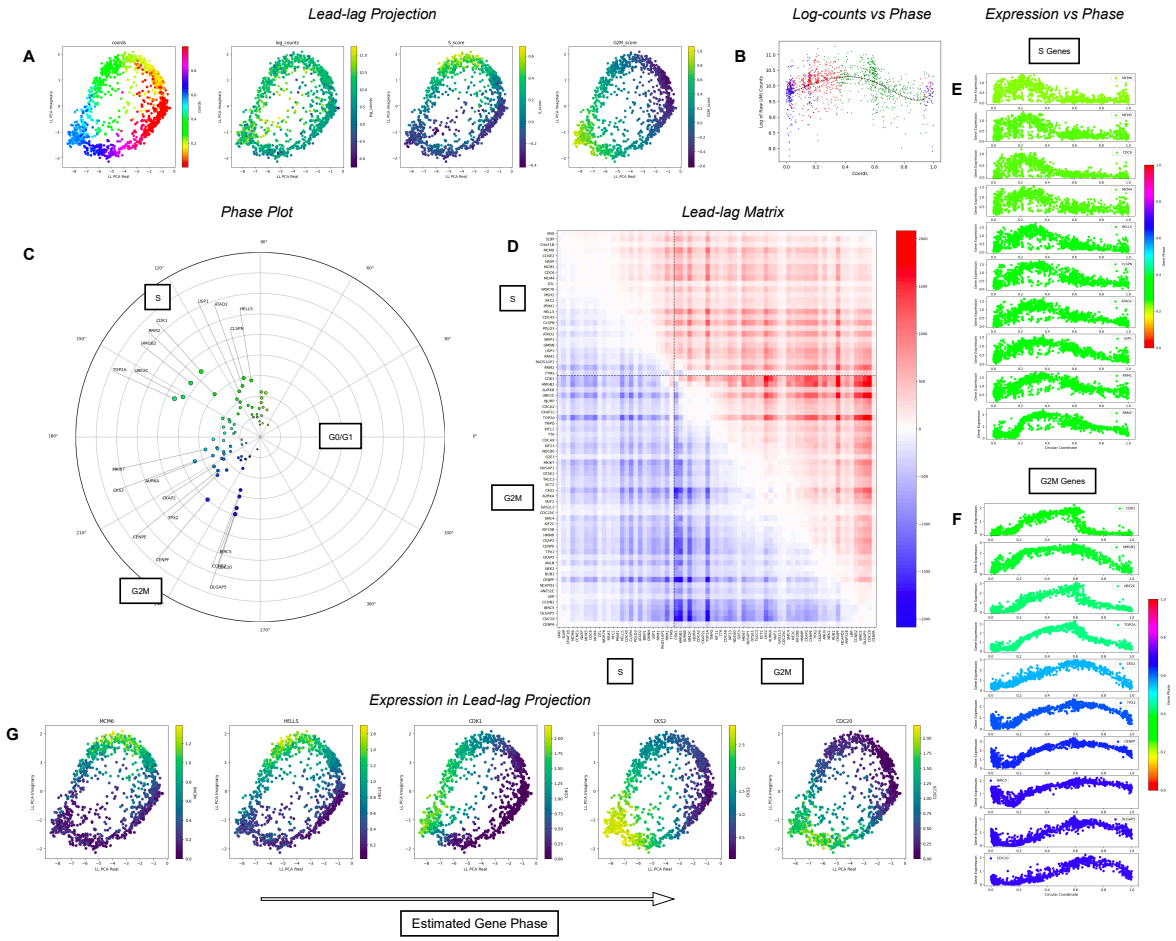

Figure 4: Analysis of cycling human dermal fibroblasts. **(A)** Lead-lag projection and estimated circular coordinate phases with standard Seurat S and $G2M$ cell phase scoring. **(B)** The log of the raw UMI counts per cell against the circular coordinates. **(C)** Estimated gene phases from harmonic lead-lag analysis. **(D)** The lead-lag matrix, sorted by estimated gene phase and groups by S and $G2M$ genes. **(E,F)** Expression of top 10 S and $G2M$ genes respectively against estimated phases. **(G)** Top 5 genes (sorted by lead-lag amplitude) sorted by phase, plotted over lead-lag projection.

2.5 AR silencing, senescence and deviation from the cell cycle

In the previous experiment, we validated our tool's ability to capture the cell cycle and its phases. In this section, we aimed to explore its potential application in detecting *deviations* from the cell cycle. Specifically, we investigated senescence, a state characterized by permanent cell cycle arrest. Our objective was to determine whether senescence was observable geometrically as a deviation from a circular structure in mRNA expression space.

Dataset We leveraged our previously established knowledge that the androgen receptor (AR) plays a crucial role in regulating dermal fibroblast homeostasis and early steps of cancer associated fibroblast (CAF) activation. AR acts specifically as a key negative regulator of cellular senescence–effector genes, like CDKN1A, and of genes of the Senescence Associated Secretory Phenotype (SASP), coding for a battery of proinflammatory cytokines and matrix remodeling proteins that are also expressed in fully established CAFs [CGP+18]. Specifically, we analyzed a single-cell RNA-seq dataset of human dermal fibroblasts (HDFs) infected with two lentiviruses expressing different AR-silencing shRNAs versus a control expressing scrambled shRNAs. Following puromycin selection, cells were collected and processed for single cell RNA-seq using the 10x Chromium 3' kit according to the manufacturer's instructions. Sequencing was subsequently performed at the Lausanne Genomic Technologies Facility.

Preprocesing Pre-processing and filtering were performed by grouping the shAR conditions with the control. This dataset consisted of 5,607, 3,663, and 4,410 cells in the shAR1, shAR2, and shCTRL subgroups, respectively. The rationale for grouping the conditions together was to (1) identify shared circular structures before (2) locating perturbations or deviations specific to each condition. In the combined dataset, cells expressing at least 200 genes and genes expressed in at least three cells were retained. Raw UMI counts were normalized per cell and log-transformed. MAGIC [vDSN+18] was applied to the data with a nearest-neighbor parameter of knn = 4 for one step ($t = 1$). Highly variable genes were then identified using `scanpy.tl.highly-variable`. Finally, max-min sampling was applied in the first 50 principal components to downsample to 2^12 landmark cells.

Cell cycle gene phase estimation exhibits robustness across datasets We first aimed to validate the findings from the previous case study on fibroblasts and the cell cycle across different datasets. To achieve this, we restricted gene expression to the default small cell cycle (SCC) gene set in Velocycle [LLT+24] and replicated our approach on the Capolupo dataset [CKL+22]. All 98 genes in this set were identified as highly variable. Although the circular structure was no longer visible in the first three principal components (Figure 5A), a circular enrichment was observed when analyzing combinations of three out of the first five principal components. Specifically, a circular structure was found in components 0, 1, and 3 (Figure 5B,C). Lead-lag analysis revealed a striking correlation in estimated gene phases ($r \approx 0.92$) and amplitudes ($r \approx 0.87$) across this dataset and that of Capolupo [CKL+22] (77 shared, highly variable genes). Notably, significantly different pre-processing steps were employed between the two

experiments (e.g., union of conditions, k-nn smoothing), underscoring the robustness of our method.

Prominence of G1 and G1-S transition phase The absence of a visible cell cycle in the first three principal components of the gene set used for the previous dataset suggested the presence of another source of variability. Upon investigation, we found that the Reactome gene set *Mitotic G1 and G1/S Transition* (MG1/S) exhibited a clear circular structure in the first three principal components (Figure 5E,F and Figure 12D-H). After combining and pre-processing the data, both conditions displayed a similar distribution along this circular feature (Figure 5E). This revealed a higher proportion of cells in the G1 phase across both conditions compared to the Capolupo experiment, accounting for some of the variability observed between the datasets.

Comparison with sCC coordinate We also compared (Figure 12H,I) the MG1/S circular coordinate with the one estimated using the sCC gene set (Figure 12A). Interestingly, we found the two circular coordinates followed one another consecutively rather than covarying (Figure 12H). Given that the sCC gene set is primarily comprised of markers for S and G2M phases (Figure 4), this observation is consistent with the idea that MG1/S and S/G2M gene sets represent consecutive components of a larger cell cycle regulatory network.

Circular structure in SASP genes coupled with MG1/S Given the connection between AR silencing and senescence in fibroblasts [CGP+18], we tested a further Reactome gene set *Senescence-associated Secretory Phenotype (SASP)* for circular structure. This gene set ($n = 112$) intersected with 22 highly variable genes in the combined expression data, 8 of which were also in the MG1/S gene set (CDKN2C, CCNA2, CDKN1A, CDK6, CDKN2B, CDKN1B, CDK2, CDKN2D). Circular enrichment suggested the SASP circular structure was more significant than that of any hallmark set (Figure 12B,C). Further, the SASP and MG1/S circular coordinates were clearly coupled with one another (Figure 5F). We concluded that both circular structures represented sub-processes of the cell cycle primarily related to the G1/S transition.

AR silencing and senescence We then aimed to understand the factors separating the shAR and control conditions. To this end, we re-processed the count matrices without MAGIC k-nn smoothing to prevent bleeding between the conditions. Using bulk RNAsequencing analysis of HDFs with AR silencing, we devised an AR signature score based on the top 200 up-regulated genes under AR silencing. We found that this score was signifcantly more prominent in the shAR samples relative to the control (Figure 5I,J). Further, AR silencing was concentrated at a particular angle on the SASP circular coordinate (Figure 5I,J). This was accompanied by peaks in senescence regulator CDKN1A and SASP gene IL6 (Figure 12L), as determined by the lead-lag phase estimation (Figure 12Q). We confirmed the connection with senescence by comparing a more traditional Leiden cluster analysis with our lead-lag projection (Figure 12O,P). Two Leiden clusters (2 and 6) were concentrated at this angle (Figure 12P); particularly, cluster 2 primarily contained shAR cells which scored highly for SASP and AR silencing.

