The topology of cellular ontogeny

Thesis by
Emanuel Flores Bautista

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 12th, 2025

Emanuel Flores Bautista
ORCID: 0000-0002-2810-1757

A mi madre, mi padre y mis abuelos.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents Roxana and José Luis, and my grandparents Estela, Juana, Francisco and José. All of what I have is the result of your hard work, but more fundamentally, of your love towards myself during my formative years. Gracias Ma, y Pa, por todo lo que han hecho por mí. I also want to thank my siblings, Luisa Jhoana and José Luis, for all the adventures during this journey, and for helping me keep myself grounded.

I would also like to give special thanks to my advisor, Matt Thomson. Your vision, guidance, and outstanding support has forged me as a scientist. I feel incredibly lucky to have had the opportunity to work in the Thomson Lab. The amount of intelectual freedom I had to explore my ideas, and dive into the world of algebraic topology is something I will always be genuinely grateful for. The numerous conversations and discussions with Matt, ranging from Betti numbers, to biocircuits, and the Mayer-Vietoris sequence will forever be cherished in my memory. In the Thomson Lab, I have learned the power of rigor, building things from first principles, and the importance of "playing with the concepts" to get a feel for them as second nature.

I would also like to thank my committee members, Matilde Marcolli, David Prober, and Lior Pachter for providing essential feedback on this work. In particular, I want to thank Matilde for inspiring me to enter to the world of topology with her fascinating lectures, and for putting me in contact with members of her group: Juan Pablo Vigneaux, and Sita Gakkhar. My conversations with Juan Pablo and Sita were essential to my understanding of the effects of dimension in the inference of topological structures in single-cell data.

I also want to thank the members of the Thomson lab, past and present, for the numerous conversations, discussions, and comraderie. In particular, I want to give huge thanks to my collaborator, Binglun Shao, as the version of this project is a product of our conversations, discussions, and her fantastically creative and rigorous work ethic. Binglun developed the Critical Edge method, optimized the topogene analysis, and spearheaded the human immune and neuroectoderm analyses.

Conversations with other members of the group including Enrique Amaya, Rex Liu, Alex Detkov, James Gornet, Jialong Jiang, Dave Brown, Jerry Wang, Tami Khazinei, Sisi Chen, Yu-Jen Chen, and others were essential to this work, and I'm grateful to have coincided with them during my time at Caltech.

I want to thank Professors Richard Murray for accepting me into the SURF program in 2018 to work on biocircuits. Also want to thank Prof. Rob Phillips for hosting me in his group for summer research and a research rotation in his lab characterizing the promoters in the bacterium *E. coli*. Both of these experimental lab training experiences were essential for my formation as a scientist. I also want to thank my committee chair, Prof. David Prober, for a rotation at the beginning of my PhD. Naturally, the guidance from mentors during these times were essential: I want to thank first and foremost Andy Halleran, which very patiently taught me all of the pipeline for building synthetic biocircuits in the lab, and for the amazing brainstorming sessions we had when trying to model the evolutionary stability of the circuits. I would say that Andy intiated me into the world of rigorous scientific research, and I probably would not be here without his unwavering mentorship. Suzy, Bil and Andrey were also great mentors during my rotations, and my gratitude goes to them as well.

Besides the pure pursuit of knowledge, it was also an absolute honor and joy to have crossed paths with so many amazing individuals at Caltech. Importantly, I was fortunate to meet a great Mexican group of graduate students including Manuel Razo, Jorge Castillo, Andrés Ortiz, Alejandro Granados, Jesus Del Río, and my great friends and labmates Enrique Amaya and David Larios. I feel tremendously fortunate that our times at Caltech coincided. I have learned so much from all of you, not only in the academic sense, but about so many integral aspects of life; I hold all of you in the highest level of respect and honor in my spirit. Other friends and colleagues I was fortunate to meet at Caltech include Uriah Israel, Tom Röschinger, Victor García, Andy DeLaitsch, Reed McCardell, Jon Marken, and many more. I thank Justin Bois, Kostia Zuev, and Joel Tropp, for their amazing courses in Applied and Computational Mathematics, as they were integral parts of my formation. I also want to thank the Caltech community at large: Laura and Daniel at ISP, Alice

Sogomonian and Divina Bautista at the Health Center, Ernie, the staff at Red Door and Chandler, Tess in the Bursar's office, Rebecca Fox, Jessica Silva, Sue Zindle, Darrell Peterson. You were all essential to my experience at Caltech, and I am incredibly greateful for your support.

