Topology identifies concurrent cyclic processes in single-cell transcriptomics and androgen receptor function

Kelly Maggs*[1,2,3], Markus K. Youssef*[4,5], Cyril Pulver[5], Jovan Isma[8,10], Tâm J. Nguyên[5,6], Matthias Arzt[1], Wouter Karthaus[5,7], Heather A. Harrington[1,2,3], 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, École 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]International Cancer Prevention Institute, Epalinges, Switzerland
[10]Department of Immunobiology, University of Lausanne (UNIL), Switzerland.

December 6, 2025

Abstract

Standard single-cell RNA-seq analysis frameworks aggregate overlapping biological processes and impose a single parametrization, conflating distinct programs. Here, we introduce a topological framework that detects and disentangles multiple cyclic processes directly from single-cell transcriptomic data. We validate this approach on synthetic datasets and scRNA-seq profiles of human dermal fibroblasts under control conditions and following androgen receptor (AR) silencing, as well as in vivo mouse prostate regeneration under androgen receptor add-back. We show robust cell cycle structure across conditions, identify an unbiased AR-linked stress signature related to the senescence and proliferation across organisms, and uncover cholesterol homeostasis as an AR-linked program in tissue regeneration. This framework enables identification and separation of concurrent cyclic processes from snapshot single-cell data, revealing complex multi-dimensional regulatory dynamics inaccessible to standard clustering analysis.

*Shared first-authorship. Co-correspondence: maggs@mpi-cbg.de, paolo.dotto@unil.ch.

Single-cell RNA-seq gives a high-dimensional point cloud that is often analyzed as an approximation of the Waddington landscape. The long-held promise of single-cell sequencing has been to uncover both 'intercellular heterogeneity and subtle quantitative processes' [1]. The standard single-cell analysis pipeline of clustering over low-dimensional projections is remarkably successful at identifying cellular heterogeneity. However, the distillation of subtle biological and regulatory processes remains a subject of intense inquiry and debate [2].

Waddington's classical epigenetic landscape [3] is a common starting point for mathematical and computational models of development [4–6]. In the landscape model, cells roll down branching valleys carved by underlying regulatory programs. However, such a model lacks a framework for the cyclic processes, such as cell-cycle oscillations, tissue regeneration, and metabolic processes, including stress-response circuits. Additionally, the current model fails to capture that these processes occur concurrently. An updated model would therefore be a sophisticated, high-dimensional mathematical structure built from multiple processes—some one-way (irreversible) and others that are recurrent (cyclic). An unsolved problem in the field is how to identify and separate these different types of processes.

Existing methods in computational single-cell sequencing analysis reflect these conceptual issues in their mathematical modeling assumptions. The vast majority of tools assume that dynamic processes are modeled as 1-dimensional structures such as trees and paths, with some more recent models assuming that the data lies on a circle [7–13]. Projecting onto a single parameter can miss or conflate these concurrent circular programs. While more complicated mathematical models have been proposed in theory [14, 15], the data-driven formalism to describe how multiple processes coalesce to form higher order structure has yet to be laid out in a principled manner.

Here, we rely on recent advances in computational topology to develop a principled mathematical approach to identify distinct cyclic biological processes and study their higher-order interactions. Recent applications of topological data analysis in single-cell studies further support the promise of this topological/geometric perspective [16–21] – particularly the use of cohomology-based circular coordinates in concurrent recent work [22, 23]. Our topological approach relies on persistent cohomology as a mathematical signature of recurrent biological processes. These measurements come equipped with stability theorems that guarantee robustness – a useful feature when working with noisy transcriptomic data. Based on the integration of differential forms and harmonic cocycle [24–27], we introduce new mathematical quantities to measure multi-dimensional recurrent dynamics within the Waddington landscape. We develop a geometric language for analyzing identified topological features to recover pseudo-temporal orderings of both cells and gene expression, allowing an in-depth biological interpretation of the discovered structure.

We first validate that topological and geometric invariants can robustly detect and analyze biological phenomena to yield reliable insight. Using the cell cycle as an example, we introduce a topological-geometric framework to recover the cyclic ordering of cells and dynamic ordering of gene expression consistent with known literature. We show that our predicted ordering is robust across both cell type and organism. Furthermore, this topological approach shows that key cell cycle genes in proliferating human dermal fibroblasts function normally, regardless of whether the androgen receptor is present.

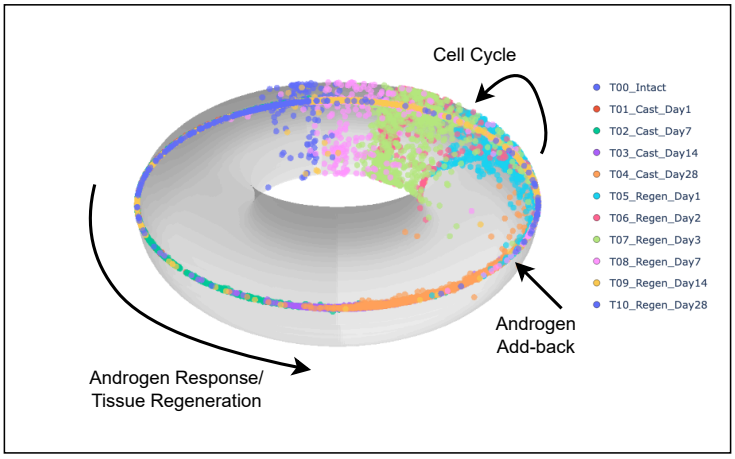


Figure 1: **Multi-temporal processes in AR-modulated prostate regeneration**. Topological re-analysis of androgen-receptor modulated prostatic regeneration data from [29]. Upon androgen add-back, regeneration and cell cycle span *multiple circular temporal dimensions* in single-cell RNA sequencing data, suggesting *temporal degrees of freedom* in cellular decision making.

We explicitly demonstrate the existence of multiple concurrent cyclic processes in both synthetic models and real-world transcriptomic data. Specifically, we focus on transcriptional circuits interlocking cell-cycle regulation with androgen receptor (AR)–dependent control of gene expression, where AR functions as a master regulator of developmental and cancer-related transcriptional programs [28]. Topological analysis of androgen modulated prostate regeneration data from [29] reveals that tissue regeneration and cell cycle generate separate circular dimensions Figure 1. The existence of such *multiple recurrent temporal dimensions* challenges the current paradigm of modeling dynamics from single-cell RNA-seq data.

Our topological approach shows that key cell-cycle genes in proliferating human dermal fibroblasts remain robustly and coherently regulated regardless of AR status, even though AR acts in these cells as a repressor of senescence-associated, protumorigenic activation [30–32]. We derive an AR stress signature that comprises of upregulation of senescence and pro-inflammatory genes with protumorigenic potential. Cells with AR loss heterogenously exhibit such a signature that is reduced along progression through the cell cycle topology. We further extend our analysis to an in vivo mouse prostate regeneration model under AR control, revealing a consistent interlocking of AR stress signature and topological cell cycle dynamics.

Finally, we apply our cohomology-based circular gene set enrichment to topologically identify cholesterol homeostasis and androgen response as key interlocking sub-processes of tissue regeneration in mouse prostatic epithelial cells. Together, these results establish the broad applicability of our topological framework for identifying and separating complex cyclic biological programs across organisms, tissue contexts, and experimental modalities.

Results

Robust topological pipeline for identifying statistically significant circular biological processes

In this work, we introduce *Cocycle Hunter* (`cHunter`) (Figure 2) as a topological framework for single-cell analysis. Circular geometries in high-dimensional single-cell RNA-seq data involve the coordinated behavior of many genes. We extend gene set enrichment, a widely used method in transcriptomic analysis [33], to *circular* gene set enrichment (Figure 2A), which tests whether projecting single-cell data onto a specific gene set produces a statistically significant cyclic pattern. We focus on *gene set projections* by restricting our analysis to groups of biologically relevant genes (e.g., those from the Molecular Signatures Database [34]). This strategy facilitates the detection of recurrent sub-processes and provides a clear basis for downstream biological interpretation.

To quantify circularity, we apply a robust mathematical technique, persistent cohomology [35], to automatically detect circular structures within gene set projections and evaluate their significance using a summary statistic, the *ring score* ((Figure 2B). Upon detection of such a significant circular structure, we assign each cell a phase on the inferred cycle, thereby providing an ordering that captures pseudo-dynamics through circular coordinates defining a harmonic signal. These quantities are computed across pairwise gene projections to construct a *lead-lag matrix* (Figure 2F), whose entries encode directional relationships between pairs of genes via integrals describing their orientation in certain subspaces (Figure 2E). Positive and negative values of the ij-th entry respectively model expression of gene i peaking before or after gene j along the circular structure. Spectral analysis of this matrix (Figure 2F) yields an inferred temporal ordering of gene activation, with the amplitude and phase of complex spectral components furnishing quantitative estimates of each gene's relative contribution and peak activity along the harmonic cycle.

Topological approach excels at concurrent process analysis

In single-cell RNA-seq data, gene sets may contain more than one dynamic transcriptomic process (Figure 2B). For example, the cell cycle can co-occur with an independent signaling process that induces transcription. In such cases, state-of-the-art methods [8] for estimating circular processes are not designed to disentangle circular dynamics from orthogonal processes. In order to test this hypothesis, we designed a synthetic test dataset (Appendix A.2) consisting of uncoupled circular and linear processes (Figure 3).

The topological approach taken by `cHunter` was able to reliably and robustly (Figure 3B) decouple these processes. We compared `cHunter` with a state-of-the-art method `velocycle` [8] for analyzing circular structures in single-cell data. Across various parameter settings (Figure 3E), `cHunter` consistently outperformed `velocycle` and the PCA baseline, both of which tended to conflate the circular and monotonic processes (Figure 3B). Notably, `cHunter` uses persistent cohomology to separate out circular and monotonic dynamics based on the topology of the data, whereas `velocycle` relies on a fixed single cycle prior that can conflate recurrent dynamics with other processes.