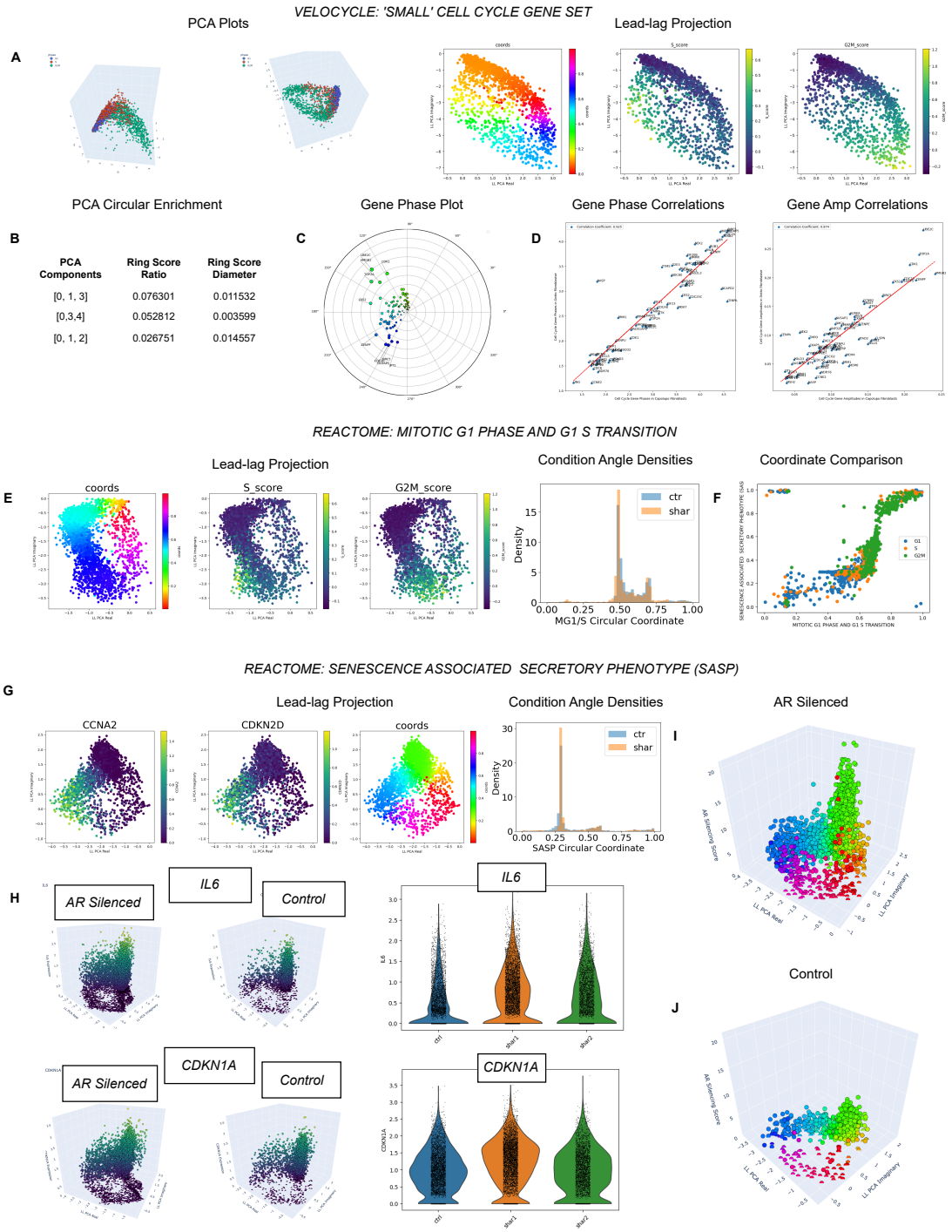

Figure 5: Analysis of AR silencing data. **(A)** Left: PCA projection in dimensions one, two and three and one, two and four respectively of the data restricted to the velocycle 'small' cell cycle gene set. Right: Lead-lag projection of the circular structure identified in PCA dimensions one, two and four. **(B)** The ring scores of different PCA combinations. **(C)** Phase plot corresponding to the lead-lag projection. **(D)** The correlation between gene phases and amplitudes in the Dotto data with those estimated in the Capolupo data. **(E)** Lead-lag projection of MG1/S gene set. **(F)** Circular coordinate comparison between MG1/S gene-set and SASP gene set. **(G)** Lead-lag projection of *SASP* gene set. **(H)** Left: Expression of IL6 and CDKN1A plotted against SASP lead-lag projection, separated by shAR and control conditions. Right: the associated violin plots. **(I,J)** AR silencing score against SASP lead-lag projection, separated by shAR and control conditions.

Discussion Our analysis suggests that the effects of AR silencing are (1) most pronounced at a specific point in the G1-S transition and (2) orthogonal to the corresponding circular structure. We hypothesize that cells exhibiting a high AR silencing score exit the cell cycle at this point Figure 5I — effectively entering a different valley in the epigenetic landscape — through an epigenetic/transcriptional process related to senescence and AR signaling. This process is observable as a geometric feature Figure 5I; specifically, as a continuous path of cells that deviate from the circular structure as AR silencing score increases. We plan to examine spliced and unspliced abundances to ultimately determine the directedness and orientation of this path in future work. If the directionality of the deviation is confirmed, the connection with AR silencing suggests that AR agonists or BET inhibitors may potentially drive cells back to a normal cyclic program.

2.6 cHunter recovers circular dynamics in androgen-induced prostate regeneration

To complement the previous sections, we sought to assess the capability of our approach in (1) the in vivo setting, (2) mouse models and (3) multiple scRNA-seq samples. To this end, we examined a publicly available data-set [KHC+20] where the authors study the role of androgen in prostatic epithelial cell regeneration in mice (Details A.3). We hypothesized that the regenerative cycle of tissue generation would manifest geometrically in certain gene sets.

Dataset Mice are castrated and within 28 days the prostate involutes to 90% of its size. At day 28 mice are transplanted subcuteneously with dihydrotestosterone (DHT) to initiate androgen mediated regeneration. After 28 days of regeneration the prostate returns to homeostasis. Multiple mice received the treatment concurrently, with single-cell samples of their respective prostrates taken over the full course of castration and regeneration. Our hypothesis was that the deviation from and return to homeostasis would create a loop structure identifiable via our persistent cohomology algorithm.

Circular enrichment identifies biologically meaningful processes Our first question was whether circular gene set enrichment was able to identify relevant biological processes. We first subset the gene set by density, before performing maxmin sampling (Methods 4.1). We then performed circular gene set enrichment using two variants (diameter and ratio) of our ring score, testing over the 50 hallmark sets of genes. [STM+05] We found that androgen response was the top circularly enriched subspace (Figure 6K) of statistical significance against both ring score variants, validating our hypothesis that gene expression associated with androgen response would generate a closed biological loop in the single-cell phase subspace.

Lead-lag analysis suggests ordering on androgen-response genes We then computed the circular coordinate and performed the associated lead-lag analysis (Figure 6C-F). We noted that several androgen-regulated genes in the matrix have key roles in prostate development and maintenance. Notably NKX3.1, a transcription known to

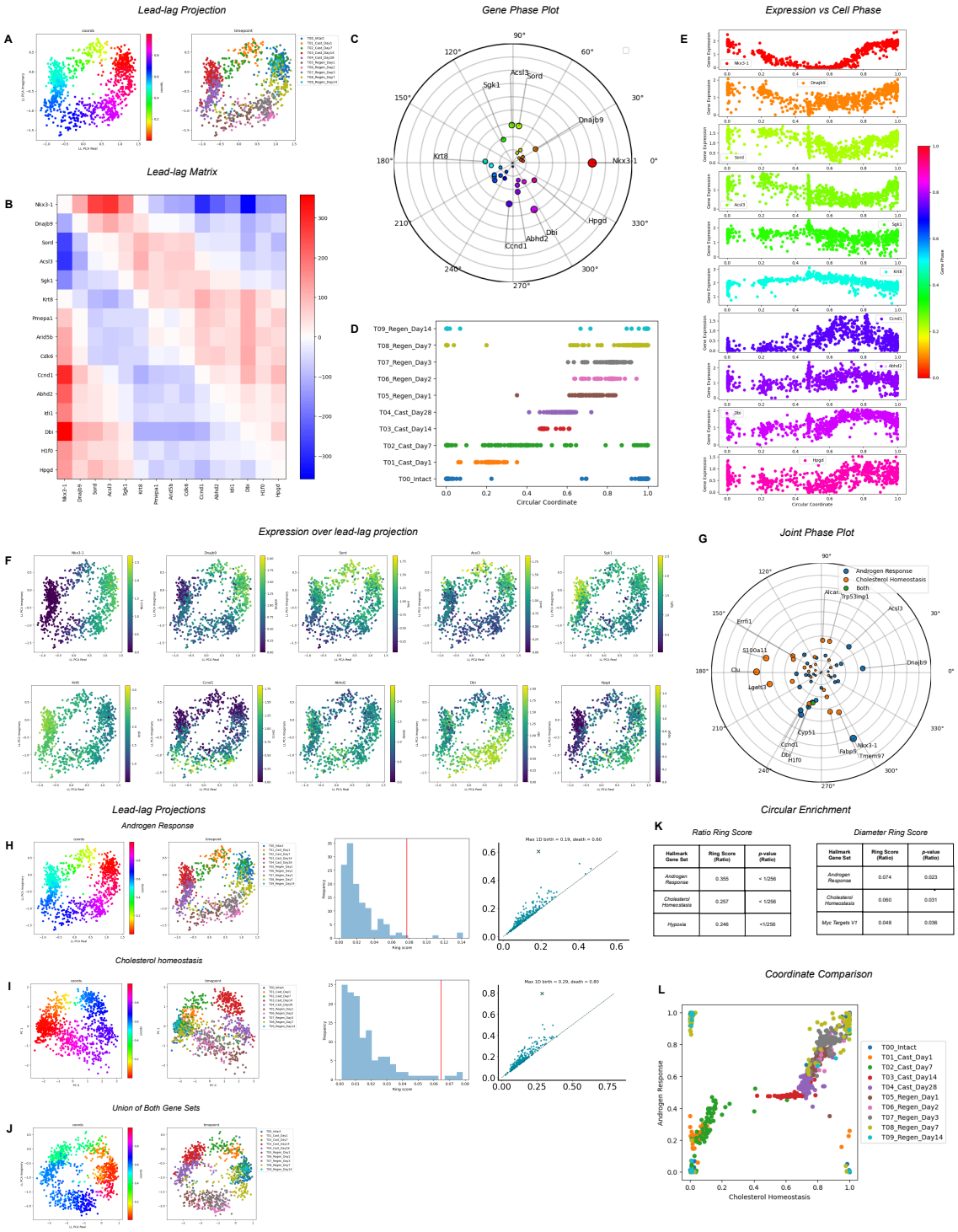

Figure 6: Analysis of prostate regeneration data. **(A)** The lead-lag projection colored by circular coordinates and sample time, respectively. **(B)** Lead-lag matrix. **(C)** The estimated gene phases. **(D)** Histogram plots of sample times distributed against the circular coordinates.**(E)** Gene expression of primary genes vs circular coordinate. **(F)** Gene expression of primary genes vs lead-lag projection. **(G)** Phase plot of genes against the circular coordinate attained in the union of the two gene sets. **(H-J)** The lead-lag projection against two enriched gene sets and their union, and corresponding ring score permutation tests and persistence diagrams. **(K)** Circular gene set enrichment results. **(L)** Circular coordinates of Cholesterol Homeostasis vs Androgen Response gene sets.

be essential for prostate development and maintenance of luminal cell [BGDS+99] identity, showed strong enrichment/correlation to the homeostasis timepoints. We also noted the high amplitude of the androgen-regulated SGK1, which plays a role in cellular stress response. Lastly, we observed a strong effect of androgen-regulated cell-cycle genes in the early phase of regeneration, in line with the high proliferation observed. Of note these androgen-response genes have been generated in prostate cancer cell lines that are dependent on androgens for growth, which may suggest these cell lines have hijacked as part of the regenerative response to androgens.

Phase estimates recover ground truth time labels The experiment provided a ground-truth labeling of samples ordered by time in the underlying biological process. These labels were respectively ordered by: intact; 1 day, 7 days, 14 days after castration; 1 day, 2 days, 7 days, 14 days and 28 days after regeneration. We observed that the estimated circular coordinate was coherent with the ground truth ordering of the samples (Figure 6D).

Circular enrichment identifies multiple interlocking sub-processes In addition to the androgen-response gene set, circular gene set enrichment identified the cholesterol homeostasis gene set as statistically significant with respect to the two ring score variants (Figure 6H-K). Plotting the circular coordinates against each other (Figure 6L) showed that both circular coordinates were dependent on each other, which suggests that the biological processes associated with cholesterol homeostasis and androgen responses are sub-processes of the master tissue regeneration cycle. Potentially this is an intrinsic adaptation to the absence of androgens, as cholesterol is a precursor to androgens. Additionally, the union of the gene sets also exhibited a dominant circular structure, from which the lead-lag analysis allowed the ordering of the genes between the sets to be compared. This demonstrated that the two processes are dominant at different points of the process, as per the phase plot in Figure 6G.

3 Discussion

In this work, we have presented a novel topological approach to the discovery of closed biological processes. Our method provides measures of statistical uncertainty, and further, generates quantitative estimates of specific gene's activity and their order in the cascade of gene expression associated with the cyclic process. In contrast to other methods, it does not involve any machine learning techniques or parameter fitting, focusing instead on generating meaningful, interpretable, and reproducible linear projections based on global geometric and topological arguments.

3.1 Biological insights

Our new method revealed biological insights that warrant further investigation.

Robust cell-cycle gene cascades in dermal fibroblasts. By focusing on a well-characterized cell culture system of primary human dermal fibroblasts, we first validated the solidity of our method to detect a circular distribution of cells and gene peaks along

the phases of the cell cycle. We then expanded our analysis to human dermal fibroblasts in a different experimental setting based on silencing of the AR gene, which triggers early steps of CAF activation. In agreement with the analysis of the previous data set, we found a remarkably similar structure of gene phases and peaks along the cell cycle, highlighting the robustness of our approach.

AR silencing and cell cycle exit. Juxtaposed to this, our analysis suggests that silencing of the AR gene caused an exit from the circular distribution that coincided with increased expression of the key senescence effector gene CDKN1 and of the SASP component IL6. Overall, the findings are consistent with a Waddington model of epigenetic programming of the transition from normal stromal to CAFs, which may be amenable to reversal by epigenetic reprogramming. We speculate that this may be achieved by already proven experimental approaches such as treatment of cells with inhibitors of BET proteins or with AR agonists.