Finally, I would like to thank some people that were essential before coming to Caltech. My advisor from my undergraduate research, Ernesto Pérez-Rueda, was a fundamental character in my path to choose a career in science. I am grateful for his mentorship, and for being able to interact and work with him and his group for two years prior to coming to Caltech. I also want to thank my friend Jorge Flores, who introduced me to the world of algebraic topology. Our conversations about the subject were really fun and in hidsight seem like the aura of fun in pure discovery we had during these conversations extended into the whole of my PhD research.

ABSTRACT

A fundamental goal of modern biology is to build global, predictive models of gene regulation that encompass diverse physiological contexts. Single-cell transcriptomics has enabled the creation of developmental cell atlases—detailed catalogs of gene expression patterns and differentiation trajectories at an organismal scale. The widespread availability of cell atlases across metazoan model organisms presents an opportunity to construct global theories of cell-state control. In this thesis, we introduce a framework that uses persistent homology to decompose cell atlases into topological structures that provide signatures of gene regulation at the scale of an organism. Using this framework, we found that the topological structure of a broad set of developmental atlases contains only a discrete set of topological structures—such as clusters, trees, and loops—revealing the recurrent use of global gene regulatory strategies. Our analysis revealed that the tree topology, while predominant, is not universal. Indeed, we identified non-trivial topologies containing loops in the development of human immune cells, seam-hypodermal cells in *C. elegans*, and the cnidocytes of multiple cnidarians. Analysis of cell-state manifolds with non-trivial topology demonstrated an important role of convergent structures in increasing cellular diversity along paths to a common cell fate, and of cyclic structures in self-renewal of progenitor-like states. Together, this work provides a global perspective on principles of cell-state regulation, and suggests that loops are important organizing structures for controlling cell differentiation.

PUBLISHED CONTENT AND CONTRIBUTIONS

Flores-Bautista, Emanuel and Matt Thomson (2023). "Unraveling cell differentiation mechanisms through topological exploration of single-cell developmental trajectories". In: *bioRxiv*. DOI: 10.1101/2023.07.28.551057. URL: https://www.biorxiv.org/content/early/2023/11/01/2023.07.28.551057.

E.F.B. participated in the conception of the project, developed the mathematical and computational framework, analyzed the data, and wrote the manuscript. Chapter 2 is an update to this article.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

Unlike other scientific disciplines, biology underuses the power of representing observables as mathematical objects to build predictive models. In this sense, Caltech has been a notable dissident in this regard, placing computational and mathematical thinking at the center stage for biological research, and this thesis is a reflection of that ethos. Professor Rob Phillips aptly notes that "biology is in the era of Tycho Brahe," referring to the field's current focus on data collection with limited emphasis on discovering mathematical laws governing observed phenomena. This thesis aims to establish a theoretical framework that leverages topological invariants to test hypotheses of development and make global predictions of developmental structures. Throughout the following chapters, I will present advances towards a mathematical theory of cell fate control—an *esquisse* for a program that harnesses topology to guide a predictive dynamical theory of development.

1.1 Cell differentiation is the foundation of metazoan complexity

The remarkable diversity of animal forms—spanning bilaterians, cnidarians, sponges (porifera), comb jellies (ctenophores), and placozoa—shares a fundamental characteristic: the division of labor at the cellular level (Richter and King, 2013). This multicellular specialization forms the cornerstone of metazoan complexity (Sebé-Pedrós, Degnan, and Ruiz-Trillo, 2017).