We additionally compared the estimation of circular coordinates against another, more traditional synthetic model presented in `velocycle` [8], which simulates a single

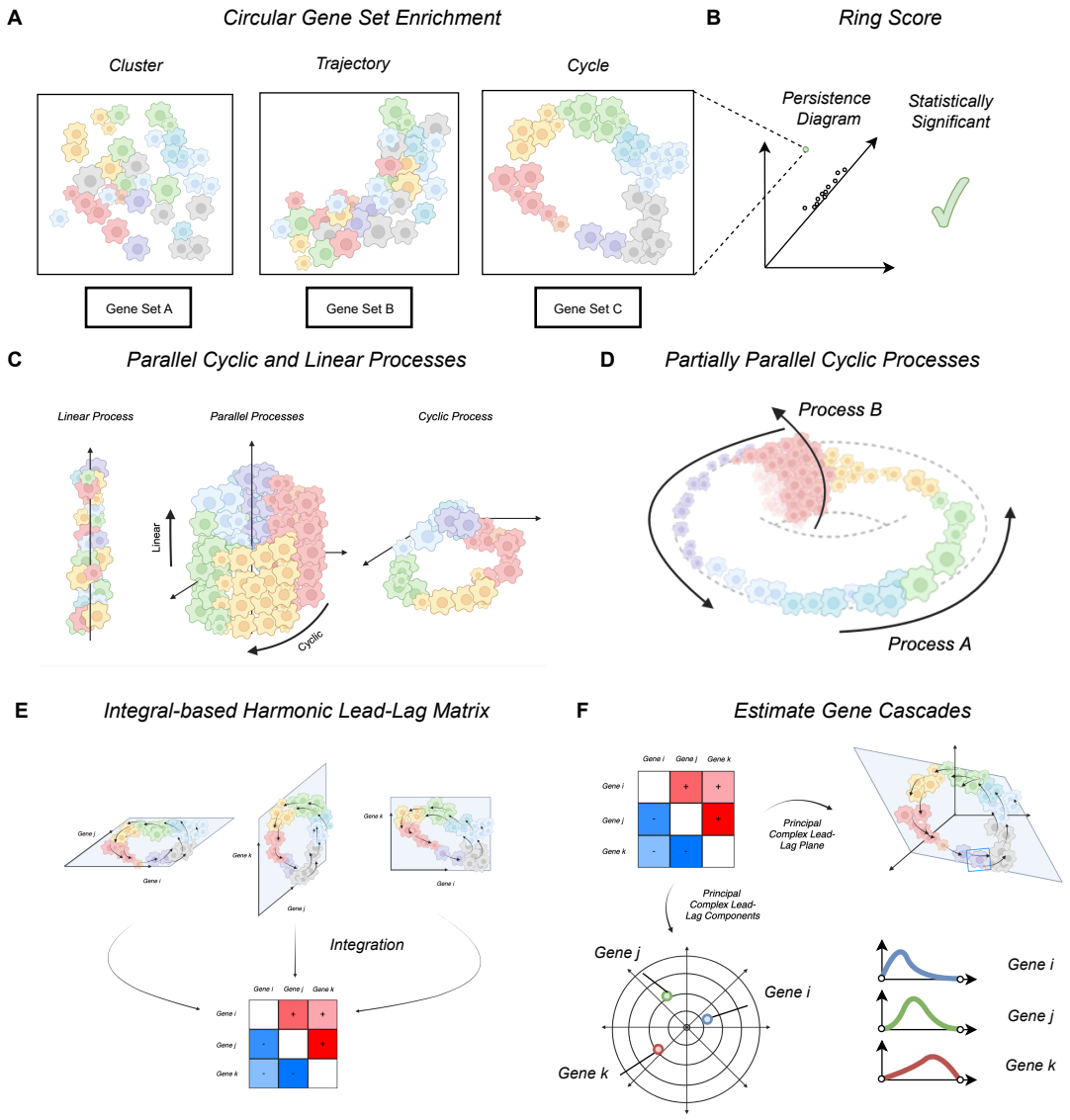


Figure 2: **Conceptual overview of `cHunter` topological and geometric pipeline.** **(A)** Circular gene set enrichment tests whether the projection of a single-cell experiment onto a given gene set reveals a cyclic pattern. **(B)** Persistent cohomology is used to quantify circular geometry and statistical significance. **(C)** Topological model of parallel uncoupled linear and circular processes as forming a cylindrical product space. **(D)** Topological model of partially parallel cyclic processes as embedded on the torus. **(E)** The data is further analyzed by projecting onto pairwise gene planes and integrating to compute a lead-lag matrix. **(F)** Finally, eigenvector analysis of this matrix estimates the order of gene activation, revealing gene cascade phases associated with the cycle.

cyclic process (see Appendix A.2). Sensitivity Figure S3A and robustness analysis Figure S3C over the parameters in [8] showed that `cHunter`, `velocycle` and the PCA baseline produced roughly equivalent circular coordinate predictions with the ground-truth (Figure S3A,C,D). Given that the circular structure was clearly visible in the first two principal components (Figure S3B), our analysis justified the need for more challenging settings to distinguish the approaches.

Synthetic model shows how topology emerges from transcriptional negative feedback loops

To lay the theoretical groundwork for how topological structure can emerge from RNA regulation, we introduced a synthetic model of coordinated negative feedback loops in RNA transcription (Appendix A.1). In our model, RNA transcription was viewed as a bursty, switch-like process in which genes toggled between inactive and active states based on global gene expression, reflecting the intermittent nature of transcription observed in cells [36, 37]. Rather than assuming continuous gene expression, our approach introduced a threshold mechanism that triggered transcription only when specific conditions were met, thereby capturing the finite and episodic production of RNA. To mimic the inherent variability of single-cell data, we incorporated a sampling process that simulated the random fluctuations observed in real experiments [38–41]. When incorporating negative feedback loops into the interactions between RNA and proteins, our model naturally gave rise to cyclic patterns that were observable in low-dimensional projections (see Figure S2D). Moreover, the geometric lead-lag analysis of such cycles was able to faithfully recover the underlying ground-truth leader-follower relationships between genes S2D).

Topology and lead-lag analysis recovers known cell cycle dynamics of human fibroblasts in culture

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 fibroblasts [42] (Appendix A.3), which has been used previously as a validation dataset on cell-cycle tools [8]. We projected the data onto the gene set spanned by known S and G2M markers (we use the markers in [8]). 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 was assigned based on known S and G2M orderings, and cell cycle start point was assigned minimize total counts – both choices are independent of lead-lag coefficients. Further, the logarithm of the total number of molecules was consistent with passage of the cell cycle through the S phase into mitosis.

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 [8]. The human fibroblasts dataset contains the additional information of spliced and

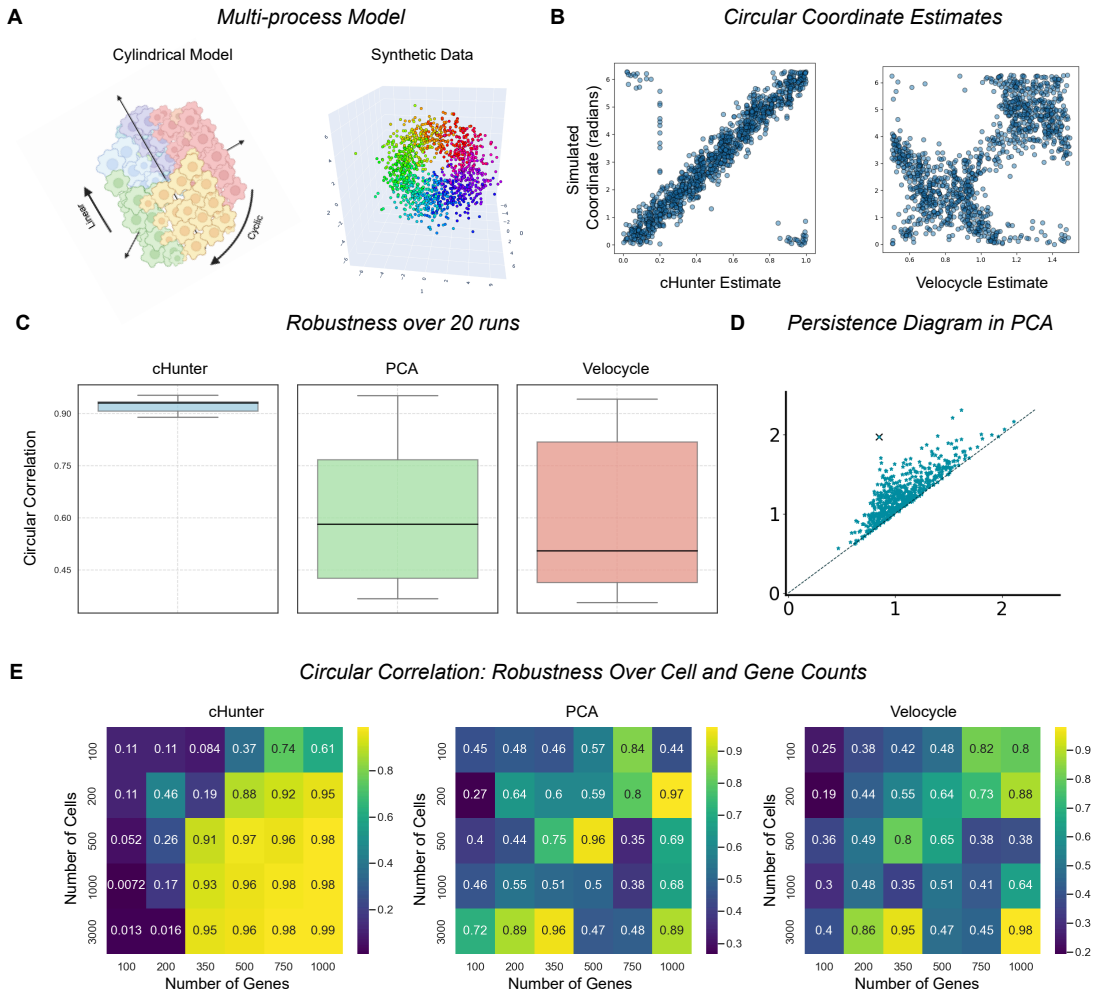


Figure 3: **Topological methods disentangle synthetic parallel processes.** **(A)** Conceptual cylindrical model and 3D PCA of the multi-process benchmark with 1200 simulated cells and 300 genes associated to each process. **(B)** Predicted circular coordinates by cHunter and Velocycle. Velocycle confuses the linear and circular processes while cHunter disentangles them. **(C)** Box-plot of ground-truth circular correlation of 20 instances of the multi-process synthetic data set with simulated 1200 cells, 300 circular-process genes, and 300 linear-process genes with cHunter, PCA baseline and Velocyle circular coordinate estimates. **(D)** Example persistence diagram of PCA 0,1,2 of such an instance. **(E)** Circular correlations with ground truth over varying cell and gene counts in the multi-process synthetic benchmark.