Cholesterol homeostasis in prostate regeneration. We further extended the validity of our method to analyze the topological distribution of cells in an in vivo model of prostate gland development, a process under AR control. We examined a large, comprehensive single-cell dataset from prostate tissue of mice subjected to androgen ablation (via castration) and subsequent androgen administration, which results in normal prostate reconstitution. Our analysis clustered single-cell data based on an AR gene signature, revealing a circular arrangement of cells that strongly correlated with the score of a "canonical" androgen response signature. Further analysis using 50 gene hallmark signatures uncovered an overlapping distributions of cells with modulation of the androgen response and of cholesterol homeostasis. This previously unrecognized connection has significant biological implications, including the involvement of enzymes shared between cholesterol and androgen synthesis, which could be targeted in disease contexts.

3.2 Key methodological points

Applicability across multiple experimental settings. We show the applicability of our method through diverse settings and generate coherent biological hypotheses in each. Firstly, we connect our model with established synthetic models of RNA production and benchmark against a state-of-the-art method for extracting circular coordinates (Section 2.3). In a dataset of cycling human dermal fibroblast cells, we were able to identify the cell cycle and recover sensible estimates on the dynamic ordering of genes (Section 2.4). The same dynamic ordering was observed in the next section (Section 2.5), which examined human dermal fibroblasts in a different experimental condition. Finally, in Section 2.6 we identified and exploited circular structure in a mouse model dataset, which required analysis over multiple heterogeneous samples. We view these results as preliminary validation of topological methods as a flexible tool for identifying circular structure ab initio in diverse experimental single-cell RNA sequencing settings.

Statistical geometric analysis relative to gene sets. Projection onto a gene set corresponds to a choice of geometry. Downstream geometric quanties — harmonic

circular coordinates, leadlag coefficients and estimated gene phases and amplitudes — should *always* be considered relative to such a choice. Further, there is no reason to expect there exists a canonical gene set for studying a given biological process. For example, the circular geometries of cholesterol homeostasis and androgen response are sub-processes of prostate tissue regeneration exhibiting a complicated non-linear relationship (Figure 6L). Our approach of testing circular geometries across multiple gene sets is a first step toward systematically addressing this issue directly via statistical hypothesis testing over geometric quantities.

3.3 Limitations

Hyper-parameter selection. One of the strengths of our model is that it does require any parameter fitting or machine learning techniques, which can be difficult to reproduce or interpret. However, it does hinge on hyper-parameter choices in the pre-processing stage. These include normalizing the global counts or counts within genes, restricting or not to highly variable genes, the choice of gene sets to enrich over, filtering and sub-sampling choices, etc. Unfortunately, there did not seem to be reliable set of such hyper-parameters that worked across every setting, meaning that some experimentation is required on the part of the user.

Outlier robustness. One of the problems that plagues persistent homology is that a single outlier can dramatically impact the persistence diagram. In some circumstances, this necessitates using a k-nn diffusion-based filter such as MAGIC to control outliers in the quest to identify a biologically meaningful circular structure, which may affect the global geometry and downstream estimates. We are currently examining the mathematical effect of such averaging on lead-lag analysis in concurrent work.

3.4 Future directions

AR-dependent vs AR-response signatures AR exhibits a duality of functions: its constitutive role in regulating gene transcription (and potentially non-transcriptional processes) and its ligand-dependent activation. This duality is evident in the limited overlap between gene sets influenced by AR loss and those modulated by ligand-dependent AR activity. Additionally, AR effects are highly context- and cell type-specific, adding complexity to its role. A future application of this method will be to compare the topological distribution of cells in different contexts as a function of different AR-dependent versus AR-responsive gene signatures.

RNA velocity integration In Section 2.4 we gave a basic demonstration that lead-lag analysis was able to generate consistent ordering of spliced and unspliced peaks around the cell cycle. There exist many algorithms for estimating RNA velocity dynamics – in particular around the cell cycle – which are rooted in ODE models of RNA spliced and unspliced transcription dynamics [BLP+20, LMSZ+18, ROD+22, LLT+24]. In contrast, the chain-of-offsets model we use here is time-reparametrization invariant and does not involve ODEs — rather, it focuses on the geometry of the phase space to estimate gene orderings. A future direction is the integration of our pipeline with such models, where we believe that the ab initio use of topology as a preprocessing step may

identify new settings where advanced RNA velocity models may be applied in addition to complement our model. Notably, our algorithm still identifies meaningful dynamics in cases (such as tissue regeneration) that are potentially outside of the time-scale of RNA velocity, given that the half-life of spliced mRNA is roughly 30 minutes on average [RLF+11].

A general theory of identifying concurrent processes Our example in Section 2.6 identified cholesterol homeostasis and androgen response as multiple interlocking sub-processes in tissue regeneration. We envision that this is a first step in a general algorithm for separating complex biological phenomena into simpler sub-processes – or 'process times' [FGP24] – and studying their interaction based on the geometry of gene expression space. As part of this, a broader theory of geometric gene set enrichment could be useful to study geometries other than circular ones — for instances, paths or more general 1-dimensional manifolds — as a basis for identifying 'basic' processes ab initio through geometric testing.

4 Method

4.1 Single Cell RNA-seq

Basic Background. Single-cell sequencing is a technology that measures the number of RNA molecules in each cell in a given tissue. For our purposes, the mathematical output of a single-cell experiment consists of a matrix `expression` whose rows correspond to the cells and columns to genes. Equivalently, the matrix corresponds to an integer point cloud whose points are cells and dimensions are genes.

Pre-processing. Single-cell data is well-known to be noisy. There are several common pre-processing steps to filtering and processing the initial expression matrix that we apply in our experiments. Following the Seurat [HSK+23] recommendations, our pipeline is as follows:

1. Filter genes and cells by a minimal number of molecules.
2. Library normalization and variance stabilization. In our case, we used an ℓ_1-normalization scaled by 10000, followed by a log-transformation, as suggested in [BHS+18]
$$\texttt{expression[i,g]} = \log\left(1 + \frac{10000 \cdot \texttt{count[i,g]}}{\sum_g \texttt{count[i,g]}}\right)$$
3. We optionally apply a diffusion scheme using the MAGIC[1] algorithm [vDSN+18].
4. Taking subsets of the remaining gene set determined by highly variable genes.

4.2 Mathematical background

4.2.1 Persistent cohomology

Simplicial complexes A (finite) *simplicial complex* is a (finite) set X and collection $\mathcal{S}$ of subsets of X satisfying

$$T \subset S \quad \text{and} \quad S \in \mathcal{S} \qquad \Rightarrow \qquad T \in \mathcal{S}.$$

The sets $S \in \mathcal{S}$ are *simplices* and sets $S \in \mathcal{S}^k$ of cardinality $k+1$ are *k-simplices*. If X has n elements, the *geometric realization* of $(X, \mathcal{S})$ is topological subspace

$$|\,\mathcal{S}\,| = \bigcup_{\{x_0,\ldots,x_k\}\in\mathcal{S}} \mathrm{Hull}\{e_{x_0}, \ldots, e_{x_k}\} \subset \mathbb{R}^n,$$

where to each element x of X, we associate a different choice e_x of standard basis vector of $\mathbb{R}^n$, and Hull denotes the convex hull. A simplicial complex is *oriented* if there is a total order on X, in which case simplices are written as ordered tuples

$$[x_0, \ldots, x_k] \in \mathcal{S}^k.$$

[1] Magic is a low-pass filter which replaces the gene expression of each cell with a weighted average of its nearest neighbours.

Cochains. Let $(X, \mathcal{S})$ be an oriented simplicial complex and R a commutative ring. The *k-cochains* of $(X, \mathcal{S})$ is the R-module

$$C^k(\mathcal{S}; R) = \{\ f : \mathcal{S}^k \to R\ \}$$

of R-valued functions on the set $\mathcal{S}^k$ of k-simplices. The *k-th coboundary operator* is R-linear map

$$d_k : C^k(\mathcal{S}; R) \to C^{k+1}(\mathcal{S}; R)$$

defined by

$$d_k f\big([x_0, \ldots, x_{k+1}]\big) = \sum_i (-1)^i f([x_0, \ldots, \hat{x}_i, \ldots, x_{k+1}])$$

where $\hat{x}_i$ is omitted from $[x_0, \ldots, x_{k+1}]$. The *simplicial cochain complex* $C^\bullet(\mathcal{S}; R)$ of $(X, \mathcal{S})$ is the sequence of R-modules

$$C^0(\mathcal{S}; R) \xrightarrow{d_0} C^1(\mathcal{S}; R) \xrightarrow{d_1} \ldots$$

and satisfies the relation $d_{k+1} d_k = 0$ for all k. The *k-th cohomology* of the complex is given by the quotient

$$H^k(\mathcal{S}; R) = \mathrm{Ker} d_k / \mathrm{Im} d_{k-1}$$

of the *k-cocycles* $\mathrm{Ker} d_k$ modulo the *k-coboundaries* $\mathrm{Im} d_{k-1}$. A map $\varphi : (X, \mathcal{S}) \to (Y, \mathcal{T})$ is *simplicial* if $\varphi(S) \in \mathcal{T}$ for all $S \in \mathcal{S}$. If φ is simplicial, then it induces an R-module homorphism on cohomology

$$H^*(\varphi) : H^*(\mathcal{T}; R) \to H^*(\mathcal{S}; R) : [f] \mapsto f \circ \varphi]$$

for any ring R.

Persistent cohomology. Let (X, d) be a metric space, i.e., a set X equipped with a metric (distance) d. The *Vietoris-Rips complex* at distance $\varepsilon \in \mathbb{R}_+$ is the simplicial complex X_ε with k-simplices

$$X_\varepsilon^k = \big\{[x_0, \ldots, x_k] \mid d(x_i, x_j) \leq \varepsilon\big\}.$$

For each $\varepsilon < \delta$, there is an inclusion $X_\varepsilon \hookrightarrow X_\delta$, thus given $0 = \varepsilon_0 < \varepsilon_1 < \ldots < \varepsilon_n$, we obtain the *Rips filtration*

$$X_{\varepsilon_0} \hookrightarrow X_{\varepsilon_1} \hookrightarrow X_{\varepsilon_2} \hookrightarrow \ldots \hookrightarrow X_{\varepsilon_n}.$$

The *k-th persistent cohomology* $H^k(X_\bullet)$ of (X, d) over a field $\mathbb{F}$ is then the sequence

$$H^k(X_{\varepsilon_0}; \mathbb{F}) \leftarrow H^k(X_{\varepsilon_1}; \mathbb{F}) \leftarrow H^k(X_{\varepsilon_2}; \mathbb{F}) \leftarrow \ldots \leftarrow H^k(X_{\varepsilon_n}; \mathbb{F})$$

obtained from applying cohomology to the Rips filtration.

$$\begin{array}{ccccccccccccc} \cdots & \longleftarrow & 0 & \longleftarrow & \mathbb{F} & \xleftarrow[id]{} & \mathbb{F} & \xleftarrow[id]{} & \cdots & \xleftarrow[id]{} & \mathbb{F} & \longleftarrow & 0 & \longleftarrow & \cdots \\ & & & & \varepsilon_i & & & & & & \varepsilon_j & & & & \end{array}$$

The following theorem forms the basis of practical applications of persistent cohomology, and is attributed to different references depending on the conditions [Web85,

ZC04, ELZ00]. It is formulated in terms of *intervals* $\mathbb{I}_{[\varepsilon,\delta]}$, for *varepsilon* $\leq \delta$, which is the following sequence of $\mathbb{F}$-vector spaces and linear maps.

$$\cdots \longleftarrow 0 \longleftarrow \underset{\varepsilon_i}{\mathbb{F}} \xleftarrow[id]{} \mathbb{F} \xleftarrow[id]{} \cdots \xleftarrow[id]{} \underset{\varepsilon_j}{\mathbb{F}} \longleftarrow 0 \longleftarrow \cdots$$

Theorem 2 (Interval decomposition)**.** *For each k, the persistent cohomology $H^k(X_\bullet;\mathbb{F})$ of a finite metric space (X,d) is isomorphic to a direct sum of interval modules*

$$H^k(X_\bullet;\mathbb{F}) \cong \bigoplus_{\gamma\in\Gamma_k} \mathbb{I}_{[\varepsilon^\gamma,\delta^\gamma]} \tag{3}$$

which is unique up to permutation of summands.