At its core, animal development proceeds through cell differentiation: a sophisticated series of processes through which cell identity changes, orchestrated by biomolecular regulatory networks (Liberali and Schier, 2024). Through this remarkable process, animals *develop* from a single-celled zygote into structured and organized multicellular entities.

The fascinating question of how cells in the animal embryo diversify, self-organize, and coordinate with precision has captivated scientists since antiquity. Historical records show that even Aristotle studied chick embryos, while numerous other philosophers and scientists were captivated by

developmental phenomena.

1.2 A brief history of developmental biology: the advent of the genomic era

The study of animal development advanced significantly in the late 19th and early 20th centuries through pioneering embryological studies. Although largely observational, these ingenious investigations provided fundamental insights into cell fate control. In 1891, Driesch (Sander, 1997) demonstrated totipotency at the two-cell stage of the sea urchin embryo *Echinus microtuberculatus* by separating two-celled embryos and observing that each cell developed into a complete, albeit smaller, animal. Another landmark experiment was conducted by Mangold in 1924, revealing that a specific region of the frog embryo (now known as the Mangold-Spemann organizer) could induce the formation of a complete head and spinal cord when transplanted into another embryo.

By the time of Mangold's discovery, Mendel's theory of inheritance had been rediscovered, inspiring researchers to study development through direct genetic manipulation. However, the biochemical composition of genes remained elusive in the early 20th century. A paradigm shift was necessary to understand the molecular underpinnings of animal development. In the 1920s, Morgan and his group demonstrated that genes resided in chromosomes and discovered genetic linkage. In 1941, Beadle and Tatum showed that genes code for enzymes, advancing the maturation of the field of genetics. It was until 1952, that Hershey and Chase provided evidence that genes were biochemically composed of DNA by showing that only phage DNA was necessary for bacterial infection.

Another paradigm shift occurred in 1961, when Jacob and Monod proposed a model for gene regulation, based on work on the lactose system of *E.coli* and the lysogenic circuit of the λ phage. In their model regulatory genes control the activity of other genes upon binding to a DNA sequence upstream of the controlled gene(s). In parallel, in 1969, Britten and Davidson published a theory on the role of gene regulation in cell differentiation during development, though the molecular mechanisms remained mysterious. Together, these studies established that the cellular diversity in animal development could driven by the spatiotemporal control of gene expression. Thus, it became

increasingly clear that mapping the expression patterns of key developmental genes in time and space would aid in our understanding of animal development. Direct visualization methods such as *in situ* hybridization enabled the discovery of spatial patterns of developmental genes (Lyons, Hogan, and Robertson, 1995). However, *in situ* hybridization was limited by the number of genes that could be studied simultaneously.

The 1990s witnessed the flourishing of genomics, with next-generation sequencing catalyzing the development of modern molecular biology methods. One of the most groundbreaking advances in genomics became possible: genome-wide gene expression profiling through mRNA sequencing (RNA-seq) (Mortazavi et al., 2008). In contrast to previous approaches, RNA-seq provided a comprehensive view of gene expression. This technology facilitated numerous discoveries in developmental biology, including the characterization of gene modules driving cell identity during fate commitment and the identification of novel regulators for cell fate specification.

The study of specific cell populations in embryos using RNA-seq relied on techniques like fluorescence-activated cell sorting (FACS), but still presented limitations for biological discovery. Recognizing the cellular heterogeneity within tissues inspired the development of single-cell RNA-seq (scRNA-seq), which enabled the study of transcriptional programs across different cell types within a sample.

After the technology matured, several methods in 2015 (Klein et al., 2015; Macosko et al., 2015) leveraged microfluidics to profile thousands of cells simultaneously. An important application of scRNA-seq became the creation of comprehensive catalogs of cell states across an animal's developmental history—or *developmental cell atlases*. Following 2015, numerous developmental atlases were published for various organisms including frogs, zebrafish, and the nematode *C. elegans*.