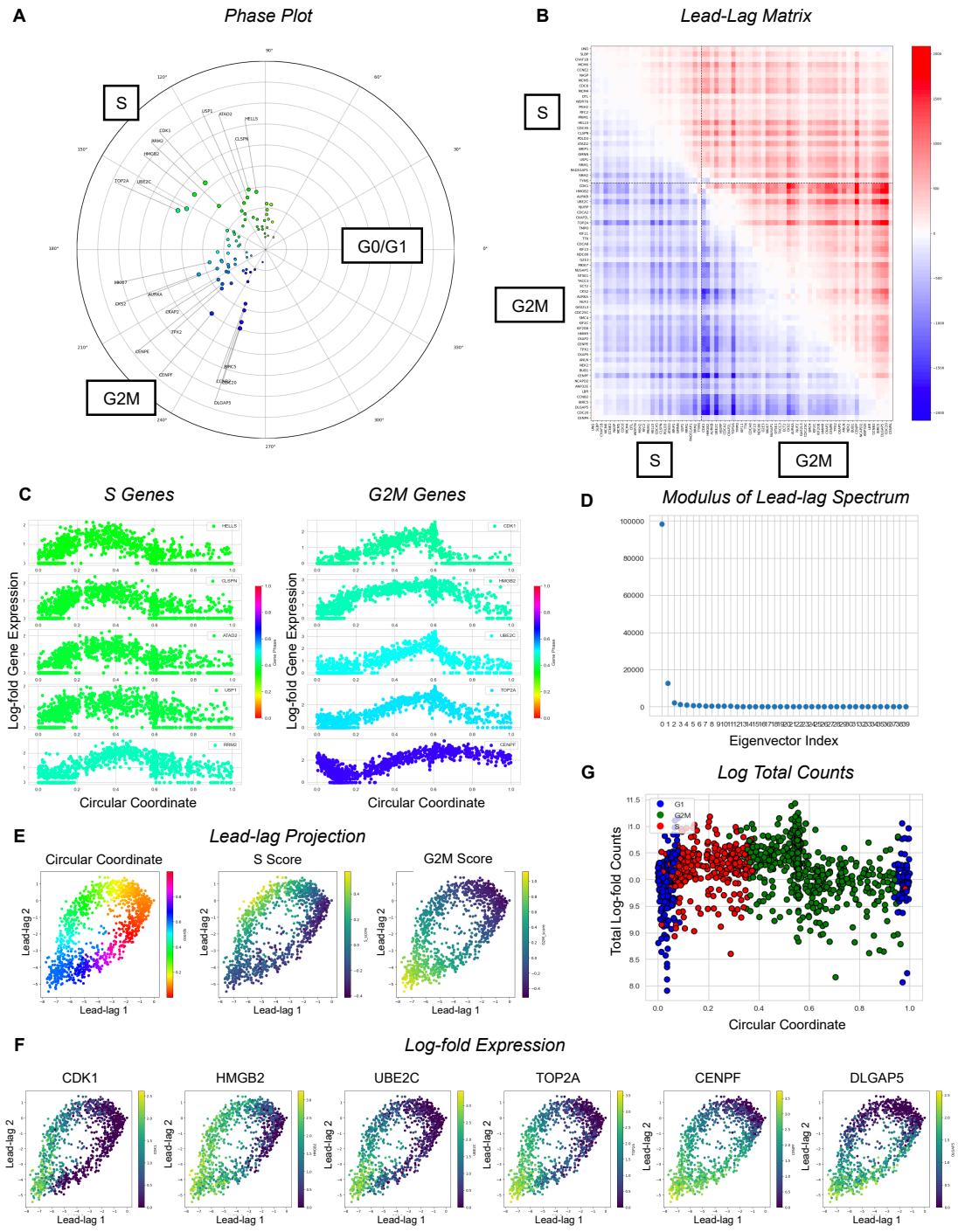


Figure 4: **Topological methods recover known cell cycle dynamics.** **(A)** Estimated gene phases from harmonic lead-lag analysis. **(B)** The lead-lag matrix, sorted by estimated gene phase and groups by S and $G2M$ genes. **(C)** Normalized log-fold expression of oscillatory genes in the S and G2M phases plotted against the estimated circular coordinates. **(D)** Sorted absolute values of the complex eigenvalues of the lead-lag matrix. **(E)** The linear lead-lag projection onto the principal complex eigenplane of the lead-lag matrix colored by circular coordinates, `scanpy`-derived S- and G2M-scores. **(F)** The linear lead-lag projection onto the principal complex eigenplane of the lead-lag matrix colored by oscillatory gene expression in order of estimated phase around the cell cycle. **(G)** The log of the total reads per cell (y-axis) against the circular coordinates (x-axis).

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 S4A). The lead-lag matrix showed that unspliced RNA levels led spliced RNA levels of corresponding genes as expected (Figure S4A). Further, the estimated phases of unspliced molecules proceeded those of their corresponding spliced isoforms broadly across the gene set (Figure S4B,D).

Topology shows that S and G2M cell cycle dynamics are robust under AR silencing

Existing literature suggests that the androgen receptor (AR) plays a crucial role in regulating human dermal fibroblasts (HDF) homeostasis and can influence early steps of cancer-associated fibroblast (CAF) activation. In this setting, AR can act as a key negative regulator of cellular senescence–effector genes (like CDKN1A) and of genes comprising the Senescence-Associated Secretory Phenotype (SASP) [43, 44], which includes proinflammatory cytokines and matrix remodeling proteins also found in fully established CAFs [30]. Having attained a baseline of normal cell cycle functioning in HDFs, we set out to investigate whether AR-silencing in these cells produces biologically meaningful topological and/or geometric perturbations.

Our novel single-cell RNA-seq dataset consisted of human dermal fibroblasts (HDFs) infected with two lentiviruses either expressing different AR-silencing shRNAs versus a control expressing scrambled shRNAs (Appendix A). We first aimed to compare the findings from the previous case study on fibroblast cell cycle to our second multi-conditioned (control and AR-silenced) HDF dataset. To replicate the approach, we used the same S and G2M marker genes as the previous section, all 98 of which were identified as highly variable. Lead-lag analysis on a circular structure identified in the first three principal components of this gene set revealed a striking correlation in estimated gene phases ($R \approx 0.96$) and amplitudes ($R \approx 0.84$) across this dataset and that of Capolupo et. al (77 shared, highly variable genes) (Fig. 5B).

We then set about comparing the key cell cycle genes in our gene set across the two conditions within these identified cycling populations. We devised a method to re-fit normalized gene expression as a first-order sine wave over the circular coordinate using multilinear regression over the lead-lag plane. By applying this pipeline to shAR and control populations, we were able to attain separate estimates of the first-order Fourier coefficient for each cell cycle gene across conditions. We found a striking correlation between estimates of both first-order amplitude ($R \approx 0.98$) and phase ($R \approx 0.99$) (Fig. 5C). This suggests AR-silenced HDFs that progress through cell cycle, the core transcriptional oscillations of these key S and G2M genes are largely maintained (Fig. 5C,D). We concluded that the primary consequence of AR-silencing stress appears not to be a gross disruption of the cycle's mechanics in still-proliferating cells, but rather an exit from the cycle in a distinct subpopulation, as explored next.

AR stress signature decreases along S-phase of cell cycle for human dermal fibroblasts

Continuing on with the dataset consisting of AR-silenced vs control HDFs, we found that the first principal component (PC zero) captured a continuous phenotypic transition from control to shAR HDFs 5A. We termed this PC zero axis the "AR-silencing stress signature." Crucially, when we examined this continuous signature in relation to cell cycle status, we observed that cells with high PC zero scores (indicating high AR-silencing stress) were almost exclusively found outside of the S and G2M phases, consistent with an exit from active proliferation 5D. This observation reconciles the apparently overall cell cycle distributions in bulk populations with a potent cell cycle exit occurring specifically within the high-stress subpopulation.

Examining the PC zero factor loadings provided molecular insight into this stress signature. Genes with strong positive loadings, indicative of upregulation in the high-stress/arrested state, included compelling markers of cellular senescence and the SASP. Notably, TGFBI (Transforming Growth Factor Beta Induced), a protein robustly linked to senescence and induced by TGF-β signaling [45], was the 5th highest-ranked contributor to this signature among the top 2000 highly variable genes, with a strong positive loading of 0.147. Further supporting this was TGFBR1, a receptor involved in TGF-β signaling and CAF marker [46] (rank 144, loading 0.020). The key cell cycle inhibitor and senescence marker CDKN1A (p21) [47] also exhibited a positive loading of 0.029 (rank 88). Finally, IL6, a canonical SASP cytokine that can relay and sustain senescence [44], was also highly ranked (rank 210, loading 0.014). The coordinated upregulation of these genes strongly suggests that the AR-silencing stress signature identifies HDFs undergoing G0 arrest in a state highly characteristic of cellular senescence.

Circular enrichment identifies cholesterol homeostasis as cyclic sub-process of AR-regulated tissue regeneration ab initio

We next evaluated whether our analysis of cell cycle topology and AR function produces similar results in other settings, including (1) in vivo, (2) mouse models, and (3) across multiple scRNA-seq samples. To this end, we examined a publicly available dataset [29] where the authors study the role of androgen in prostatic epithelial cell regeneration in mice (Details A.5). In the study, mice are castrated, and within 28 days, the prostate involutes to 90% of its size. At day 28, mice are transplanted subcutaneously with dihydrotestosterone (DHT) to initiate androgen-mediated regeneration. After 28 days of regeneration, the prostate returns to homeostasis, notably characterized by quiescence and expression of the key prostate transcription factor NKX3.1. Multiple mice received the treatment concurrently, with single-cell samples of their respective prostrates taken over the full course of castration and regeneration.