The intervals $\mathbb{I}_{[\varepsilon^\gamma,\delta^\gamma]}$ in the decomposition are called *persistent cohomology classes*. The start ε^γ and end δ^γ points are the *birth* and *death* times of the persistent cohomology class. The *persistence diagram* of (X,d) is the multi-set

$$\mathcal{D}^k_{\mathbb{F}}(X,d) = \{(\varepsilon^\gamma,\delta^\gamma) \mid \gamma\in\Gamma_k\} \subset \mathbb{R}^2_+$$

using the interval decomposition in Equation 3.

Stability. The *bottleneck distance* between two persistence diagrams is

$$d_B\big(\mathcal{D}^k_{\mathbb{F}}(X,d_X),\mathcal{D}^k_{\mathbb{F}}(Y,d_Y)\big) = \inf_\gamma \sup_p \|p-\sigma(p)\|_\infty$$

where $p\in\mathcal{D}^k_{\mathbb{F}}(X,d_X)$, and σ ranges over all bijections $\sigma:\mathcal{D}^k_{\mathbb{F}}(X,d)\to\mathcal{D}^k_{\mathbb{F}}(Y,d_Y)$. Let

$$d_{GH}\big((X,d_X),(Y,d_Y)\big)$$

denote the Gromov-Hausdorff distance between two finite metric spaces (X,d_X) and (Y,d_Y) (see [CCG+09] for the definition). The following theorem states that the persistence diagram is stable with respect to small perturbations of the underlying metric space.

Theorem 3 ([CCG+09], Thm. 3.1)**.**

$$d_B\big(\mathcal{D}^k_{\mathbb{F}}(X,d_X),\mathcal{D}^k_{\mathbb{F}}(Y,d_Y)\big) \leq d_{GH}\big((X,d_X),(Y,d_Y)\big).$$

4.2.2 Harmonic circular coordinates

In this section we review the circular coordinates algorithm from [dSMVJ11], emphasizing the density-weighted [PP23] and sparse [Per20] variants.

Representability. For any pair X,Y of pointed topological spaces, let $[X,Y]_*$ denote the set of pointed homotopy classes of basepoint-preserving continuous maps from X to Y. The following classical theorem provides an important interpretation of cohomology over $\mathbb{Z}$. Let S^1 denote the usual circle of radius 1 in the plane, centered at the origin.

Theorem 4. *(cf. Thm. 4.57 [Hat02]) For any simplicial complex $\mathcal{S}$, there is a bijective correspondence*

$$H^1(\mathcal{S};\mathbb{Z}) \cong \big[|\mathcal{S}|,S^1\big]_*.$$

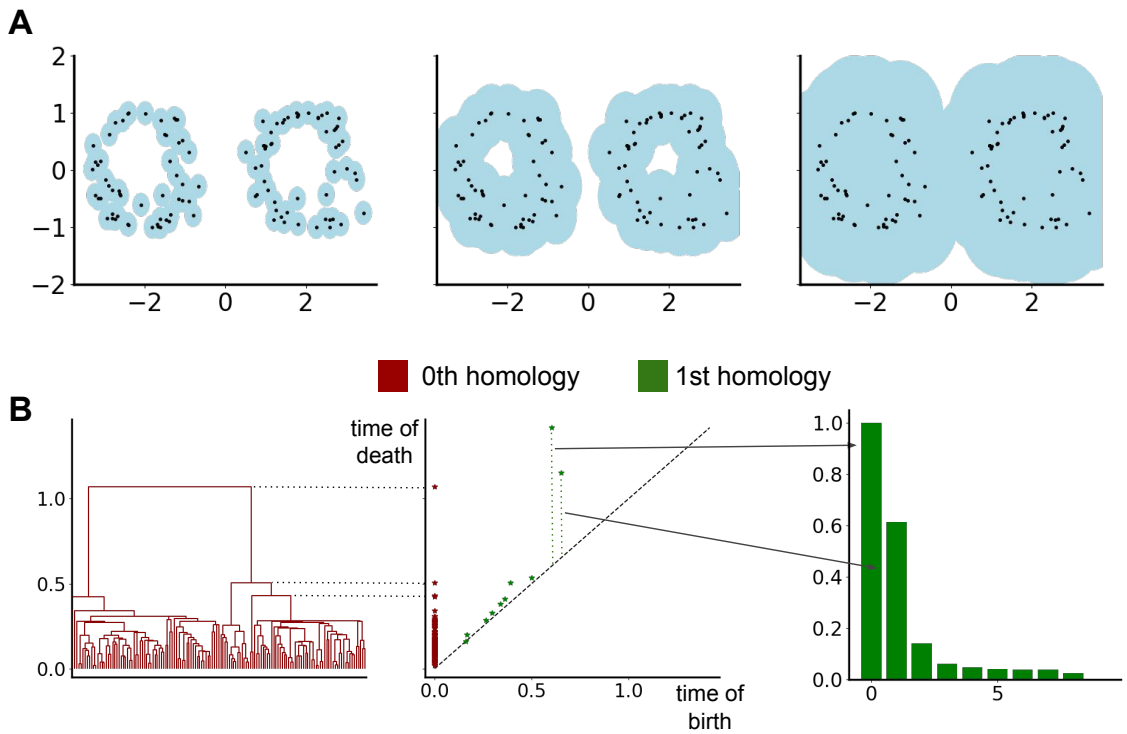

Figure 7: Introduction to topological data analysis. **(A)** Features that persist across a wide range of ε are more likely to be part of an underlying shape. Throughout the process of expanding balls around each point in a point cloud, many topological features emerge and disappear again, like connected components or "holes". Many "short-lived" holes appear throughout this process. However, the two holes that persist the longest are more likely to be a feature of interest than the short-lived ones. **(B)** Different ways to encode persistent topological features. The middle pane is the so-called *persistence diagram*, in which each dot represents a topological feature described by its time of birth and time of death. For 0-dimensional homology classes a dendrogram is a common data structure to keep track of merging clusters (left pane). For 1-dimensional cohomology classes we consider the *persistence sequence*, i.e, the sequence of differences between birth and death times of the persistent classes, ordered by length. The first two bars indicate that there are two clear "signals" of cycles, whereas the rest of the bars indicate cycles that are considered to be "topological noise".

Harmonics. The bijective correspondence above between cohomology classes in degree 1 and homotopy classes of maps into the circle begs the question of which map to pick within the homotopy class to best represent the cohomology class. A particularly useful approach begins with the choice a family of inner-products

$$\langle -, - \rangle_k : C^k(\mathcal{S}; \mathbb{R}) \otimes C^k(\mathcal{S}; \mathbb{R}) \to \mathbb{R}$$

on the cochain complex. The *combinatorial Laplacian* is then the linear map

$$\Delta_k = d_k^\dagger d_k + d_{k-1} d_{k-1}^\dagger : C^k(\mathcal{S}; \mathbb{R}) \to C^k(\mathcal{S}; \mathbb{R}).$$

where $d_k^\dagger : C^{k+1}(\mathcal{S}; \mathbb{R}) \to C^k(\mathcal{S}; \mathbb{R})$ is the adjoint map satisfying

$$\langle d_k \alpha, \beta \rangle_{k+1} = \langle \alpha, d_k^\dagger \beta \rangle_k$$

for all α, β. According to he fundamental theorem of Hodge theory, proven for the simplicial case in [Eck44], there is an isomorphism

$$H^k(\mathcal{S};\mathbb{R}) \cong \operatorname{Ker} \Delta_k$$

between the *space of harmonics* $\operatorname{Ker}\Delta_k$ and the cohomology. The space of harmonics depends on the choice of inner product and has a basis the k-cochains that minimize the induced norm within their cohomology class.

Simplicial circular coordinates. A simplicial cochain $\alpha \in C^1(\mathcal{S};\mathbb{R})$ has *integral periods* if there exists a integral cocycle $\alpha_{\mathbb{Z}} \in C^1(\mathcal{S};\mathbb{Z})$ such that

$$\alpha = \alpha_{\mathbb{Z}} + d\tilde{\theta}$$

for some $\tilde{\theta} \in C^0(\mathcal{S};\mathbb{R})$. In this case, there is a well defined function

$$\theta : \mathcal{S}^0 \to S^1 = \mathbb{R}/\mathbb{Z}; \qquad \theta(x) = \exp\left[2\pi i\big(\tilde{\theta}(x)\big)\right]$$

satisfying

$$\alpha\big([x_p, x_q]\big) = d\tilde{\theta}\big([x_p, x_q]\big) \mod \mathbb{Z}. \tag{4}$$

The map θ extends to a continuous map defined on the geometric realization $|\mathcal{S}| \to S^1$ (see [dSMVJ11]).

Harmonic smoothing. Equation 4 implies that the harmonic representative

$$\alpha_h = \alpha_{\mathbb{Z}} + d\tilde{\theta}_h$$

corresponds to a circular coordinate $\theta_h : \mathcal{S}^0 \to S^1$ with the 'smoothest' gradient (depending on choice of inner product). This is attained as a solution to the Dirichlet equation

$$d_0^{\dagger}\alpha_{\mathbb{Z}} = -\Delta_0\tilde{\theta}_h,$$

which is computed via least-squares regression. We refer to the associate continuous map $\theta_h : |\mathcal{S}| \to S^1$ as the *circular coordinate associated to* $\alpha_{\mathbb{Z}}$ and refer the reader to [dSMVJ11, PP23] for further details.

Remark 5. *The map θ_h depends only on the inner products in dimensions* 0 *and* 1.

Persistence-based coordinates. Persistent cohomology algorithms are computed over cofficients in the field $\mathbb{Z}/p$ for some prime p. This issue is mitigated by the following observation from [dSMVJ11].

Proposition 6. *For a finite simplicial complex $\mathcal{S}$, the map*

$$H^1(\mathcal{S};\mathbb{Z}) \to H^1(\mathcal{S};\mathbb{Z}/p)$$

induced by the unique ring map $\mathbb{Z} \to \mathbb{Z}/p$ is surjective whenever $H^2(\mathcal{S};\mathbb{Z}$ has no p-torsion.

Since torsion is remarkably rare in data, we can assume that we have a valid lift of a $(\mathbb{Z}/p)$-class to a $\mathbb{Z}$-class. For an interval $\mathbb{I}_{[\varepsilon,\delta]}$ in $H^1(X_\bullet;\mathbb{Z}_p)$, we lift the cocycle representative $\alpha \in C^1(X_\varepsilon;\mathbb{Z}_p)$ to a $\mathbb{Z}$-valued cocycle $\alpha_{\mathbb{Z}} \in C^1(X_\varepsilon;\mathbb{Z})$. The corresponding circular coordinate is the harmonic circular coordinate

$$\theta_h : X_\varepsilon \to S^1$$

corresponding to the harmonic representative $\alpha_h = \alpha_{\mathbb{Z}} + d\tilde{\theta}_h$. See [Per20] for details.

Density weighting. The single-cell data we work with has highly non-uniform density in practice. In [PP23], the authors define cochain inner products that correct for density and metric information, following the graph-Laplacian manifold-approximation literature [BN05].