The scale and complexity of single-cell atlases introduced new computational challenges. Among these, cell differentiation trajectory inference is crucial for interpreting developmental atlases, and will be the topic for discussion in the next section.

1.3 Trajectory inference is a topological problem

During differentiation, cells traverse specific paths in gene expression space. Single-cell RNA-seq data provides snapshots of these paths, and a fundamental goal in computational biology is to reconstruct these differentiation trajectories. Consequently, significant effort has been devoted to developing computational methods for trajectory inference. These methods aim to identify the global structure of the cell state manifold, which encompasses all cell differentiation trajectories within a single-cell atlas.

Since reconstructing the global structure of a cell state manifold is an inference problem, a reasonable approach involves making assumptions about the underlying structure. Cell differentiation has traditionally been conceptualized as a branching process, wherein cells make fate decisions along their path toward mature cell types, forming a tree-like structure. This conceptual model is encapsulated in the Waddington landscape metaphor, which depicts development as cells rolling down a hill, with plateaus representing branch points and valleys representing terminal cell types. As a result, many trajectory inference methods have been designed to identify tree structures from data, e.g. Wishbone (Setty et al., 2016), Monocle (Trapnell et al., 2014), and URD (Farrell et al., 2018) (see Saelens et al., 2019 for a review).

Other approaches have used both topological and geometric methods to infer the global structure of cell state manifolds. In particular, much work has been based on analyzing the eigenvectors of the Laplacian of a k-Nearest Neighbors graph of the data (e.g. (Angerer et al., 2015) and PHATE (Moon et al., 2019)). A different approach was used in scTDA (Rizvi et al., 2017), where the authors proposed a method to represent the cell state manifolds using Mapper (Singh, Mémoli, Carlsson, et al., 2007), a dimensionality reduction method that doesn't assume a presecribed tree-like topology for representing the data manifold. Mapper falls under the umbrella of topological data analysis (TDA), which has gained traction in recent years as a powerful tool for analyzing complex data sets. TDA provides a framework for extracting meaningful features from high-dimensional data, enabling the identification of *structure* by harnessing concepts from Algebraic Topology.

Persistent homology (PH) is a powerful technique that captures the topological features of data at multiple scales. Notably, persistent homology, the core tool used in our work, is also a TDA method. PH has been succesfully applied to the study of biological data, e.g. in (Benjamin et al., 2024) the authors used PH to classify cancer images. More relevant to our work, PH has also been used for studying cyclic expression patterns (Maggs et al., 2025). However, there has been, to the best of our knowledge, no study has used PH to study the topology of cell state manifolds in development. Therefore it remains poorly understood if there are higher order topological features, like loops or voids, that are relevant for understanding the topology of cell state manifolds.

1.4 Organization of the thesis

Chapter 2 constitutes the core work of this thesis, presenting a novel framework for analyzing the topology of cell state manifolds. To make this thesis self-contained, in Chapter 3, I provide a detailed and pedagogical exposition of the mathematical framework, assuming only familiarity with Linear Algebra.

Chapter 2

TOPOLOGICAL SIGNATURES OF GENE REGULATION REVEAL GLOBAL PRINCIPLES OF CELL STATE CONTROL

ABSTRACT

Embryo development requires precise control of a plethora of biological networks across space and time. Fundamentally, biomolecular networks are dynamical systems and topological methods can provide insights into their design principles. For example, multistable and oscillatory designs have distinct topological signatures: fixed points and periodic orbits. Crucially, these signatures can be quantified by the topological notion of homology groups, and their associated dimension, the Betti numbers. Moreover, single-cell developmental atlases provide an unprecedented view into the dynamics of development. Despite numerous efforts, it is still poorly understood what types of complex topological structures are present in developmental atlases and what functional roles they play during the development of multicellular organisms. To address these questions, we developed `totopos`, a computational framework to examine the topological features of transcriptome spaces by quantitative analysis of their homology groups. First, we showed that we can identify genetic drivers of topological structures in simulated datasets. We then applied our topological approach to more than ten single-cell developmental atlases discovering that transcriptome spaces are predominantly path-connected and only sometimes simply connected. Finally, we applied `totopos` to examine gene expression patterns in specific topological loops. These loops represented stem-like and convergent cell circuits, observed in a wide array of developmental systems such as in the human immune system, the seam cells of *C. elegans*, and the cnidocytes of cnidarians. Our results show that differentiation mechanisms can use complex topological modules and that these modules can be selected during evolution. Thus, our approach to studying the topological properties of developmental transcriptome atlases opens new possibilities for understanding the development of multicellular organisms.