Our first question was whether circular gene set enrichment was able to identify relevant biological processes. After pre-processing (See methods), we performed circular gene set enrichment using our cohomology-based ring score, testing over the 50 hallmark sets of genes [33]. We found that androgen response was the top circularly enriched subspace of statistical significance under projection onto both the first three and first two principal components (Figure 6E), validating our hypothesis that gene expression associated with androgen response would generate a closed biological loop in the single-

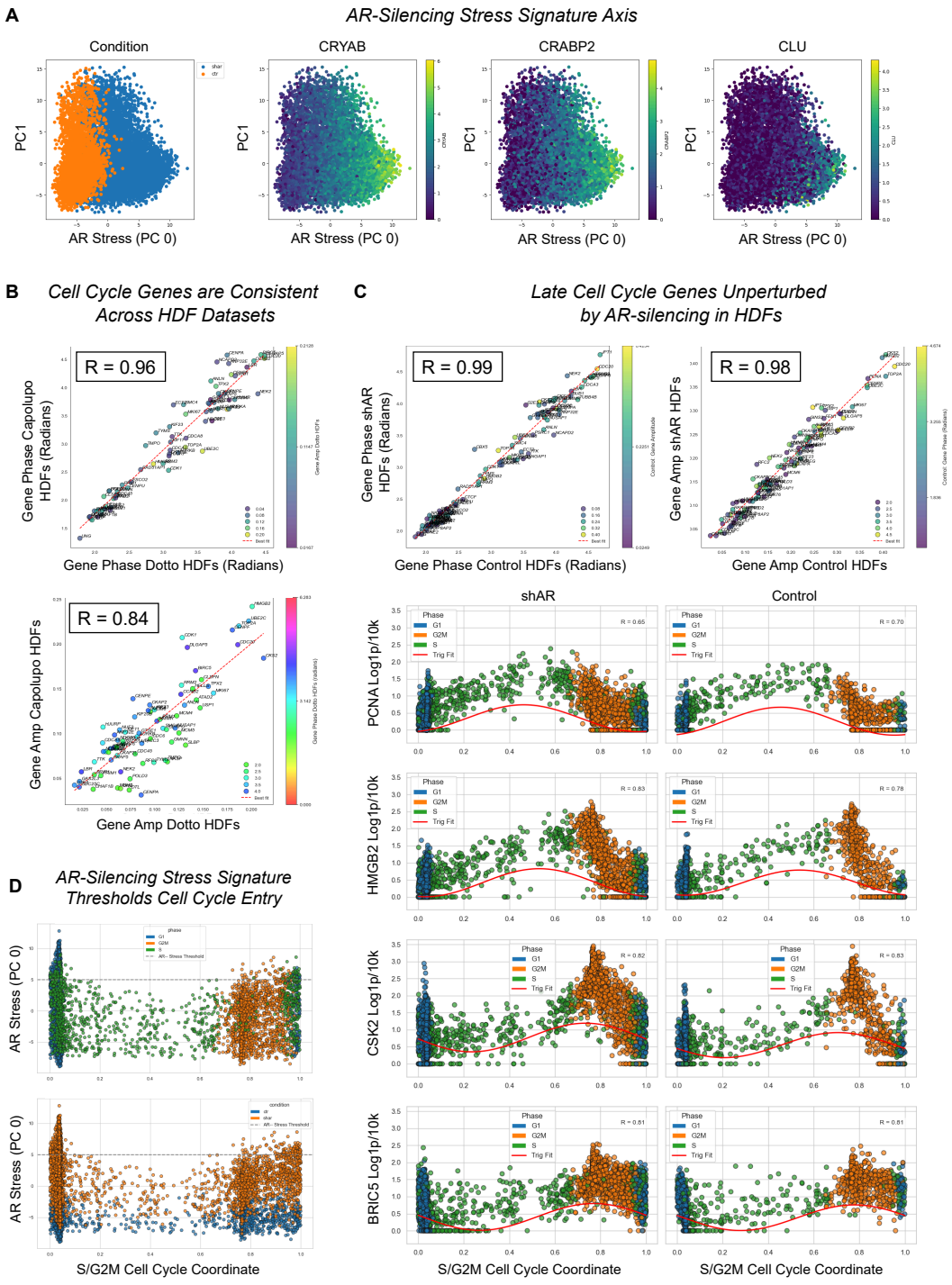
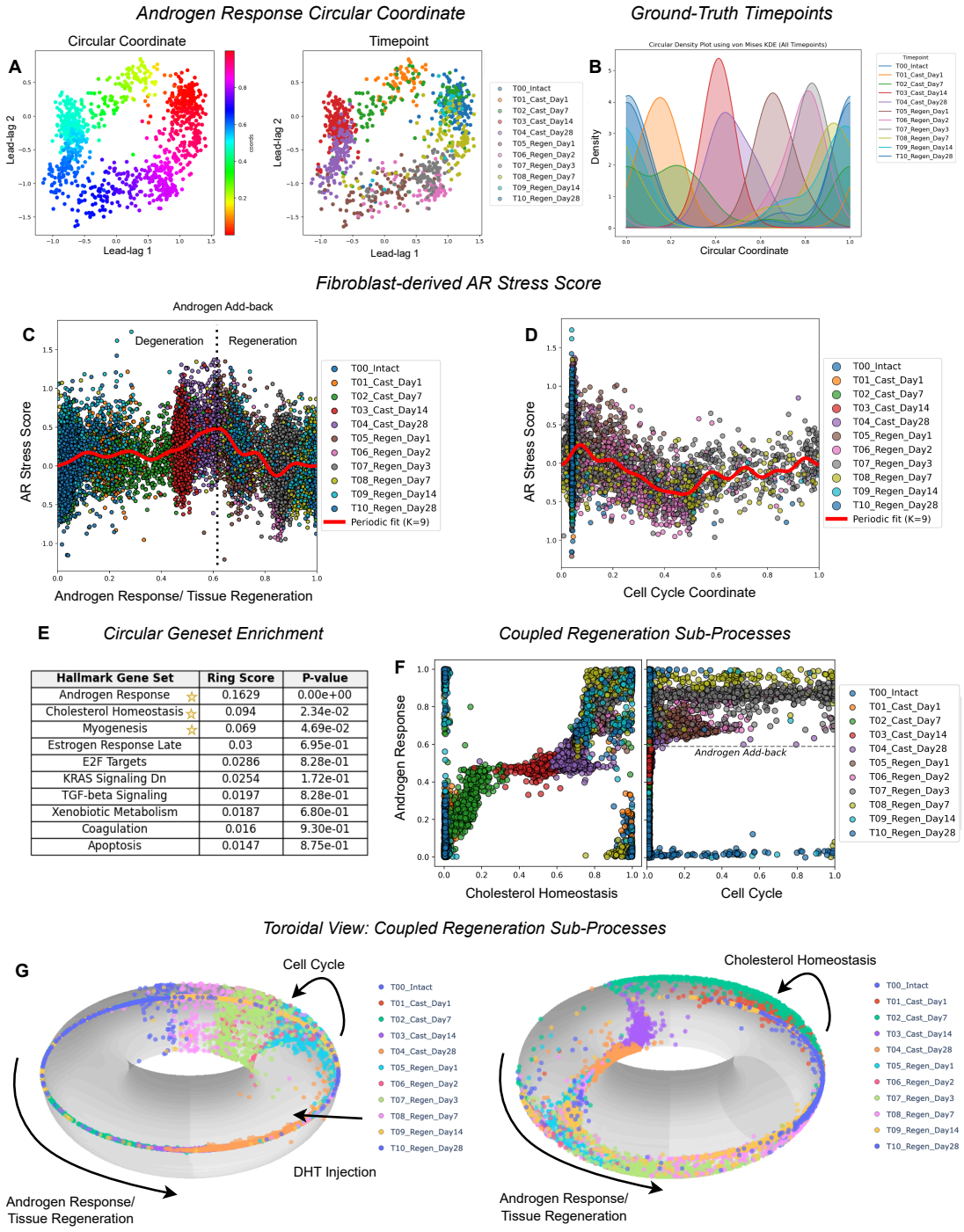


Figure 5: **Key cell cycle genes are robust under AR silencing, but AR stress can threshold cell cycle entry.** **(A)** PCA plot of first two principal components over highly-variable genes in control and AR-silenced HDFs. PC zero is taken as a continuous signature parameter indexing AR-silencing stress response. **(B)** Comparison of cHunter first-order gene-wise amplitude and phase estimates across HDFs in Dotto and Capolupo et al. datasets. **(C)** Comparison of cHunter first-order gene-wise amplitude and phase estimates across AR-silenced vs control HDFs in Dotto dataset. The expression of cell cycle genes against cell cycle coordinate in AR-silenced vs control HDFs. **(D)** AR-silencing stress response signature against cell cycle circular coordinates.



Hallmark Gene Set	Ring Score	P-value
Androgen Response ☆	0.1629	0.00e+00
Cholesterol Homeostasis ☆	0.094	2.34e-02
Myogenesis ☆	0.069	4.69e-02
Estrogen Response Late	0.03	6.95e-01
E2F Targets	0.0286	8.28e-01
KRAS Signaling Dn	0.0254	1.72e-01
TGF-beta Signaling	0.0197	8.28e-01
Xenobiotic Metabolism	0.0187	6.80e-01
Coagulation	0.016	9.30e-01
Apoptosis	0.0147	8.75e-01

Figure 6: **Coupling of cyclic tissue regeneration sub-processes has a topological description.** **(A)** The lead-lag projection colored by androgen response circular coordinates and sample time, respectively. **(B)** The Von-Mises-based kernel density estimate of each time-point against the androgen response circular coordinate. **(C)** The androgen response coordinate against the AR stress score derived from the fibroblast data. **(D)** The cell cycle coordinate against the AR stress score derived from the fibroblast data. **(E)** The top results of circular gene set enrichment by ring score – gene sets with a p-value < 0.05 are starred. **(F)** Coupling of tissue regeneration sub-processes circular coordinates: androgen response vs cholesterol homeostasis (left) / cell cycle (right). **(G)** Toroidal view of coupling of tissue regeneration sub-processes circular coordinates: androgen response vs cell cycle (left)/cholesterol homeostasis (right).

cell phase subspace. Circular enrichment identified two other hallmark gene sets – cholesterol homeostasis and myogenesis – as ab initio gene sets with circular structure.

The experiment provided a ground-truth labeling of samples ordered by time in the underlying regeneration 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. Without using explicit knowledge the timepoints in the computation, the estimated circular coordinate associated to androgen response gene set was coherent with the ground truth ordering of the samples (Figure 6B). We concluded that the circular structure found in the androgen-response gene set was an effective continuous parametrization of the tissue regeneration process, and highlights that temporal processes can be estimated directly from the geometry of the data.

We then performed the associated lead-lag analysis (Figure S6A-C). We noted that several androgen-regulated genes in the matrix have key roles in prostate development and maintenance. Notably, NKX3.1, a transcription factor known to be essential for prostate development and maintenance of luminal cell [48] identity, was estimated to peak at homeostasis time-points before castration and after full regeneration. 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 gene gets 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.

The experiment provided a ground-truth labeling of samples ordered by time in the underlying regeneration 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 androgen response circular coordinate was coherent with the ground truth ordering of the samples (Figure 6B), despite being calculated independently of this information. We concluded that the circular structure found in the androgen-response gene set was an effective continuous parametrization of the tissue regeneration process.

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 ring score (Figure 6E). Plotting the circular coordinates against each other (Figure 6F,G) showed that both circular coordinates were dependent on each other, which highlights that these biological sub-processes associated with the master tissue regeneration cycle are strongly coupled. 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 S6G.