For a finite point cloud (X, d) in $\mathbb{R}^n$, define the *kernel* and *density* functions as

$$K_t(x_i, x_j) = \frac{1}{(4\pi t)^{n/2}} e^{\frac{-d(x_i,x_j)^2}{4t}} \qquad \text{and} \qquad P_t(x_i) = \frac{\sum_j K_t(x_i, x_j)}{\sum_{k,j} K_t(x_k, x_j)}$$

for a parameter $t \in \mathbb{R}_+$. For a cut-off distance $\varepsilon \in \mathbb{R}_+$, let

$$K_t^\varepsilon(x_i, x_j) = \begin{cases} K_t(x_i, x_j) & d(x_i, x_j) \leq \varepsilon \\ 0 & \text{else,} \end{cases} \tag{5}$$

and define P_t^ε similarly. Define the *kernel-density weighting* $W : X \times X \to \mathbb{R}$ by

$$W_t^\varepsilon(x_i, x_j) = \frac{K_t^\varepsilon(x_i, x_j)}{P_t^\varepsilon(x_i) P_t^\varepsilon(x_j)}$$

The inner products on $C^0(\mathcal{S};\mathbb{R})$ and $C^1(\mathcal{S};\mathbb{R})$ given in [PP23] are defined via the formulae

$$\langle f, g\rangle_0 = \sum_{x \in X} \frac{1}{P_t^\varepsilon(x)} f(x) g(x) \qquad \langle \alpha, \beta\rangle_1 = \sum_{[x_i,x_j] \in X_\varepsilon^1} W_t^\varepsilon(x_i, x_j) \alpha\big([x_i, x_j]\big) \beta\big([x_i, x_j]\big). \tag{6}$$

Remark 7. *The inner-products correspond to density-normalized correlations of* 0*- and* 1*-cochains.*

When X is sampled from a distribution on an embedded Riemannian manifold, the combinatorial Laplacian Δ_0 under these inner products approximates (for an appropriate choice of t) the intrinsic Laplace-Beltrami operator, irrespective of distribution (Thm. 3.1 [PP23]). For an illustrated example of the re-weighting effect, see Figure 13 or [PP23].

Sparse circular coordinates. In practice, one mitigates the computational complexity of persistent cohomology by working over landmark subsets [Per20]. Briefly, a landmark set

$$L = \{x'_1, \ldots, x'_\ell\} \subset X$$

is attained by inductive maxmin sampling

$$\{x'_1, x'_2\} = \arg\max_{x_i, x_j}\{d(x_i, x_j)\} \qquad x'_k = \arg\max_{x_i} \ \min\left\{d(x_i, x'_j) \mid j < k\right\} \tag{7}$$

until reaching a specified distance

$$d(L, X) < \delta \cdot \mathrm{Diam}(X) \qquad \text{implying} \qquad X \subseteq \bigcup_{1 \le k \le \ell} B_\delta(x'_k)$$

for a parameter $\delta \in \mathbb{R}_+$ (we use 0.01 by default). Assuming the longest 1-interval $\mathbb{I}_{[\varepsilon, \varepsilon']}$ in $H^1(L_\bullet; \mathbb{Z}_p)$ satisfies $\delta < \varepsilon$, we compute the harmonic circular coordinate

$$\theta'_h : L_\varepsilon \to S^1 \qquad \alpha'_h = \alpha'_{\mathbb{Z}} + d\tilde{\theta}'_h \in C^1(L_\varepsilon; \mathbb{R})$$

for L at resolution ε using the inner products from Equation (6). Following [Per20], the circular coordinate is extended from L to X by the formula

$$\theta : X \to S^1 \tag{8}$$

$$x \in B_\varepsilon(x'_j) \mapsto \exp\left[2\pi i\left(\tilde{\theta}'_h(x'_j) + \sum_{k=1}^{\ell} \mu_k(x)\alpha'_h([x'_i, x'_j])\right)\right] \tag{9}$$

where $\mu_i : B_\varepsilon(x'_i) \to \mathbb{R}_+$ is a partition of unity. This is well-defined, independent of choice of covering neighbourhood for each x (see [Per20] Thms 3.2 and 4.2).

4.3 Circular Gene Set Enrichment

Gene set enrichment analysis [STM+05] is a ubiquitous and useful tool in single-cell RNA seq analysis. In our work, the goal is to expand this concept to that of *circular enrichment*, where a gene set is said to be circularly enriched in a population if there is a significant circular structure in the RNA expression data restricted to that gene set.

Gene set projections. Let X be an single-cell RNA expression experiment. We will assume that X has already been pre-preprocessed, including restricting to subsets for a specific cluster or cell type.

Definition 8. *For a given gene set $\mathcal{G} = \{g_1, \ldots, g_\ell\}$ define:*

1. *$X_\mathcal{G}$ to be the projection of X onto $\mathcal{G}$,*
2. *$X^k_\mathcal{G}$ as the projection of $X_\mathcal{G}$ onto the first k PCA-components.*

Note that X, $X_\mathcal{G}$, and $X^k_\mathcal{G}$ all contain the same number of points, with a change of metric. In practice, we take $k = 5$ somewhat arbitrarily. Our rationale for performing PCA is that we wish to find only gene sets where the circular structure is the strongest signal in the subspace. We take $X^k_\mathcal{G}$ as a basic geometric model of the dominant biological phenomena associated with $\mathcal{G}$.

Ring score. To sift efficiently through gene sets, we define a summary statistic that detects the existence of dominant circular structure in a subspace $X^k_{\mathcal{G}}$. Intuitively, a point cloud is *ring-structured* if it has a single pronounced "hole". Prominent examples of such point clouds are sampled versions of a circle, an annulus (e.g., a "ring"), or a (hollow) cylinder. The philosophy of the ring score is to measure to what extent a point cloud deviates from these shapes in terms of its overall circular structure. A natural choice is thus to define this quantity as a function of the 1-dimensional persistence diagram.

Suppose we have a persistence diagram consisting of n birth-death pairs

$$\mathrm{dgm}^1(X^k_{\mathcal{G}}) = \left\{ (b_i, d_i) \mid 0 \leq i \leq n \right\}$$

where the birth-death pairs are sorted in such a way that their corresponding persistence life spans, $p_i = d_i - b_i$, are sorted in decreasing order, i.e., $p_0 \geq p_1 \geq \dots$.

Definition 9 (Generalized Ring-Score Formula)**.** *Let $\mathbf{p} = (p_i)_{i \in \mathbf{N}}$ be a sequence of input-parameters in descending order, $p_0 \geq p_1 \geq \dots \geq 0$, and let $N \geq p_0$ be a normalization constant. Then we define the ring-score formula as*

$$s(\mathbf{p}, N) = \frac{1}{N} \left(p_0 - \sum_{i=1}^{\infty} \frac{p_i}{2^i} \right) .$$

Remark 10. *Note that $s \in [0, 1]$, where a higher ring score indicates a more dominant circular feature. The idea is that p_0 represents the main circular feature of interest, i.e., "the ring", whereas all subsequent persistence points "obscure" the topology of this ring. Indeed, by definition $0 \leq p_i \leq p_0$ for all i and $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$, and thus*

$$\begin{aligned} 0 = \frac{1}{N}(p_0 - p_0) &= \frac{1}{N} \left(p_0 - \sum_{i=i}^{\infty} \frac{p_0}{2^i} \right) \\ &\leq s(\mathbf{p}, N) \\ &\leq \frac{1}{N} \left(p_0 - \sum_{i=i}^{\infty} \frac{0}{2^i} \right) = \frac{p_0}{N} \leq 1 . \end{aligned}$$

The exponentially decaying weights are chosen to reflect the fact that less persistent features are less significant and to ensure that the range of s is contained in $[0, 1]$ (see Fig.14A).

Let X be a point cloud. We say we *scale* the point cloud by a factor $c > 0$ and write $c \cdot X$ by multiplying all pairwise distances by this factor while leaving the center of mass as it is. Formally, if we denote the coordinates of a point i by X_i we can calculate the coordinates of the resulting point of $c \cdot X$ by

$$(c \cdot X)_i := cX_i - (c - 1)\mu(X) ,$$

where $\mu(X)$ denotes the coordinate vector of the center of mass of X and here multiplication with a scalar is the usual coordinate-wise multiplication.
Assuming that $\mathbf{p}(X) = (d_i(X) - b_i(X))_{i \in \mathbb{N}}$ is the vector of persistence life spans of a point cloud X and that the normalization constant $N(X)$ scales with $p_0(X)$, i.e.,

that $\frac{p_0(c\cdot X)}{N(c\cdot X)} = \frac{p_0(X)}{N(X)}$ for all $c > 0$, then the ring score itself is invariant under metric scaling. In particular, one can always choose as a normalization the most persistent life span $N = p_0(X)$ or the diameter of the point cloud $N = \mathrm{diam}(X)$. In Euclidean space, $\frac{p_0(X)}{\mathrm{diam}(X)}$ attains its maximum when X is isometric to a circle, in which case we get $\frac{p_0(X)}{\mathrm{diam}(X)} = \frac{\sqrt{3}}{2}$.

Finally, observe that the generalized ring score formula from Def. 9 does not require the input sequence $\mathbf{p}$ to be persistent life spans from a given point cloud. In fact, there has been a growing interest in studying *persistence ratios*, $\pi_i = \frac{\mathrm{death}_i}{\mathrm{birth}_i}$, as opposed to persistence life spans, $p_i = \mathrm{death}_i - \mathrm{birth}_i$ [BKS17]. We note that when using persistent ratios as opposed to life spans, there is no shape of a maximal π_0-to-diameter ratio and so we simply set $N(X) = \pi_0(X)$. Thus we define the following two families of ring scores

Definition 11 (Diameter and Ratio Scores)**.** *Let $X \subset \mathbb{R}^n$ be a countable subset of an n-dimensional Euclidean space with diameter $d = \mathrm{diam}(X) < \infty$ and denote by $\mathbf{p}$ and by $\boldsymbol{\pi}$ the corresponding persistence life spans and persistence ratios respectively. Then we define*

- *the* diameter ring-score *as* $s\left(\mathbf{p}, \frac{d\sqrt{3}}{2}\right)$,
- *the* ratio ring-score *as* $s(\boldsymbol{\pi}, \pi_0)$.

Ring-score significance estimation. We estimate the statistical significance of ring scores in terms of nominal p-value, using an empirical gene set-based permutation test procedure that preserves the correlation structure of the gene expression data. For any observed gene set $\mathcal{G}^{\mathrm{obs}}$ of size ℓ, we randomly generate a gene set $\mathcal{G}$ of size ℓ to create and recompute ring scores of the random gene sets, generating a null distribution S_{NULL} for the ring score. The empirical p-value of the observed ring score is then calculated relative to this null distribution. We note that the permutation of gene labels preserves cell-cell correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting both genes and cells.

Explicitly, given an observed gene set with ℓ genes $\mathcal{G}^{\mathrm{obs}} = \{g_1^{\mathrm{obs}}, \ldots, g_\ell^{\mathrm{obs}}\}$ and preprocessed data X, we calculate significance of the ring score $s(X^k_{\mathcal{G}^{\mathrm{obs}}})$ compared to a set of scores $S_{\mathrm{NULL}}(k, \ell)$ computed with randomly selected gene sets of size l in three steps.

1. Randomly select ℓ distinct genes, $\mathcal{G} = \{g_1, \ldots, g_\ell\}$ and compute $s(X^k_{\mathcal{G}})$.
2. Repeat step 1 for 2^{10} randomly generated gene sets, and create a null-distribution of the corresponding ring scores $S_{\mathrm{NULL}}(k, \ell)$.
3. Use $S_{\mathrm{NULL}}(k, \ell)$ to estimate the nominal p-value for $s\left(X^k_{\mathcal{G}^{\mathrm{obs}}}\right)$.

This procedure is conceptually similar to estimating p-values for gene set enrichment analysis. [STM+05]

4.4 Cyclicity Analysis

4.4.1 Path cyclicity analysis

Cyclicity analysis is a method introduced by [BS16] to analyse the interaction between periodic paths in $\mathbb{R}^n$ and their embedding dimensions. In this section, we provide an overview of the definitions and basic theory.

Lead-lag matrix. Let $\gamma : S^1 \to \mathbb{R}^n$ be a closed path. The *lead-lag matrix* of γ is the $n \times n$ matrix

$$\mathbf{L}_{i,j}(\gamma) = \int_{S^1} \gamma^*(x_i dx_j - x_j dx_i) = \int_{S^1} \left(\gamma^i \frac{d\gamma^j}{dt} - \gamma^j \frac{d\gamma^i}{dt}\right) dt. \tag{10}$$

The (i, j)-entry of this matrix is called the (i, j) *lead-lag coefficient*. In simple models (cf. [14), this coefficient measures the extent to which coordinate i leads or lags behind coordinate j. Following from the properties of integration, the lead-lag matrix is

1. skew-symmetric, and
2. invariant under orientation-preserving reparametrization of γ.