2.1 Introduction

Metazoan development involves the progressive specialization of a totipotent cell state to a plethora of cell types constituting the complex tissues present in the adult organism. This process involves both progressive fate restriction and dynamic spatial organization, with cells adopting specific gene expression patterns at precise locations. Single-cell RNA sequencing (scRNA-seq) has transformed our ability to study this process, enabling the creation of single-cell developmental atlases across metazoans (Briggs et al., 2018; Packer et al., 2019; Siebert et al., 2019; Steger et al., 2022; Liao et al., 2022; Plass et al., 2018; Calderon et al., 2022; Lange et al., 2023; Zhang et al., 2020; Qiu et al., 2024).

Modern single-cell developmental atlases reveal the landscape of regulatory states that cells traverse during differentiation, yet extracting fundamental principles from these complex datasets remains challenging. Dynamical systems theory suggests that topologically equivalent phase spaces exhibit a globally similar behavior (Arnold, 1991; Hopf, 1927). Thus, identifying the topology of developmental atlases across organisms could shed light on shared principles of cell state control. Indeed, several efforts have aimed to analyze the topology of single-cell data (Rizvi et al., 2017; Vipond et al., 2021). However, there has been no systematic topological study in the context of developmental atlases.

Current methods for developmental trajectory inference often assume a tree-like topology (Farrell et al., 2018; Setty et al., 2016; Briggs et al., 2018), or use non-linear dimensionality reduction techniques for analysis (Packer et al., 2019). These choices may influence the misinterpretation of developmental trajectories by distorting the topology of data (see (Wattenberg, Viégas, and Johnson, 2016) for t-SNE), instead of directly inferring the topology in an unsupervised way. Therefore, it has remained unclear to what extent complex topological structures are present in the developmental transcriptome of different organisms.

In this study, we introduce a computational framework that uses algebraic topology to decode the global structure of developmental atlases. By leveraging simplicial homology groups and their as-

sociated Betti numbers we developed a topological classification for cell differentiation trajectories. We specifically focused on characterizing closed loop trajectories, as these understudied features could represent crucial functions in development.

To study topological loops in developmental atlases, we developed Trajectory Outlining for Topological Loops in Single-cell data (`totopos`), a suite of tools that enable: 1) identifying the topology of a single-cell dataset, 2) retrieving cells that belong to a system containing a closed loop trajectory (`topoCells`), 3) identifying critical genes that drive the loop topological signature (`topoGenes`), and 4) building pseudotime coordinates for loop topologies. The foundation of our methodology is persistent homology (PH) (Methods, Box 2), a multiscale approach that analyzes the dynamics of how topological loops appear and cease at different scales.

We first validated our computational framework by demonstrating its ability to identify closed loops in simulated scRNA-seq datasets that exhibit cyclic dynamics and a cell cycle dataset. Building on this validation, we screened over ten transcriptomic atlases from different species, uncovering the presence of loops in the development of multiple metazoan lineages. These findings challenge traditional developmental models based on tree-like branching structures composed of discrete cell fate decisions.