Human fibroblast-derived AR-silencing stress signature functions consistently in regenerating mouse epithelia

Finally, we sought to compare whether the transcriptional AR-deprivation response was governed by comparable genes across HDFs and prostatic mouse epithelial cells. We observed that our HDF-derived AR stress signature was maximally valued on prostatic

epithelial cells at the peak of androgen deprivation, 28 days after castration, before gradually decreasing after androgen add-back (Figure 6C and Figure 6K). This suggests that our AR stress signature was measuring a consistent transcriptional response related to AR in two settings: (1) where HDFs are AR-silenced in culture and (2) where mouse prostate epithelial cells are deprived of androgen via castration.

We also wanted to examine the relationship between AR stress signature in the prostate data. Interestingly, cell cycle gene sets were not found to be circularly enriched in the prostate data. We were, however, able to identify a persistent cohomology class corresponding to the cell cycle by restricting to the first 7 days of regeneration and the 28th day after castration, suggesting that the geometry of the tissue regeneration cycle obscures the cell cycle in the full dataset. Analogous to the HDFs, androgen add-back mediated progression through the S-phase of the cell cycle Figure 6F,G was associated with a lower AR stress signature, suggesting a potential common transcription pathway across human and mouse cell-types regulating cell-cycle progression and AR-signaling.

Discussion

Our work introduces a principled topological-geometric framework that detects and disentangles multiple concurrent cyclic processes in single-cell transcriptomic data. By relaxing the traditional one-dimensional manifold assumptions that dominate current single-cell analysis pipelines, we demonstrate that topological and geometric techniques are capable of capturing the multidimensional and recurrent structure of biological dynamics. This approach provides a unified mathematical formalism to recover and interpret cyclic processes that are typically conflated or obscured by standard clustering or trajectory inference methods.

The identification of multiple, concurrent cycles as separate circular dimensions in single-cell data highlights a fundamental limitation of Waddington's classical landscape metaphor. While the landscape model elegantly describes differentiation as a branching and irreversible process, it lacks the formalism to represent concurrent cyclic phenomena that govern many biological systems, such as the cell cycle and tissue regeneration. Our results establish that the transcriptomic manifold of single cells is not simply tree-like, or even circular, but is better described by a richer collection of higher-order topological dimensions—circles, tori, and potentially more complex substructures—that reflect the coexistence of multiple biological clocks. The example in Figure 6G demonstrates a key observation about the Waddington landscape: it is *multi-temporal*, i.e., it exhibits multiple dimensions corresponding to temporal processes.

We first validated our method using synthetic datasets where we designed a cylindrical model of concurrent linear and cyclic processes. In this controlled setting, our framework—implemented as the `cHunter` software package—robustly separated circular from linear dynamics, outperforming state-of-the-art methods that assume a single global cycle. This benchmark confirmed that topological methods generalize beyond prior heuristics by allowing the data itself to define the number and structure of recurrent processes. Importantly, our synthetic modeling of negative transcriptional feedback loops provided mechanistic intuition for how cyclic topology can emerge naturally from gene-regulatory interactions, linking geometric structure to biological causality for scRNA-seq data.

Focusing on primary human dermal fibroblasts, we validated the use of topology and lead-lag analysis to detect circular cell and gene transcription dynamics across the cell cycle, with consistent estimates underscoring the robustness of our approach. The harmonic coordinates and lead–lag matrices recovered coherent transitions from S to G2/M phases, faithfully reproducing known cell-cycle gene relationships. We found that cells responded heterogeneously to silencing of the AR gene, and that AR-stress–related cell-cycle arrest coincided with increased expression of key senescence-related genes, including TGFBI, CDKN1A, and IL6. The relationship between our AR-stress signature and cell-cycle obstruction was observed in multiple settings—in vivo mouse epithelia and in vitro human fibroblasts—suggesting a shared underlying molecular mechanism. These findings are consistent with a continuous phenotypic transformation from normal stromal fibroblasts to cancer-associated fibroblasts (CAFs), a process that may be reversible through epigenetic reprogramming.

Our topological approach of circular gene set enrichment revealed cholesterol homeostasis as a key process in AR-regulated tissue regeneration. This previously unrecognized connection has significant biological implications, including the involvement of enzymes shared between cholesterol and androgen biosynthesis, which could be therapeutically targeted in disease contexts such as prostate cancer or stromal reprogramming. This example demonstrates the viability of circular enrichment as a generic topological technique for the unbiased, de novo identification of gene sets associated with cyclic transcriptional processes. In the in vivo prostate regeneration model, the harmonic ordering of androgen-response and cholesterol-homeostasis gene sets suggested that these processes are interdependent sub-processes of a broader regenerative program, supporting the concept of nested cyclic hierarchies within tissue-level regeneration.

Through these analyses, we demonstrated the applicability of our ab initio topological approach across diverse settings—including synthetic RNA models and benchmarks, in vitro human cells, and heterogeneous in vivo mouse data. We have introduced topological and geometric statistics to estimate the multidimensional geometry of the Waddington landscape directly from data, extending beyond clustering and single-parameter methods. This framework allows practitioners to isolate multiple cyclic processes and study their interactions, capturing both irreversible and recurrent dimensions of cellular state transitions. We envision that this represents a first step toward a general geometric theory for factorizing dynamic transcriptomic phenomena into simpler sub-processes—or "process times" [49]—and for estimating the underlying multi-temporal topological structure of the Waddington landscape.

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. Research in the G.P.D laboratory was supported by grants from the Swiss National Science Foundation (310030B_176404 "Genomic instability and evolution in cancer stromal cells") and the NIH (R01CA269356, the contents do not necessarily represent he official views of the NIH). K.M., M.A. and H.A.H. are members of the Max Planck/Oxford Center to Center collaboration supported by the Mathematical Foundations of Intelligence: An "Erlangen Programme" for AI, reference: EP/Y028872/1. K.M. and H.A.H. are members of the EPSRC Erlangen programme for AI Hub EP/Z531224/1.

References

1. Linnarsson, S. Single-cell biology meeting marks rebirth of an old science. *Genome Biology* **14,** 305. ISSN: 1465-6906. http://dx.doi.org/10.1186/gb-2013-14-4-305 (2013).
2. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLOS Computational Biology* **18** (ed Nie, Q.) e1010492. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1010492 (Sept. 2022).
3. Waddington, C. *The Strategy of the Genes* ISBN: 9781315765471. http://dx.doi.org/10.4324/9781315765471 (Routledge, Apr. 1957).
4. Bhattacharya, S., Zhang, Q. & Andersen, M. E. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Systems Biology* **5,** 85. ISSN: 1752-0509. http://dx.doi.org/10.1186/1752-0509-5-85 (2011).
5. Schiebinger, G. Reconstructing developmental landscapes and trajectories from single-cell data. *Current Opinion in Systems Biology* **27,** 100351. ISSN: 2452-3100. http://dx.doi.org/10.1016/j.coisb.2021.06.002 (Sept. 2021).
6. Li, H., Long, C., Hong, Y., Luo, L. & Zuo, Y. Characterizing Cellular Differentiation Potency and Waddington Landscape via Energy Indicator. *Research* **6.** ISSN: 2639-5274. http://dx.doi.org/10.34133/research.0118 (Jan. 2023).
7. Liang, S., Wang, F., Han, J. & Chen, K. Latent periodic process inference from single-cell RNA-seq data. *Nature Communications* **11** (2020).
8. Lederer, A. R. *et al.* Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. *Nature Methods.* ISSN: 1548-7105. http://dx.doi.org/10.1038/s41592-024-02471-8 (Oct. 2024).
9. Riba, A. *et al.* Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature Communications* **13.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-022-30545-8 (May 2022).

10. Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M. & Rajewsky, N. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology* **16,** e9946. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20209946. https://www.embopress.org/doi/abs/10.15252/msb.20209946 (2020).

11. Auerbach, B. J., FitzGerald, G. A. & Li, M. Tempo: An unsupervised bayesian algorithm for circadian phase inference in single-cell transcriptomics. *Nature Communications* **13** (Nov. 2022).

12. Talamanca, L., Gobet, C. & Naef, F. Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. *Science* **379,** 478–483 (2023).

13. Phillips, N. E. *et al.* The circadian oscillator analysed at the single-transcript level. *Molecular Systems Biology* **17** (2021).

14. Winfree, A. T. *The geometry of Biological Time* (Springer, 2001).

15. Thom, R. *Structural Stability and Morphogenesis: An Outline of a General Theory of Models* ISBN: 9780805392791. https://books.google.ch/books?id=3jtRAAAAMAAJ (W. A. Benjamin, 1975).

16. Rabadán, R. *et al.* Identification of relevant genetic alterations in cancer using topological data analysis. *Nature Communications* **11** (July 2020).

17. Carrière, M. & Rabadán, R. *Topological Data Analysis of Single-Cell Hi-C Contact Maps* in *Topological Data Analysis* (eds Baas, N. A., Carlsson, G. E., Quick, G., Szymik, M. & Thaule, M.) (Springer International Publishing, Cham, 2020), 147–162. ISBN: 978-3-030-43408-3.

18. Stolz, B. J. *et al. Relational persistent homology for multispecies data with application to the tumor microenvironment* 2023. arXiv: 2308.06205 [math.AT].

19. Kuchroo, M. *et al.* Single-cell analysis reveals inflammatory interactions driving macular degeneration. *Nature Communications* **14** (May 2023).

20. Imoto, Y. & Hiraoka, Y. V-mapper: Topological data analysis for high-dimensional data with velocity. *Nonlinear Theory and Its Applications, IEICE* **14,** 92–105 (2023).

21. Govek, K. W., Yamajala, V. S. & Camara, P. G. Clustering-independent analysis of genomic data using spectral simplicial theory. *PLOS Computational Biology* **15,** 1–20. https://doi.org/10.1371/journal.pcbi.1007509 (Nov. 2019).

22. Blumberg, A. J., Carrière, M., Fung, J. H. & Mandell, M. A. *Subsampling, aligning, and averaging to find circular coordinates in recurrent time series* 2024. arXiv: 2412.18515 [stat.ML]. https://arxiv.org/abs/2412.18515.

23. Damrich, S., Berens, P. & Kobak, D. *Persistent Homology for High-dimensional Data Based on Spectral Methods* 2024. arXiv: 2311.03087 [cs.LG]. https://arxiv.org/abs/2311.03087.

24. De Silva, V., Morozov, D. & Vejdemo-Johansson, M. Persistent Cohomology and Circular Coordinates. *Discrete & Computational Geometry* **45,** 737–759. ISSN: 1432-0444. https://doi.org/10.1007/s00454-011-9344-x (June 2011).

25. Perea, J. A. Sparse circular coordinates via principal $\mathbb{Z}$-bundles. *Topological Data Analysis,* 435–458 (2020).

26. Paik, T. & Park, J. Circular coordinates for density-robust analysis. *Foundations of Data Science.* https://www.aimsciences.org/article/id/68932a37a55b204b1c83c5fc (2025).