If $\overline{\gamma} : D \to \mathbb{R}^n$ is differentiable map of a disk into $\mathbb{R}^n$ such that the composite

$$S^1 \hookrightarrow D \xrightarrow{\overline{\gamma}} \mathbb{R}^n$$

is equal to γ, then Stoke's theorem implies that

$$\mathbf{L}_{i,j}(\gamma) = \int_{\partial D} \gamma^*(x_i dx_j - x_j dx_i) = \frac{1}{2} \int_D \overline{\gamma}^*(dx_i \wedge dx_j) \tag{11}$$

i.e., the lead-lag is proportional to the signed-area of the disk in the (ij)-plane in $\mathbb{R}^n$. It follows that the lead-lag matrix is invariant under translation of γ.

Chain-of-offsets model. For intepretability, [BS16] presents the following model. Suppose $\chi : S^1 \to \mathbb{R}$ is a periodic function with Fourier decomposition given by the real part of

$$\chi(t) = \sum_{m=0}^{\infty} z_m \exp(2\pi i m t) \tag{12}$$

for complex Fourier coefficients $z_m \in \mathbb{C}$. A path $\gamma : S^1 \to \mathbb{R}^n$ is a *chain-of-offsets model* if there exist $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in [0, 1)^n$ such that the components satisfy

$$\gamma^i(t) = \mathbf{a}_i \cdot \chi(t - \mathbf{b}_i). \tag{13}$$

In terms of biology, the chain-of-offsets model assumes that gene expression values in different coordinates around a circular expression pattern are offset copies of one another — up to a reparametrization.

Lead-lag of the chain-of-offsets. The upside of the chain-of-offsets model is that the lead-lag matrix has the following closed form, due to [BS16]. Suppose that $\gamma : S^1 \to \mathbb{R}^n$ has offsets as in Section 4.4.1 and that the generating function $\chi : S^1 \to \mathbb{R}$ has Fourier decomposition in Equation (12).

Proposition 12 (Prop. 5.1., [BS16])**.** *The lead-lag matrix of γ is*

$$\mathbf{L}_{i,j}(\gamma) = 2\pi \mathbf{a}_i \mathbf{a}_j \left(\sum_m m|z_m|^2 \sin\left(2\pi m(\mathbf{b}_i - \mathbf{b}_j)\right) \right)$$

Ellipse model Following [BS16], suppose we have a simple case where the primary function

$$\chi(t) = z_1 \exp(2\pi i t) \tag{14}$$

has a Fourier series concentrated in the first harmonic, i.e., that the path is an ellipse centered about the origin. In this case, applying Proposition 12 to the chain-of-offsets $\gamma : S^1 \to \mathbb{R}^n$ shows that the lead-lag matrix is

$$\mathbf{L}_{i,j}(\gamma) = 2\pi |z_1|^2 \sin\Big(2\pi(\mathbf{b}_i - \mathbf{b_j})\Big), \tag{15}$$

we can derive an important interpretation:

the (ij)-leadlag coefficient is proportional to the sine of the difference in offsets of dimension i and j.

Such a quantity is maximized and minimized when coordinate i is offset by a quarter and a negative quarter of the total period from coordinate j respectively. As is argued in [BS16], if the Fourier series of $\chi(t)$ is dominated by z_1, i.e., if

$$|z_1|^2 \gg \sum_{m \geq 2} |z_m|^2$$

then

$$\mathbf{L}_{i,j}(\gamma) \approx 2\pi |z_1|^2 \sin\Big(2\pi(\mathbf{b}_i - \mathbf{b_j})\Big)$$

providing similar interpretation of the lead-lag coefficients up to noise.

Complex diagonalization. Motivated by the fact that the lead-lag matrix is skew-symmetric, we review some basic results about the spectral theory of skew-symmetric matrices.

Let $\mathbf{A}$ be a real $n \times n$ skew-symmetric matrix. A basic result from linear algebra is that there exist a real orthonormal matrix $\mathbf{P} \in O(n)$ and a block diagonal $n \times n$ matrix

$$\Sigma = \begin{bmatrix} 0 & -\lambda_1 & & & & & \\ \lambda_1 & 0 & & & & & \\ & & 0 & -\lambda_2 & & & \\ & & \lambda_2 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & 0 \end{bmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \cdots$$

satisfying $\mathbf{A} = \mathbf{P}^{-1}\Sigma\mathbf{P}$. Equivalently, treating $\mathbf{A}$ as an $n \times n$ complex matrix, there exists a unitary matrix $\mathbf{U} \in U(n)$ and matrix

$$\widetilde{\Sigma} = \begin{bmatrix} i\lambda_1 & & & & & & \\ & -i\lambda_1 & & & & & \\ & & i\lambda_2 & & & & \\ & & & -i\lambda_2 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & 0 \end{bmatrix}$$

satisfying $\mathbf{A} = \mathbf{U}^*\widetilde{\Sigma}\mathbf{U}$. The $2k$-th complex column vector $\mathbf{u}+i\mathbf{v}$ of $\mathbf{U}$ corresponds to the $2k$-th and $(2k+1)$-th real column vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$ of the orthonormal matrix $\mathbf{P}$. Geometrically, $\mathbf{A}$ acts on the real plane spanned by $\mathbf{u}$ and $\mathbf{v}$ via a $\pi/2$-radian (resp. $-\pi/2$-rotation) plus stretching by λ_j for each pair of complex conjugate eigenvectors $\mathbf{u}(j) + i\mathbf{v}(j)$ (resp. $\mathbf{u}(j) - i\mathbf{v}(j)$) in the columns of $\mathbf{U}$.

Spectral theory and offset recovery. Suppose that a chain of offsets $\gamma : S^1 \to \mathbb{R}^n$ comes from the ellipse model (14) with components

$$\gamma^i(t) = \mathbf{a}_i \cdot \chi(t - \mathbf{b}_i).$$

The difference of angles formula applied to the lead-lag matrix in 15 yields

$$\begin{aligned} \mathbf{L}_{i,j}(\gamma) &= 2\pi\mathbf{a}_i\mathbf{a}_j|z_1|^2 \sin\Big(2\pi(\mathbf{b}_i - \mathbf{b_j})\Big) \\ &= 2\pi\mathbf{a}_i\mathbf{a}_j|z_1|^2\Big[\sin(2\pi\mathbf{b}_i)\cos(2\pi\mathbf{b}_j) - \sin(2\pi\mathbf{b}_j)\cos(2\pi\mathbf{b}_i)\Big]. \end{aligned}$$

Denote amplitude-offset vectors by

$$\mathbf{u} = \mathbf{a}\sin(2\pi\mathbf{b}) \qquad \text{and} \qquad \mathbf{v} = \mathbf{a}\cos(2\pi\mathbf{b})$$

respectively. At the matrix level, the relation above implies that

$$\mathbf{L}(\gamma) = 2\pi|z_1|^2\Big(\mathbf{u}\otimes\mathbf{v} - \mathbf{v}\otimes\mathbf{u}\Big). \tag{16}$$

In this case, the lead-lag matrix is a rank 2 matrix, whose coimage (i.e., the span of its row vectors) is concentrated in the $\mathbf{uv}$-plane. As per the discussion above, the fact that $\mathbf{L}(\gamma)$ is skew-symmetric implies that it must be a rotation in this plane. The eigenvalues corresponding to a non-zero complex-conjugate pair of eigenvectors can be calculated in terms of $\mathbf{u}$ and $\mathbf{v}$ as

$$\pm i\sin(\theta)\|\mathbf{u}\|\|\mathbf{v}\|$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Following [BS16], the non-trivial complex eigenvector $\mathbf{z} = \mathrm{Re}(\mathbf{z}) + i\mathrm{Im}(\mathbf{z}) \in \mathbb{C}^n$ of $\mathbf{L}(\gamma)$ has components

$$\begin{bmatrix}\mathrm{Re}(\mathbf{z})_k \\ \mathrm{Im}(\mathbf{z})_k\end{bmatrix} = \begin{bmatrix} \frac{-\cos\theta}{\|\mathbf{u}\|} & \frac{1}{\|\mathbf{v}\|} \\ \frac{\sin\theta}{\|\mathbf{u}\|} & 0 \end{bmatrix}\begin{bmatrix}\mathbf{u}_k \\ \mathbf{v}_k\end{bmatrix} = 2\pi|z_1|^2\mathbf{a}_k\begin{bmatrix} \frac{-\cos\theta}{\|\mathbf{u}\|} & \frac{1}{\|\mathbf{v}\|} \\ \frac{\sin\theta}{\|\mathbf{u}\|} & 0 \end{bmatrix}\begin{bmatrix}\cos(2\pi\mathbf{b}_k) \\ \sin(2\pi\mathbf{b}_k)\end{bmatrix}. \tag{17}$$

This formula implies the following important point, as noted in [BS16]:

the amplitudes and offsets of dimensions in the ellipse model are related to the principal eigenvector of the lead-lag matrix by a linear transformation.

Hence, one recovers the cyclic order of the coordinates.

Remark 13. *By 17, the extent to which the amplitude is recoverable depends on the angle θ between $\mathbf{u}$ and $\mathbf{v}$. As a heuristic justification, we note that* sin *and* cos *are orthogonal as functions, with a further mathematical discussion of when the amplitude estimates are reliable to furture work.*

4.4.2 Simplicial cyclicity analysis

We present here our generalization of cyclicity analysis to embedded simplicial complexes.

Manifold model. Let M be an m-dimensional oriented Riemannian manifold with volume form Vol_M, and denote its Hodge star operator by

$$* : \Omega^{m-k}(M) \to \Omega^k(M).$$

This induces an inner product on k-forms defined by

$$\langle \omega, \eta \rangle_{\Omega^k(M)} = \int_M *\omega \wedge \eta \, d\mathrm{Vol}_M.$$

For a 1-form $\eta \in \Omega^1(M)$ and a smooth map $\varphi : M \to \mathbb{R}^n$, the *manifold lead-lag* is given by

$$\mathbf{L}_{i,j}(\varphi, \eta) = \langle \eta, \varphi^* \omega_{i,j} \rangle_{\Omega^1(M)}. \tag{18}$$

where

$$\omega_{i,j} = x_i dx_j - x_j dx_i \in \Omega^1(\mathbb{R}^n).$$

Path lead-lag is a special case of manifold lead-lag. Let $d\theta = *1$ be the volume form on S^1 with the standard Riemannian structure, $\gamma : S^1 \to \mathbb{R}^n$ a closed path, and $\mathbf{L}_{i,j}(\gamma)$ the standard lead-lag matrix of the path γ given in Equation (10).

Proposition 14. $\mathbf{L}_{i,j}(\gamma) = \mathbf{L}_{i,j}(\gamma, d\theta)$.

Proof. We have

$$\mathbf{L}_{i,j}(\gamma) = \int_{S^1} \gamma^* \omega_{i,j} = \int_{S^1} 1 \wedge \gamma^* \omega_{i,j} = \int_{S^1} (*d\theta) \wedge \gamma^* \omega_{i,j} = \langle d\theta, \gamma^* \omega_{i,j} \rangle_{\Omega^1(S^1)}.$$

□

Simplicial model. Let $\varphi : \mathcal{S} \to \mathbb{R}^n$ be an oriented simplicial complex affinely embedded in $\mathbb{R}^n$. For an edge $[x_p, x_q] \in \mathcal{S}$, define

$$\varphi_{p,q} : I \xrightarrow{[x_p,x_q]} \mathcal{S} \xrightarrow{\varphi} \mathbb{R}^n$$

to be the path from $\varphi(x_p)$ to $\varphi(x_q)$. Integration induces a simplicial cochain

$$\int \varphi^* \omega_{i,j} : [x_p, x_q] \mapsto \int_I \varphi_{p,q}^* \omega_{i,j} \in \mathbb{R} \tag{19}$$

in $C^1(\mathcal{S}; \mathbb{R})$ for each $1 \leq i, j \leq n$. Equip $C^1(\mathcal{S}; \mathbb{R})$ with the inner product $\langle -, - \rangle_{C^1}$ from Equation (6) which is induced by the geometry of the ambient $\mathbb{R}^n$.