Through extensive analysis of single-cell developmental atlases, we identified a small set of topological building blocks that serve as fundamental architectural elements of cell differentiation processes. We refer to these building blocks as "topological motifs" (Figure 2.2 C). These motifs are characterized by unique combinations of the β_0 and β_1 Betti numbers and include structures such as clusters, linear trajectories, trees, and loops. Remarkably, we found that all analyzed single-cell developmental atlases can be classified according to these topological motifs or their combinations (Fig 2.3). This suggests the existence of a universal topological code underlying cellular differentiation programs. Furthermore, we demonstrate how these topological motifs are connected with different classes of gene regulatory dynamics, providing a global theory of cell state control.

In addition, we identified specific gene modules that drive the dynamics associated with these topological features. For instance, we present evidence suggesting that seam cells maintain their stemness through a gene expression topological loop in *C. elegans*. A topological loop in the developing human immune system bridges the erythroid and myeloid lineages via erythroblastic macrophages and a population of promyelocytes in the fetal liver with erythroid signatures. Finally, we identified a convergent differentiation topology conserved across three cnidarian species, indicating that topological gene expression features can be stable across evolutionary time.

Altogether, our results show that topological loops are present in single-cell developmental atlases and can uncover and describe mechanisms to maintain homeostatic differentiation, stemness, and regeneration. Our framework enables the identification of such loops from scRNA-seq datasets along with the characterization of the associated gene expression programs. By integrating topological and gene regulatory analyses we provide a new approach for understanding the principles governing cell state transitions across biological systems.

Box 1. Introduction to algebraic topology. In this section, we introduce the theoretical concepts of the `totopos` framework.

- Topological space: A mathematical structure that abstracts away the notion of distance and instead focuses on *proximity*. A topological space captures essential features of nearness by a collection of "open sets" — the topology. We will not provide a formal definition, but interested readers can refer to (Hatcher, 2001). Rather, we will think of a topological space as a device that enables us to quantify the shape and connectivity of single-cell transcriptomes.

- Manifold: An $n-$dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ is a topological space that locally *looks like* $\mathbb{R}^n$. For example, smooth curves and surfaces are 1 and $2-$D manifolds, respectively. Manifolds generalize Euclidean space: they can be curved, and importantly for our

discussion, they can have non-trivial topology or shape. Some canonical 2-dimensional manifolds are the sphere, the torus, and the Klein bottle.

- Simplex: An $n-$dimensional simplex is generalization of a polyhedron consisting of $n+1$ vertices (Fig 2.1 A). For instance, a 0-simplex is a point, 1-simplex is a line, a 2-simplex is a triangle and so forth. The dimension of a simplex is given by $\dim \sigma = |\sigma| - 1$. We will denote an n-simplex as a tuple $\sigma = [v_0, v_1, ..., v_n]$ where the ordering of vertices matters for defining orientation. Simplices are the building blocks of discrete versions of manifolds called simplicial complexes.

- Abstract simplicial complex (ASC): Abstract simplicial complexes are, intuitively, a computational scaffold for calculating the homology of manifolds. An ASC K over a vertex set V is a collection of subsets of V, that is closed under the subset relation—if a simplex σ is in K and $\tau \subset \sigma$ (τ is a subset of σ) then τ is in K (Fig 2.2, A). The dimension of K is just the maximal dimension of its simplex, $\dim K = \max_{\sigma \in K} \dim \sigma$. This simple algebraic definition provides a computational model for studying the topology of manifolds.

- Chains: An n-chain is a linear combination of n-simplices. The concept of chains allows us to *algebrize* simplices. For example, adding multiple $1-$simplices forms a path which is a $1-$chain. The power of the concept of chains is that under this definition, the space of n-chains C_n forms a commutative group, allowing us to leverage the rich algebraic properties developed in Group Theory.

- Boundary map: A boundary map can be intuitively understood as a manual on how to *glue* together simplices of consecutive dimensions. For each dimension, the boundary map is a linear map $\partial_n : C_n \to C_{n-1}$, with the property that $\partial_{n-1} \circ \partial_n = 0$. That is, applying two consecutive boundary mappings annihilates the geometric object. For

example, the boundary of a solid 3-ball is a sphere $\partial_3(B^3) = \mathbb{S}^2$, and a sphere has no boundary: $\partial_2\partial_3(B^3) = \partial_2(\mathbb{S}^2) = \emptyset$.