27. Baryshnikov, Y. & Schlafly, E. *Cyclicity in multivariate time series and applications to functional MRI data* in *2016 IEEE 55th Conference on Decision and Control (CDC)* (2016), 1625–1630.

28. Dotto, G. P., Buckinx, A., Özdemir, B. C. & Simon, C. Androgen receptor signalling in non-prostatic malignancies: challenges and opportunities. *Nature Reviews Cancer* **25,** 93–108. ISSN: 1474-1768. http://dx.doi.org/10.1038/s41568-024-00772-w (Nov. 2024).

29. Karthaus, W. R. *et al.* Regenerative potential of prostate luminal cells revealed by single-cell analysis. *Science* **368,** 497–505. eprint: https://www.science.org/doi/pdf/10.1126/science.aay0267. https://www.science.org/doi/abs/10.1126/science.aay0267 (2020).

30. Clocchiatti, A. *et al.* Androgen receptor functions as transcriptional repressor of cancer-associated fibroblast activation. *The Journal of clinical investigation* **128,** 5531–5548 (2018).

31. Ghosh, S. *et al.* Nuclear lamin A/C phosphorylation by loss of androgen receptor leads to cancer-associated fibroblast activation. *Nature Communications* **15.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-024-52344-z (Sept. 2024).

32. Mazzeo, L. *et al.* ANKRD1 is a mesenchymal-specific driver of cancer-associated fibroblast activation bridging androgen receptor loss to AP-1 activation. *Nature Communications* **15.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-024-45308-w (Feb. 2024).

33. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).

34. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* **1,** 417–425. ISSN: 2405-4712. http://dx.doi.org/10.1016/j.cels.2015.12.004 (Dec. 2015).

35. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Proceedings of the twentieth annual symposium on Computational geometry* (2004).

36. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440,** 358–362. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature04599 (Mar. 2006).

37. Paulsson, J. & Ehrenberg, M. Random Signal Fluctuations Can Reduce Random Fluctuations in Regulated Components of Chemical Regulatory Networks. *Physical Review Letters* **84,** 5447–5450. ISSN: 1079-7114. http://dx.doi.org/10.1103/PhysRevLett.84.5447 (June 2000).

38. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123,** 1025–1036. ISSN: 0092-8674. http://dx.doi.org/10.1016/j.cell.2005.09.031 (Dec. 2005).

39. Suter, D. M. *et al.* Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* **332,** 472–474. ISSN: 1095-9203. http://dx.doi.org/10.1126/science.1198817 (Apr. 2011).

40. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560,** 494–498 (Aug. 2018).

41. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38,** 1408–1414. ISSN: 1546-1696. http://dx.doi.org/10.1038/s41587-020-0591-3 (Aug. 2020).

42. Capolupo, L. *et al.* Sphingolipids control dermal fibroblast heterogeneity. *Science* **376,** eabh1623. eprint: https://www.science.org/doi/pdf/10.1126/science.abh1623. https://www.science.org/doi/abs/10.1126/science.abh1623 (2022).

43. Coppé, J.-P. *et al.* Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. *PLoS biology* **6,** e301 (2008).

44. Kuilman, T. *et al.* Oncogene-induced senescence relayed by an Interleukin-dependent inflammatory network. *Cell* **133,** 1019–1031 (2008).

45. Matsuda, S. *et al.* TGF- in the microenvironment induces a physiologically occurring immune-suppressive senescent state. *Cell Reports* **42,** 112129. ISSN: 2211-1247. http://dx.doi.org/10.1016/j.celrep.2023.112129 (Mar. 2023).

46. Kim, D. E. *et al.* Convergent roles of ATF3 and CSL in chromatin control of cancer-associated fibroblast activation. *Journal of Experimental Medicine* **214,** 2349–2368. ISSN: 1540-9538. http://dx.doi.org/10.1084/jem.20170724 (July 2017).

47. Hassan, S. N. *et al.* p21-Dependent Senescence Induction by BMP4 Renders Glioblastoma Cells Vulnerable to Senolytics. *International Journal of Molecular Sciences* **25,** 4974 (2024).

48. Bhatia-Gaur, R. *et al.* Roles for Nkx3.1 in prostate development and cancer. *Genes & Development* **13,** 966–977. ISSN: 0890-9369. http://dx.doi.org/10.1101/gad.13.8.966 (Apr. 1999).

49. Fang, M., Gorin, G. & Pachter, L. Trajectory inference from single-cell genomics data with a process time model. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2024/05/31/2024.01.26.577510.full.pdf. https://www.biorxiv.org/content/early/2024/05/31/2024.01.26.577510 (2024).

Methods

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 [1] 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 [2]
$$\texttt{expression[i,g]} = \log\left(1 + \frac{10000 \cdot \texttt{count[i,g]}}{\sum_g \texttt{count[i,g]}}\right)$$
3. We optionally apply a diffusion scheme using the MAGIC[1] algorithm [3].
4. Taking subsets of the remaining gene set determined by highly variable genes.

Mathematical background

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.

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

The following theorem forms the basis of practical applications of persistent cohomology, and is attributed to different references depending on the conditions [4–6]. 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 1 (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*

$$H^k(X_\bullet;\mathbb{F}) \cong \bigoplus_{\gamma\in\Gamma_k} \mathbb{I}_{[\varepsilon^\gamma,\delta^\gamma]} \tag{1}$$

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 (1).

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 [7] for the definition). The following theorem states that the persistence diagram is stable with respect to small perturbations of the underlying metric space.

Theorem 2 ([7], Thm. 3.1)**.**

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

Harmonic circular coordinates

In this section we review the circular coordinates algorithm from [8], emphasizing the density-weighted [9] and sparse [10] 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 3. *(cf. Thm. 4.57 [11]) For any simplicial complex $\mathcal{S}$, there is a bijective correspondence*

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

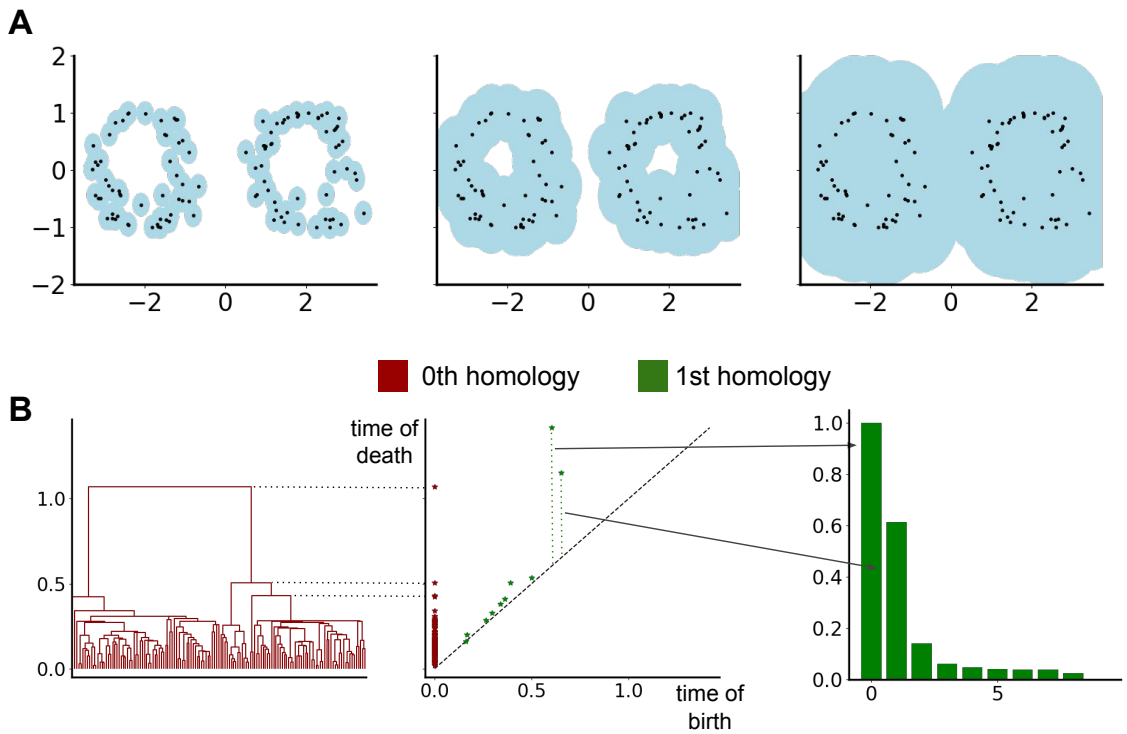


Figure 7: Introduction to topological data analysis. **(A)** Features that persist across a wide range of ε are more likely to be part of an underlying shape. Throughout the process of expanding balls around each point in a point cloud, many topological features emerge and disappear again, like connected 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 of 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 [12], 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 consists of 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

$$\theta : \mathcal{S}^0 \to S^1 = \mathbb{R}/\mathbb{Z}; \qquad \theta(x) = \tilde{\theta}(x) \mathrm{mod} \mathbb{Z}$$

satisfying

$$\alpha\big([x_p, x_q]\big) = d\tilde{\theta}\big([x_p, x_q]\big) \mod \mathbb{Z}. \tag{2}$$

The map θ extends to a continuous map defined on the geometric realization $|\mathcal{S}| \to S^1$ (see [8]).

Harmonic smoothing. Equation (2) 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 [8, 9] for further details.

Remark 4. *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 [8].

Proposition 5. *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 $H^2(\mathcal{S};\mathbb{Z})$, or equivalently $H_1(\mathcal{S})$, 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 [10] for details.

Density weighting. The single-cell data we work with has highly non-uniform density in practice. In [9], the authors define cochain inner products that correct for density and metric information, following the graph-Laplacian manifold-approximation literature [13].

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

$$K_t^\varepsilon(x_i, x_j) = \begin{cases} K_t(x_i, x_j) & d(x_i, x_j) \leq \varepsilon \\ 0 & \text{else,} \end{cases} \tag{3}$$

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 [9] are defined via the formulae

$$\langle f, g \rangle_0 = \sum_{x \in X} \frac{1}{P_t^\varepsilon(x)} f(x) g(x) \qquad \langle \alpha, \beta \rangle_1 = \sum_{[x_i, x_j] \in X_\varepsilon^1} W_t^\varepsilon(x_i, x_j) \alpha\big([x_i, x_j]\big) \beta\big([x_i, x_j]\big). \tag{4}$$

Remark 6. *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 [9]).