Definition 15. *The simplicial lead-lag matrix over $\alpha \in C^1(\mathcal{S}; \mathbb{R})$ is*

$$\mathbf{L}_{i,j}(\varphi, \alpha) = \left\langle \alpha, \int \varphi^* \omega_{i,j} \right\rangle_{C^1}$$

We reason that since $\langle -, - \rangle_{C^1}$ is an approximation of the underlying Riemannian metric (when it exists), the simplicial lead-lag is an approximation of the manifold model (when the data indeed lies on a manifold).

Geometric interpretation. For an affinely embedded simplicial complex, a simple application of Stokes theorem shows that the integrated cochain in Equation 19 has the following interpretation.

Proposition 16.

$$\int_I \varphi_{p,q}^* \omega_{i,j} = \det \begin{bmatrix} \varphi_i(x_p) & \varphi_i(x_q) \\ \varphi_j(x_p) & \varphi_j(x_q) \end{bmatrix}.$$

This corresponds to the signed area in the ij-plane of the parallelogram spanned by vectors $\phi(x_p)$ and $\phi(x_q)$. In practice, the simplicial complex will be the Rips complex $\mathcal{X}_\varepsilon$ of a normalized count matrix $\mathbf{X}$ conceptualized as a point cloud in ambient gene expression space. If we use the inner product on simplicial 1-cochains described by Equation 6, the coefficients of lead-lag matrix are given by the following formula:

$$\mathbf{L}_{i,j}(\varphi, \alpha) = \sum_{[x_p,x_q] \in \mathcal{X}_\varepsilon} \frac{K_t^\varepsilon(x_p, x_q)}{P_t^\varepsilon(x_p) P_t^\varepsilon(x_q)} \alpha([x_p, x_q]) \cdot \det \begin{bmatrix} \varphi_i(x_p) & \varphi_i(x_q) \\ \varphi_j(x_p) & \varphi_j(x_q) \end{bmatrix}. \tag{20}$$

This quantity represents a type of *α-weighted signed area.* In other words, it is a linear combination of the signed areas in the ij-plane of parallelograms spanned by $\varphi(x_p)$ and $\varphi(x_q)$ for edges $[x_p, x_q]$ in the Rips complex, weighted by the 1-cochain α and density specified by the inner product.

Change-of-basis. Let $(x_1, \ldots, x_n)$ be the standard coordinate functions on $\mathbb{R}^n$, and

$$X = [x_1, \ldots, x_n] \in M^{1\times n}\big(C^\infty(\mathbb{R}^n)\big) \qquad dX = [dx_1, \ldots, dx_n] \in M^{1\times n}\big(\Omega^1(\mathbb{R}^n)\big)$$

These matrices generate the lead-lag forms

$$\mathbf{L}^X := X^T dX - (dX)^T X = [x_i dx_j - x_j dx_i]_{i,j} \in M^{n\times n}\big(\Omega^1(\mathbb{R}^n)\big).$$

Suppose $y_1, \ldots, y_n$ is another orthornormal coordinate system on $\mathbb{R}^n$ such that $Y = X\mathbf{P}$ for some orthonormal matrix $\mathbf{P} \in \mathrm{O}(n)$. Similarly, define the matrix of lead-lag forms with respect to the system Y as

$$\mathbf{L}^Y := Y^T dY - (dY)^T Y = [y_i dy_j - y_j dy_i]_{i,j} \in M^{n\times n}\big(\Omega^1(\mathbb{R}^n)\big).$$

Let $\varphi : \mathcal{S} \to \mathbb{R}^n$ be an affine embedding and $\alpha \in C^1(\mathcal{S};\mathbb{R})$ a simplicial 1-cochain. The lead-lag matrices with respect to the two coordinate systems are given by

$$\mathbf{L}^X_{i,j}(\varphi,\alpha) = \left\langle \alpha, \int \varphi^*(x_i dx_j - x_j dx_i) \right\rangle_{C^1}$$

and

$$\mathbf{L}^Y_{i,j}(\varphi,\alpha) = \left\langle \alpha, \int \varphi^*(y_i dy_j - y_j dy_i) \right\rangle_{C^1}.$$

They are related by the following formula.

Proposition 17. $\mathbf{L}^Y(\varphi,\alpha) = \mathbf{P}^T\mathbf{L}^X(\varphi,\alpha)\mathbf{P}$

Proof. We have

$$\mathbf{L}^Y = Y^T dY - (dY)^T Y = \mathbf{P}^T X^T dX\mathbf{P} - \mathbf{P}^T (dX)^T X\mathbf{P} = \mathbf{P}^T\mathbf{L}^X\mathbf{P}.$$

The linearity of integration, pullbacks and the operator $\langle \alpha, -\rangle_{C^1}$ implies that

$$\mathbf{L}^Y(\varphi,\alpha) = \mathbf{P}^T\mathbf{L}^X(\varphi,\alpha)\mathbf{P}.$$

□

Spectral theory. Now suppose we have computed the lead-lag matrix $\mathbf{L}^X(\varphi,\alpha)$ with respect to an embedding, a 1-cochain, and the standard coordinates X on $\mathbb{R}^n$. Putting our matrix in canonical form yields a real orthonormal matrix $\mathbf{P} \in O(n)$ and block diagonal $n \times n$ matrix

$$\Sigma = \begin{bmatrix} 0 & -\lambda_1 & & & & & \\ \lambda_1 & 0 & & & & & \\ & & 0 & -\lambda_2 & & & \\ & & \lambda_2 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & 0 \end{bmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \tag{21}$$

satisfying $\mathbf{L}^X(\varphi,\alpha) = \mathbf{P}^T\Sigma\mathbf{P}$. Define a new coordinate system on $\mathbb{R}^n$ by

$$Y = [y_1, \ldots, y_n] = X\mathbf{P}^T$$

for $X = [x_1, \ldots, x_n]$ as above. A consequence of Proposition 17 is that we can interpret Σ from Equation (21) as a lead-lag matrix

$$\Sigma_{i,j} = \mathbf{L}^Y_{i,j}(\varphi,\alpha) = \left\langle \alpha, \int \varphi^*(y_i dy_j - y_j dy_i) \right\rangle_{C^1}$$

in the Y coordinates. In light of the discussion above of α-weighted signed area (20), we have the following interpretation:

the value λ_i in Equation (21) corresponds to the α-weighted signed area in the plane spanned by y_{2i} and y_{2i+1}.

Thus, putting our lead-lag matrix in canonical form identifies the plane spanned by y_0 and y_1 as the *principal plane*, i.e., that which maximizes the α-weighted signed area. Further, the ratio λ_1/λ_2 describes the extent to which the α-weighted signed area is fully captured by the principal plane.

Recentering. A major difference between simplicial and path cyclicity analyses is that simplicial cyclicity analysis is not invariant under translations of the emb edded simplicial complex. To mitigate the effect of translations, we first perform a procedure to "recenter" the data around the cochain α. Given a 1-cochain α, define a weighting on the vertices by $\eta(v_i) = \sum_j |\, \alpha_h(e_{i,j})\,|$ for the cochain α_h evaluated on each oriented 1-simplex $e_{i,j}$ in $\mathcal{S}$. The α-weighted mean is defined as

$$\mu(\alpha) := \frac{\sum_p \eta(x_p)\varphi(x_p)}{\sum_p \eta(x_p)} \in \mathbb{R}^n.$$

The recentered data is defined to be $\widehat{\varphi}(x_p) = \varphi(x_p) - \mu(\alpha)$ for all p.

Discussion. In cases where data is sampled from S^1, and α approximates the volume form, we claim that the simplicial lead-lag is a good model of the path lead-lag. In this case, it is reasonable to expect that the spectral decomposition of the simplicial lead-lag matrix exhibits a similar connection with the offsets in the Fourier decomposition of the chain-of-offsets model. In order to rigorously generalize this relationship, one needs a version of the chain-of-offsets model for manifolds rather than paths. We plan to define and study such a model in future work.

4.4.3 Cyclicity analysis on single-cell data

The goal of this section is contextualize Figure 2 in simplicial lead-lag theory.

Dominant geometric data. Recall that we denote the first k components of PCA projection of a single-cell experiment X calculated in the subspace spanned by genes $\mathcal{G}$ by $X^k_{\mathcal{G}}$. Our principle is that $X^k_{\mathcal{G}}$ is the basic geometric model of the dominant biological phenomena in $\mathcal{G}$. The points of the metric space $X^k_{\mathcal{G}}$ are in bijection with the points in X. with the metric induced by the PCA projection $X^k_{\mathcal{G}} \to \mathbb{R}^k$ (Figure 2A).

The following geometric data are associated to this choice of metric.

1. The persistent cohomology $\mathcal{D}_1(X^k_{\mathcal{G}})$ and ring score $s(X^k_{\mathcal{G}})$
2. The dominant interval $\mathbb{I}_{[\varepsilon,\delta]}$ and associated Rips complex $X^k_{\mathcal{G},\varepsilon}$
3. Inner-products over $C^0(X^k_{\mathcal{G},\varepsilon};\mathbb{R})$ and $C^1(X^k_{\mathcal{G},\varepsilon};\mathbb{R})$, via Equation (6)
4. Harmonic representative $\alpha_h \in C^1(X^k_{\mathcal{G},\varepsilon};\mathbb{R})$ and circular coordinate $\theta_h : X^k_{\mathcal{G},\varepsilon} \to S^1$

The harmonic representative α_h is a simplicial model for the locally-defined gradient vector field[2] of θ_h in the Riemannian case (Figure 2B). A key assumption is that this is a model for the likely circular dynamics in the data, noting that orientation is arbitrary and needs to be determined by the user based on biological information.

Lifting. The geometric data model the underlying biological process in $\mathcal{G}$. To understand which genes influence this process, we examine the lift (Figure 2B)

$$\begin{tikzcd} & \mathbb{R}^{\mathcal{G}} \arrow[d, "\text{PCA}"] \\ X^k_{\mathcal{G},\varepsilon} \arrow[ur, "\varphi_{\mathcal{G}}"] \arrow[r, hook, "\text{isometry}"'] & \mathbb{R}^k \end{tikzcd}$$

where φ_G is map that linearly extends the expression of genes G at each node to the Rips complex at parameter $\varepsilon \in \mathbb{R}$. The lead-lag matrix (Figure 2C)

$$\mathbf{L}_{i,j}(\varphi_{\mathcal{G}}, \alpha_h) = \langle \alpha_h, \int \varphi^*_{\mathcal{G}} \omega_{i,j} \rangle_{C^1}$$

for $1 \leq i, j \leq |\mathcal{G}|$ represents pairwise gene ordering relative to α_h. The lead-lag projection and and principal complex eigenvector components approximate the global order in which the genes in $\mathcal{G}$ occur along α_h (Figure 2D).

[2]Formally, the dual vector field of the associated 1-form.

References

[AFL22] Benjamin J. Auerbach, Garret A. FitzGerald, and Mingyao Li. Tempo: An unsupervised bayesian algorithm for circadian phase inference in single-cell transcriptomics. *Nature Communications*, 13(1), Nov 2022.

[BGDS+99] R. Bhatia-Gaur, A. A. Donjacour, P. J. Sciavolino, M. Kim, N. Desai, P. Young, C. R. Norton, T. Gridley, R. D. Cardiff, G. R. Cunha, C. Abate-Shen, and M. M. Shen. Roles for nkx3.1 in prostate development and cancer. *Genes amp; Development*, 13(8):966–977, April 1999.

[BHS+18] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[BKS17] Omer Bobrowski, Matthew Kahle, and Primoz Skraba. Maximally persistent cycles in random geometric complexes. *The Annals of Applied Probability*, pages 2032–2060, 2017.

[BLP+20] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, August 2020.

[BN05] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Learning Theory*, page 486–500, 2005.

[BS16] Yuliy Baryshnikov and Emily Schlafly. Cyclicity in multivariate time series and applications to functional mri data. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1625–1630, 2016.

[CCG+09] Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Memoli, and Steve Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28(5):1393–1403, Jul 2009.

[CFX06] Long Cai, Nir Friedman, and X. Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, March 2006.

[CGP+18] Andrea Clocchiatti, Soumitra Ghosh, Maria-Giuseppina Procopio, Luigi Mazzeo, Pino Bordignon, Paola Ostano, Sandro Goruppi, Giulia Bottoni, Atul Katarkar, Mitchell Levesque, et al. Androgen receptor functions as transcriptional repressor of cancer-associated fibroblast activation. *The Journal of clinical investigation*, 128(12):5531–5548, 2018.