Sparse circular coordinates. In practice, one mitigates the computational complexity of persistent cohomology by working over landmark subsets [10]. Briefly, a landmark set

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

is attained by inductive maxmin sampling

$$\{x'_1, x'_2\} = \arg\max_{x_i, x_j} \{d(x_i, x_j)\} \qquad x'_k = \arg\max_{x_i} \ \min \big\{ d(x_i, x'_j) \mid j < k \big\} \tag{5}$$

until reaching a specified distance

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

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 (4). Following [10], the circular coordinate is extended from L to X by the formula

$$\theta : X \to S^1$$
$$x \in B_\varepsilon(x'_j) \mapsto \tilde{\theta}'_h(x'_j) + \sum_{k=1}^{\ell} \mu_k(x)\alpha'_h([x'_i, x'_j]) \mathrm{mod} \mathbb{Z}$$

where $\mu_i : B_\varepsilon(x'_i) \to \mathbb{R}_+$ is a partition of unity. This is well-defined, independent of the choice of covering neighbourhood for each x (see [10] Thms 3.2 and 4.2).

Circular Gene Set Enrichment

Gene set enrichment analysis (GSEA) [14] is a mainstay of single-cell RNA-seq workflows. We extend this idea to *circular enrichment*: a gene set is *circularly enriched* in a cell population when the expression profile restricted to that set forms a statistically significant circular structure.

Gene set projections. Let X denote a single-cell RNA-seq data matrix that has already been fully pre-preprocessed (normalization, noise removal, and restriction to a single cluster or cell type).

Definition 7. *For a gene set $\mathcal{G} = \{g_1, \ldots, g_\ell\}$ define:*

1. *$X_\mathcal{G}$: the restriction of X to the coordinates indexed by $\mathcal{G}$;*
2. *$X_\mathcal{G}^{(k)}$: the projection of $X_\mathcal{G}$ onto its first k principal components.*

Note that the matrices X, $X_\mathcal{G}$, and $X_\mathcal{G}^{(k)}$ contain the same cells; only the ambient geometry changes. Intuitively, $X_\mathcal{G}$ isolates the biological program associated with $\mathcal{G}$, while $X_\mathcal{G}^{(k)}$ offers a lower dimensional representation of that program. Throughout this work, we set $k = 3$ — large enough to capture the dominant variation yet small enough to suppress noise. By examining $X_\mathcal{G}^{(k)}$, we assess whether the dominant structure of this gene set manifests as a circular pattern.

Ring score formula. To efficiently screen large collections of gene sets, we introduce the *ring score* — a single statistic that quantifies how strongly the subspace $X_\mathcal{G}^{(k)}$ exhibits circular topology. Such ring-like geometry is characteristic of cyclic programs such as the cell cycle or circadian rhythm.
A point cloud is called *ring-structured* when it contains one prominent one-dimensional hole, as in (a sampled) circle, annulus, or hollow cylinder. The ring score, therefore, gauges how closely a point cloud resembles these ideal shapes. We derive it from the one-dimensional persistence diagram of $X_\mathcal{G}^{(k)}$.
Let the persistence diagram consist of n birth-death pairs

$$\mathrm{dgm}^1(X_\mathcal{G}^k) = \left\{ (b_i, d_i) \mid 0 \le i \le n \right\}$$

and denote their lifetimes $p_i = d_i - b_i$, where the indices are ordered such $p_0 \ge p_1 \ge \ldots$. The dominant lifetime p_0 reflects the main circular feature, whereas the remaining lifetimes represent additional loops that "blur" this signal.

Definition 8 (Generalized Ring Score Formula)**.** *Let* $\mathbf{p} = (p_i)_{i\in\mathbf{N}}$ *be a non-increasing sequence of non-negative real numbers and choose a normalization constant* $N \geq p_0$*. Then we define the* generalized 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 9. *Because* $0 \leq p_i \leq p_0$ *for all* i *and* $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$*, the range of the score is given by* $s \in [0, 1]$*, with larger values indicating a more dominant circular structure. The dyadic wights* 2^{-i} *dampen the influence of less persistent features and, together with the normalization by* N*, ensure that the score lies within the unit interval. More formally,*

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

Scale invariance of ring score Let $X = \{X_i\}_{i=1}^m \subset \mathbb{R}^n$ be a finite point cloud with the center of mass $\mu(X) = \frac{1}{m}\sum_{i=1}^m X_i$. *Metric scaling* by a factor $c > 0$ keeps the centroid fixed while multiplying every pairwise distance by c:

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

Write $\mathbf{p}(X) = (p_i)_{i\in\mathbb{N}}$ for the ordered lifetimes in the one-dimensional persistence diagram of X, and let $N(X)$ be the normalization constant that appears in Definition 8. If $N(X)$ is chosen so that

$$\frac{p_0(c \cdot X)}{N(c \cdot X)} = \frac{p_0(X)}{N(X)}\,,$$

then the ring score is invariant under metric scaling. In particular, two natural and convenient choices for N satisfying this requirement are:

- $N(X) = p_0(X)$, the longest lifetime itself;
- $N(X) = \operatorname{diam}(X)$, the Euclidean diameter of the point cloud.

In fact, for point clouds embedded in $\mathbb{R}^n$ the ratio $\frac{p_0(X)}{\operatorname{diam}(X)}$ achieves its theoretical maximum of $\frac{\sqrt{3}}{2}$, when X is isometric to a perfect circle.

Two families of ring scores The ring score formula of Definition 8 merely requires a non-increasing sequence of non-negative numbers and in particular this sequence does not need to be persistence lifetimes. An increasingly popular alternative to lifetimes is to use *persistence ratios*, $\pi_i = \frac{\text{death}_i}{\text{birth}_i}$ [15].
Because no shape maximizes $\pi_0/\operatorname{diam}(X)$, a convenient scale-invariant choice is simply $N(X) = \pi_0(X)$. Thus we define the following two families of ring scores.

Definition 10 (Diameter and ratio ring scores)**.** *Let $X \subset \mathbb{R}^n$ be a finite point cloud with diameter $d = \mathrm{diam}(X)$. Denote by $\mathbf{p}$ and by $\boldsymbol{\pi}$ the sequences of lifetimes and persistence ratios, respectively. We define*

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

Iterative Density Filtering While persistent homology is robust under perturbations to the underlying point cloud, it can be somewhat sensitive to outliers. To mitigate this, we elected to perform a basic kernel-density filtering as a pre-processing step within the circular enrichment pipeline. For a choice of bandwidth parameter $\sigma \in \mathbb{R}_+$, we used the Euclidean kernel

$$K(x, y) = \exp\left(\frac{\|x - y\|^2}{2\sigma}\right) \tag{6}$$

over points x, y in Euclidean space $\mathbb{R}^n$. We used the quantity

$$\mu(x) = \sum_{x \sim y} K(x, y) \tag{7}$$

where $x \sim y$ if y is one as a rough estimate of the density function on the point. We then subset the data using upper and lower percentiles of the density function in Equation (7), repeating this process for a fixed number of times n.

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 pre-processed 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. [14]

Cyclicity Analysis

Path cyclicity analysis

Cyclicity analysis is a method introduced by [16] 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

$$\mathbf{L}_{i,j}(\gamma) = \int_{S^1} \gamma^*(x_i dx_j - x_j dx_i) = \int_0^1 \Big(\gamma^i(t)\frac{d\gamma^j}{dt}(t) - \gamma^j(t)\frac{d\gamma^i}{dt}(t)\Big)dt. \tag{8}$$

The (i, j)-entry of this matrix is called the (i, j) *lead-lag coefficient*. In simple models (cf. (12)), 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

$$\mathbf{L}_{i,j}(\gamma) = \int_{\partial D} \gamma^*(x_i dx_j - x_j dx_i) = \frac{1}{2}\int_D dx_i \wedge dx_j \tag{9}$$

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

$$\chi(t) = \sum_{m=0}^{\infty} z_m \exp(2\pi i m t) \tag{10}$$

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

$$\gamma^i(t) = \mathbf{a}_i \cdot \chi(t - \mathbf{b}_i). \tag{11}$$

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 [16]. Suppose that $\gamma : S^1 \to \mathbb{R}^n$ has offsets as in Equation (11) and that the generating function $\chi : S^1 \to \mathbb{R}$ has Fourier decomposition in Equation (10).

Proposition 11 (Prop. 5.1., [16])**.** *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 [16], suppose we have a simple case where the primary function

$$\chi(t) = z_1 \exp(2\pi i t) \tag{12}$$

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 11 to the chain-of-offsets $\gamma : S^1 \to \mathbb{R}^n$ shows that the lead-lag matrix is

$$\mathbf{L}_{i,j}(\gamma) = 2\pi |z_1|^2 \sin\Big(2\pi(\mathbf{b}_i - \mathbf{b_j})\Big), \tag{13}$$

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 [16], 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

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

provides a 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 exists 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 (12) 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 13 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

$$\mathbf{L}(\gamma) = 2\pi|z_1|^2\Big(\mathbf{u}\otimes\mathbf{v} - \mathbf{v}\otimes\mathbf{u}\Big). \tag{14}$$

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 [16], 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

$$\begin{bmatrix}\mathrm{Re}(\mathbf{z})_k \\ \mathrm{Im}(\mathbf{z})_k\end{bmatrix} = \begin{bmatrix} \frac{-\cos\theta}{\|\mathbf{u}\|} & \frac{1}{\|\mathbf{v}\|} \\ \frac{\sin\theta}{\|\mathbf{u}\|} & 0 \end{bmatrix}\begin{bmatrix}\mathbf{u}_k \\ \mathbf{v}_k\end{bmatrix} = 2\pi|z_1|^2\mathbf{a}_k\begin{bmatrix} \frac{-\cos\theta}{\|\mathbf{u}\|} & \frac{1}{\|\mathbf{v}\|} \\ \frac{\sin\theta}{\|\mathbf{u}\|} & 0 \end{bmatrix}\begin{bmatrix}\cos(2\pi\mathbf{b}_k) \\ \sin(2\pi\mathbf{b}_k)\end{bmatrix}. \tag{15}$$

This formula implies the following important point, as noted in [16]:

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 12. *By 15, 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 future work.*

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

$$\mathbf{L}_{i,j}(\varphi, \eta) = \langle \eta, \varphi^* \omega_{i,j} \rangle_{\Omega^1(M)}. \tag{16}$$

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 (8).

Proposition 13. $\mathbf{L}_{i,j}(\gamma) = \mathbf{L}_{i,j}(\gamma, d\theta)$.

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

$$\int \varphi^* \omega_{i,j} : [x_p, x_q] \mapsto \int_I \varphi_{p,q}^* \omega_{i,j} \in \mathbb{R} \tag{17}$$

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 (4) which is induced by the geometry of the ambient $\mathbb{R}^n$.

Definition 14. *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 17 has the following interpretation.

Proposition 15.

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

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 4, the coefficients of the lead-lag matrix are given by the following formula:

$$\mathbf{L}_{i,j}(\varphi, \alpha) = \sum_{[x_p,x_q] \in \mathcal{X}_\varepsilon} \frac{K_t^\varepsilon(x_p, x_q)}{P_t^\varepsilon(x_p) P_t^\varepsilon(x_q)} \alpha([x_p, x_q]) \cdot \det \begin{bmatrix} \varphi_i(x_p) & \varphi_i(x_q) \\ \varphi_j(x_p) & \varphi_j(x_q) \end{bmatrix}. \tag{18}$$

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

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 a block diagonal $n \times n$ matrix

$$\Sigma = \begin{bmatrix} 0 & -\lambda_1 & & & & & \\ \lambda_1 & 0 & & & & & \\ & & 0 & -\lambda_2 & & & \\ & & \lambda_2 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \\ & & & & & & 0 \end{bmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \tag{19}$$

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 16 is that we can interpret Σ from Equation (19) 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 (18), we have the following interpretation:

the value λ_i in Equation (19) 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 embedded 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.

Cyclicity analysis on single-cell data

The goal of this section is to 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 with 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 (4)
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 principal complex eigenvector components approximate the global order in which the genes in $\mathcal{G}$ occur along α_h (Figure 2D).

Data and code availability

The `cHunter` software package and the Jupyter Notebooks to analyze the publicly available data are available online at Github. The original human dermal fibroblast data will be deposited in GEO under accession GSE312110.

Author contributions

K.M., M.K.Y. and K.H conceived the study. K.M and M.K.Y developed the overall analysis framework and led the project. K.M. conceived of topological and geometric framework, implemented the cHunter software, performed the topological and geometric analyses and generated the figures. M.K.Y. contributed to the development of the circular enrichment framework, implemented key algorithms and performed computational analyses. T.J.N. assisted with synthetic experiments and design. J.I. designed and carried out the experimental work. J.I, W.K. and C.P provided biological expertise and interpretation of results. M.A. contributed to data curation, software development and reproducible analysis workflows. H.A.H. and K.H. supervised the mathematical and computational aspects of the work. G.P.D. supervised the biological and translational aspects of the work. K.M., M.K.Y. and G.P.D. wrote the manuscript with input from all authors. All authors discussed the results and approved the final version of the manuscript.

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

Competing interests

The authors declare no competing interests. G.P.D. has no financial or non-financial competing interests, he is the founder of EpiKare Bio.

Methods References

1. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology.* https://doi.org/10.1038/s41587-023-01767-y (2023).
2. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36,** 411–420 (2018).
3. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174** (2018).
4. Webb, C. Decomposition of graded modules. *Proceedings of the American Mathematical Society* **94,** 565–571 (1985).
5. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Proceedings of the twentieth annual symposium on Computational geometry* (2004).
6. Edelsbrunner, H., Letscher, D. & Zomorodian, A. *Topological persistence and simplification* in *Proceedings 41st annual symposium on foundations of computer science* (2000), 454–463.
7. Chazal, F., Cohen-Steiner, D., Guibas, L. J., Memoli, F. & Oudot, S. Y. Gromov-Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum* **28,** 1393–1403 (July 2009).
8. De Silva, V., Morozov, D. & Vejdemo-Johansson, M. Persistent Cohomology and Circular Coordinates. *Discrete & Computational Geometry* **45,** 737–759. ISSN: 1432-0444. https://doi.org/10.1007/s00454-011-9344-x (June 2011).
9. Paik, T. & Park, J. Circular coordinates for density-robust analysis. *Foundations of Data Science.* https://www.aimsciences.org/article/id/68932a37a55b204b1c83c5fc (2025).
10. Perea, J. A. Sparse circular coordinates via principal $\mathbb{Z}$-bundles. *Topological Data Analysis,* 435–458 (2020).
11. Hatcher, A. *Algebraic topology* ISBN: 9780521791601 (Cambridge University Press, Cambridge; New York, 2002).
12. Eckmann, B. Harmonische Funktionen und Randwertaufgaben in einem Komplex. *Commentarii Mathematici Helvetici* **17,** 240–255. ISSN: 1420-8946. http://dx.doi.org/10.1007/BF02566245 (Dec. 1944).
13. Belkin, M. & Niyogi, P. Towards a theoretical foundation for laplacian-based manifold methods. *Learning Theory,* 486–500 (2005).
14. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).
15. Bobrowski, O., Kahle, M. & Skraba, P. Maximally persistent cycles in random geometric complexes. *The Annals of Applied Probability,* 2032–2060 (2017).
16. Baryshnikov, Y. & Schlafly, E. *Cyclicity in multivariate time series and applications to functional MRI data* in *2016 IEEE 55th Conference on Decision and Control (CDC)* (2016), 1625–1630.

A Experiment and pre-processing details

A.1 Synthetic Transcriptomic Switch Model

Classical model Our main synthetic model was a modified version of the linear ODE model of gene transcription and translation in [1]. The original linear model in [1] is governed by two ODEs:

$$\frac{d\mathbf{r}}{dt} = \mathbf{A}\mathbf{p} - \mathbf{B}\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = \mathbf{C}\mathbf{r} - \mathbf{D}\mathbf{p} \tag{20}$$

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 [2, 3]. More recent models of RNA transcription account for this by designating discrete transcriptional states with constant transcription rates [4–7]. To adapt the linear ODE model in (20) to this phenomenon, 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

$$\frac{d\mathbf{r}}{dt} = \sigma(\mathbf{A}\mathbf{p};\alpha,\tau) - \mathbf{B}\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = \mathbf{C}\mathbf{r} - \mathbf{D}\mathbf{p}. \tag{21}$$

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 S1D and Figure S2D.

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 (see Cyclicity Analysis in the Method's section), we hypothesized that when the protein-RNA interaction matrix $\mathbf{A}$ in the transcriptomic switch model 21 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 $\mathbf{A}$. Despite the modified transcriptional switch model and realistic noise, the lead-lag matrix (Figure S2B) provided a relatively robust model of the relationships between different genes (Figure S2C).

Transient processes We also wanted to demonstrate the viability of using circular enrichment in studying transient processes 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 S1). Circular gene set enrichment revealed significant circular structure outside of the first principal component (Figure S1G,H). The reconstructed phase estimates reliably recovered the ordering of genes (Figure S1I,J), and the lead-lag matrix provided a reasonable estimation of the underlying gene-protein interaction matrix (Figure S1C).

Model parameters In Table 8 we present the parameters for the transient (T) and recurrent (R) examples (Figure S1 and Figure S2) of the synthetic transcriptomic switch model. Parameters were manually selected to lead to approximate ODE solutions that were sufficiently rich. Sensitivity analysis (Figure S2D) 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.

Sampling with negative binomial noise To realistically model noisy and stochastic single-cell data, we adopted the approach of [8] of negative binomial sampling, which has been observed to be a good model for the distribution of RNA molecules [2, 3]. Our approximate solutions of mRNA levels according to Table 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). Recall that 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 Table 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

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

	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

Table 8: Hyperparameters in the transcriptomic switch model experiments.

that converges to our solved system in the mean. Data were sampled from this using the Python function `scipy.stats.nbinom.rvs`. Examples of the sampling are given in Figure S1E-G and Figure S2E-G.

A.2 Benchmarking

Velocycle synthetic model. We benchmarked the estimation of circular coordinates against the synthetic model presented in [8]. In this model, mean spliced RNA expression is modeled directly as the 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. Each of the parameters $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are drawn from a normal distribution with predefined mean and correlation structure (see [8] for details), and each cell i is associated with a pseudotime t_i, drawn from a uniform distribution $\mathcal{U}(0, 1)$. 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), but this time with a stochastic dispersion parameter given by

$$n = 1/d \sim \Gamma(1, 2)\,,$$

where $\Gamma(\alpha, \beta)$ is the gamma distribution with shape parameter α and rate parameter β.

Velocycle benchmarking. The pipeline for our model in the `velocycle` benchmarking (Figure S3A and C) consisted of: separating each gene into spliced and unspliced raw UMI counts, filtering genes that occurred in fewer than three cells, normalizing the per-cell raw UMI counts, and projecting 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 preprocessing and parameters for 200 epochs [8]. The box-plots in Figure S3C were run with the same parameters 20 times with 3000 cells and 300 genes, as per [8].

Multi-process benchmarking. The multi-process benchmark was built from a minor adaptation of the `velocycle` synthetic data generation framework. This benchmark uses a dataset containing two independent processes (linear and circular) to evaluate the performance of `cHunter`, `PCA` and `velocycle` with an incorrect prior, i.e., the underlying manifold assumed is a circle, S^1, whereas by construction this dataset has the shape of a cylinder, $S^1 \times I$.

For a fixed number of cells c and genes g, we generated the standard `velocycle` dataset to attain a circular process $X_{\text{circ}} \subset \mathbb{R}^{c \times g}$. We then generated another, independent, `velocycle` dataset with $2c$ cells and g genes, before restricting to the first half c cells ordered by phase to create a non-circular process modelled by a half-circle $X_{\text{arc}} \subset \mathbb{R}^{c \times g}$. The two datasets were then shuffled and concatenated along the cell axis, to create a dataset consisting of separate uncoupled circular and linear processes with c cells and $2g$ genes $X_{\text{total}} \subset \mathbb{R}^{c \times 2g}$. All `cHunter` estimates were carried out in the concatenated unspliced-spliced phase space.

A.3 Cycling dermal fibroblasts

Dataset. The data we used from [9] is publicly available on the Gene Expression Omnibus database under accession no. GSE167209. We followed the pre-processing procedure in [8], 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.4 AR silencing in dermal fibroblasts

Dataset Our single-cell RNA-seq dataset consisted 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.

Preprocessing 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 CTRL subgroups, respectively. We first performed clustering via the Leiden algorithm and filtered three populations of cells based on abnormally low total RNA counts, mitochondrial percentage reads, and MALAT1 expression (Figure S5A,B,C). 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. Highly variable genes were then identified using `scanpy.tl.highly-variable`. For comparison, we used scanpy's default cell-cycle scoring to assign G1, S and G2M labels to each cell (Figure S5D,E,F). Notably, while this overall cell cycle scoring showed similar G1, S, and G2/M distributions across the bulk populations of control and AR-silenced cells, further analysis would reveal specific cell-state changes linked to AR silencing (see below).

A.5 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 [10] 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 [11] 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 Database [12]. The global pre-processing was first applied to the luminal cells above. We then projected onto each gene set, followed by the first two principal components, applied iterative density filtering and finally computed the circular coordinate of the most persistent feature. The p-values were computed according to the permutation procedure described in Method .