[CHC98] TING CHEN, HONGYU L. HE, and GEORGE M. CHURCH. Modeling gene expression with differential equations. In *Biocomputing '99*. WORLD SCIENTIFIC, December 1998.

[CKL+22] Laura Capolupo, Irina Khven, Alex R. Lederer, Luigi Mazzeo, Galina Glousker, Sylvia Ho, Francesco Russo, Jonathan Paz Montoya, Dhaka R. Bhandari, Andrew P. Bowman, Shane R. Ellis, Romain Guiet, Olivier Burri, Johanna Detzner, Johannes Muthing, Krisztian Homicsko, François Kuonen, Michel Gilliet, Bernhard Spengler, Ron M. A. Heeren, G. Paolo Dotto, Gioele La Manno, and Giovanni D'Angelo. Sphingolipids control dermal fibroblast heterogeneity. *Science*, 376(6590):eabh1623, 2022.

[CR20] Mathieu Carrière and Raúl Rabadán. Topological data analysis of single-cell hi-c contact maps. In Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule, editors, *Topological Data Analysis*, pages 147–162, Cham, 2020. Springer International Publishing.

[dSMVJ11] Vin de Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete & Computational Geometry*, 45(4):737–759, Jun 2011.

[Eck44] Beno Eckmann. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, December 1944.

[ELZ00] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.

[FGP24] Meichen Fang, Gennady Gorin, and Lior Pachter. Trajectory inference from single-cell genomics data with a process time model. *bioRxiv*, 2024.

[GPZC05] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, December 2005.

[GYC19] Kiya W. Govek, Venkata S. Yamajala, and Pablo G. Camara. Clustering-independent analysis of genomic data using spectral simplicial theory. *PLOS Computational Biology*, 15(11):1–20, 11 2019.

[Hat02] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge; New York, 2002.

[HSK+23] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023.

[IH23] Yusuke Imoto and Yasuaki Hiraoka. V-mapper: Topological data analysis for high-dimensional data with velocity. *Nonlinear Theory and Its Applications, IEICE*, 14(2):92–105, 2023.

[KDS+23] Manik Kuchroo, Marcello DiStasio, Eric Song, Eda Calapkulu, Le Zhang, Maryam Ige, Amar H. Sheth, Abdelilah Majdoubi, Madhvi Menon, Alexander Tong, and et al. Single-cell analysis reveals inflammatory interactions driving macular degeneration. *Nature Communications*, 14(1), May 2023.

[KHC+20] Wouter R. Karthaus, Matan Hofree, Danielle Choi, Eliot L. Linton, Mesruh Turkekul, Alborz Bejnood, Brett Carver, Anuradha Gopalan, Wassim Abida, Vincent Laudone, Moshe Biton, Ojasvi Chaudhary, Tianhao Xu, Ignas Masilionis, Katia Manova, Linas Mazutis, Dana Pe'er, Aviv Regev, and Charles L. Sawyers. Regenerative potential of prostate luminal cells revealed by single-cell analysis. *Science*, 368(6490):497–505, 2020.

[LBT+15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, December 2015.

[LLT+24] Alex R. Lederer, Maxine Leonardi, Lorenzo Talamanca, Daniil M. Bobrovskiy, Antonio Herrera, Colas Droin, Irina Khven, Hugo J. F. Carvalho, Alessandro Valente, Albert Dominguez Mantes, Pau Mulet Arabí, Luca Pinello, Felix Naef, and Gioele La Manno. Statistical inference with a manifold-constrained rna velocity model uncovers cell cycle speed modulations. *Nature Methods*, October 2024.

[LMSZ+18] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, and et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, Aug 2018.

[LWHC20] Shaoheng Liang, Fang Wang, Jincheng Han, and Ken Chen. Latent periodic process inference from single-cell rna-seq data. *Nature Communications*, 11(1), 2020.

[PE00] Johan Paulsson and Måns Ehrenberg. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review Letters*, 84(23):5447–5450, June 2000.

[Per20] Jose A. Perea. Sparse circular coordinates via principal $\mathbb{Z}$-bundles. *Topological Data Analysis*, page 435–458, 2020.

[PHY+21] Nicholas E Phillips, Alice Hugues, Jake Yeung, Eric Durandau, Damien Nicolas, and Felix Naef. The circadian oscillator analysed at the single-transcript level. *Molecular Systems Biology*, 17(3), 2021.

[PP23] Taejin Paik and Jaemin Park. Circular coordinates for density-robust analysis, 2023.

[RLF+11] Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, Ido Amit, and Aviv Regev. Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5):436–442, April 2011.

[RMR+20] Raúl Rabadán, Yamina Mohamedi, Udi Rubin, Tim Chu, Adam N. Alghalith, Oliver Elliott, Luis Arnés, Santiago Cal, Álvaro J. Obaya, Arnold J. Levine, and et al. Identification of relevant genetic alterations in cancer using topological data analysis. *Nature Communications*, 11(1), Jul 2020.

[ROD+22] Andrea Riba, Attila Oravecz, Matej Durik, Sara Jiménez, Violaine Alunni, Marie Cerciat, Matthieu Jung, Céline Keime, William M. Keyes, and Nacho Molina. Cell cycle gene regulation dynamics revealed by rna velocity and deep-learning. *Nature Communications*, 13(1), May 2022.

[SDB+23] Bernadette J. Stolz, Jagdeep Dhesi, Joshua A. Bull, Heather A. Harrington, Helen M. Byrne, and Iris H. R. Yoon. Relational persistent homology for multispecies data with application to the tumor microenvironment, 2023.

[SFJ+20] Daniel Schwabe, Sara Formichetti, Jan Philipp Junker, Martin Falcke, and Nikolaus Rajewsky. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, 16(11):e9946, 2020.

[SMG+11] David M. Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, April 2011.

[STM+05] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[TGN23] Lorenzo Talamanca, Cédric Gobet, and Felix Naef. Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. *Science*, 379(6631):478–483, 2023.

[vDSN+18] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, and et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3), 2018.

[Wad57] C.H. Waddington. *The Strategy of the Genes*. Routledge, April 1957.

[Web85] Cary Webb. Decomposition of graded modules. *Proceedings of the American Mathematical Society*, 94(4):565–571, 1985.

[WZXW11] Jin Wang, Kun Zhang, Li Xu, and Erkang Wang. Quantifying the waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, 108(20):8257–8262, 2011.

[ZC04] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Proceedings of the twentieth annual symposium on Computational geometry*, 2004.

5 Funding acknowledgements

K.M., M.K.Y and J.I. were supported by the European Union's Horizon 2020 Research and Innovation Program under Marie Skłodowska-Curie Grant Agreement No 859860. W.K was supported by Swiss Cancer League grant number KLS-5654-08-2022 and Swiss National Science Foundation, project number: 310030-215517.

A Experiment and preprocessing details

A.1 Synthetic datasets

Transcriptomic switch model. In Figure 8 we present the parameters for the transient (T) and recurrent (R) examples (Figure 3 and Figure 9) of the synthetic transcriptomic switch model. Parameters were manually selected to lead to approximate ODE solutions that were sufficiently rich. Sensitivity analysis (Figure 9D) parameters (S) of the recurrent example were adjusted as a function of the number of cells and genes. Ground-truth cell times were selected by uniformly dividing the time interval, so that the gene expression of a cell at time t corresponds to the ODE solution at time t.

Negative binomial sampling. The negative binomial distribution $\mathrm{NegBinom}(n, p)$ has probability mass function

$$P(k) = \binom{k+n-1}{n-1} p^n (1-p)^k,$$

where k is the number of failures over n trials with probability p of success. The parameters are related to the mean μ and variance σ^2 by the formulae

$$p = \frac{n}{n+\mu} \qquad \text{and} \qquad \sigma^2 = \mu + \frac{\mu^2}{n}.$$

The approximated solutions of mRNA counts $(\mathbf{r}_i(t_j))$ in Figure 8 over discrete time intervals $t_0 < t_1 < \ldots < t_n$ were exponentiated, and used as the mean in a negative binomial sampling with a constant dispersion parameter $d = 1/n$. This generated a random count matrix

$$\mathbf{X}_{i,j} \sim \mathrm{NegBinom}\left(\frac{n}{n + \mathbf{r}_i(t_j)}, 1/d\right) \tag{22}$$

that converges to our solved system in mean. Data were sampled from this using the python function `scipy.stats.nbinom.rvs`. Examples of the sampling are given in Figure 3E-G and Figure 9E-G.

Velocycle benchmarking. The pipeline for our model in the *Velocycle* bencharking (Figure 10A and C) consisted of: separating each gene into spliced and unspliced raw UMI counts, filtering genes that occured in fewer than three cells, normalizing the per-cell raw UMI counts, and projection onto the first two PCA coordinates, followed by computing the persistent cohomology circular coordinate of the most persistent feature. The PCA baseline consisted of identical preprocessing, with the circular coordinate taken as the angle between the first two principal components. For each experiment, we ran Velocycle with its suggested preprecossing and parameters for 200 epochs [LLT+24]. The box-plots in Figure 10C were run with the same parameters 20 times with 3000 cells and 300 genes, as per [LLT+24].

	g	c	t	$\mathbf{r}(0)$	$\mathbf{p}(0)$	τ	α	$\mathbf{B}$	$\mathbf{C}$	$\mathbf{D}$
T	80	1000	$[0, 6)$	$\begin{cases} 5 & i \leq 10 \\ 0 & \text{else} \end{cases}$	0.0	10.0	$\sim U(1, 4)$	$\sim U(1, 1.6)$	$2 \cdot I_{80}$	I_{80}
R	40	1000	$[0, 20)$	$\begin{cases} 5 & i \leq 10 \\ 0 & \text{else} \end{cases}$	0.0	3.0	$\sim U(1, 4)$	$\sim U(1, 1.6)$	$2 \cdot I_{40}$	I_{40}
S	g	c	$[0, 0.02c)$	$\begin{cases} 5 & i \leq 0.1g \\ 0 & \text{else} \end{cases}$	0.0	0.02g	$\sim U(1, 4)$	$\sim U(1, 1.6)$	$2 \cdot I_g$	I_g

Figure 8: Hyperparameters in the transcriptomic switch model experiments.

A.2 Cycling dermal fibroblasts

Dataset. The data we used from [CKL+22] is publicly available on the Gene Expression Omnibus database under accession no. GSE167209. We followed the pre-processing procedure in [LLT+24], where the same dataset is used as an example. Similar to our pipeline and the standard Seurat pipeline, this consisted of log normalization and mean centering, however with additional filtering for both spliced and unspliced genes.

A.3 Prostate regeneration data

Dataset. Original data profiles were from 13,398 cells of the mouse prostate, focusing primarily on the anterior lobe, using droplet-based scRNA-seq. For recovering cell type classifications, we used the author's data and annotations for [KHC+20] available at Gene Expression Omnibus `https://www.ncbi.nlm.nih.gov/geo/` (accession no. GSE146811). We restricted the data to the largest cluster of luminal epithelial cells (L1), characterized by the genes *CD24a, Krt8*, and *Krt18*.

Pre-processing and filtering. We first restricted to the subset of the Luminal 1 (L1) subpopulation, then filtered by cells expressing at least 200 genes and genes expressed in at least three cells. Raw UMI counts were normalized per cell, log-transformed, and restricted to the subset of highly variable genes using `scanpy.tl.highly-variable`. MAGIC [vDSN+18] was applied to the data with nearest-neighbour parameter `knn` $= 4$ for one step ($t = 1$). Finally, we removed 3000 outliers with low density, according to the density function

$$\mu(x) = \Big(\sum_i d(x, x_i) \Big)^{-1},$$

where d was the metric in the first five principal components. Finally, we performed maxmin sampling of the data to reduce it to a set of 2^{10} landmark cells.

Circular enrichment. For the circular gene set enrichment, we used the 2020 hallmark gene set from the Molecular Signatures Databast. The global pre-processing was first applied to the luminal cells above. We then projected onto each gene set, followed the first two principal components, and finally computed the circular coordinate of the most persistent feature. The p-values were computed according to the permutation procedure described in Method 4.3.