- Chain complex: The sequence of chain groups C_n with decreasing geometric dimension connected by corresponding boundary maps $\partial_n : C_n \to C_{n-1}$ (Fig 2.1 A).

- Homology group H_n: the homology group is the space of equivalence classes of $n-$dimensional holes, which are not boundaries of $n+1$ dimensional simplices. Mathematically the $n-$homology group is $H_n = \ker \partial_n / \text{im}\, \partial_{n+1}$. Two members $x, y \in \ker \partial_n$ are equivalent if $x - y \in \text{im}\, \partial_{n+1}$.

- Betti number(s): the $n-$th Betti number, denoted β_n is the dimension of the $n-$homology group, i.e. $\beta_n = \dim H_n = \dim \ker \partial_n - \dim \text{im}\, \partial_{n+1}$. Note that there is a Betti number per dimension: β_0 encodes the number of connected components, β_1 is the number of loops, β_2 is the number of cavities, and so forth.

- `topoCells`: Let X be a $C \times G$ scRNA-seq matrix with a closed loop trajectory $\beta_0(X) = 1, \beta_1(X) = 1$. The `topoCells` are then defined as the subset of cells $I \subseteq \{1, \ldots, C\}$ with the following condition: the subset I preserves the loop structure: $\beta_1(X[I]) = 1$ (Fig 2.1 C, top).

- `topoGenes`: Given a scRNA-seq matrix defined as above, the topoGenes are defined as the subset of genes that drive the topological loop. Mathematically, they are a subset $J \subseteq \{1, \ldots, G\}$ such that $\beta_1(X[:, -J]) = 0$ (Fig 2.1 C, bottom).

Simplicial homology computation.

Our main goal is to show how to compute homology groups and the corresponding Betti numbers. Here we assume for simplicity that an abstract simplicial complex is available to us. In Box 2 we address the scenario of inferring the simplicial complex from data. Let us

start the discussion by decomposing a 3-simplex $t_1 = [v_0, v_1, v_2, v_3]$, which would correspond geometrically to a tetrahedron. We sort $n-$simplices by the lexicographic order of their vertices as a convention. For instance, since the tetrahedron has four faces or triangles, we can directly write down the $2-$simplices: $f_1 = [v_0, v_1, v_2]$, $f_2 = [v_0, v_1, v_3]$, $f_3 = [v_0, v_2, v_3]$, $f_4 = [v_1, v_2, v_3]$. The list of faces (triangles) is naturally a basis for C_2, the group of $2-$chains. Similarly the $1-$simplices (edges) will be $e_1 = [v_0, v_1]$, $e_2 = [v_0, v_2]$, $e_3 = [v_0, v_3]$, $e_4 = [v_1, v_2]$, $e_5 = [v_1, v_3]$, $e_6 = [v_2, v_3]$. The list of edges forms a basis for C_1. In Fig. 2.1 A we depict the tetrahedron's chain complex, which shows that this algebraic structure is a sequential decomposition into its fundamental geometric building blocks. Please note that in the step $\partial_2 : C_2 \to C_1$ we zoomed in into a single face or triangle f_2, to avoid cluttering.

We can compute the boundary of the tetrahedron $\partial_3(t_1)$ using the definition: $\partial_n([v_0, v_1, ..., v_n]) = \sum_{i=0}^{n} [v_0, ..., \hat{v_i}, ..., v_n]$ where $\hat{v_i}$ denotes deleting vertex i from the simplex σ.

$$\partial_3(t_1) = \partial_3([v_0, v_1, v_2, v_3]) \tag{2.1}$$

$$:= [v_1, v_2, v_3] - [v_0, v_2, v_3] + [v_0, v_1, v_3] - [v_0, v_1, v_2] \tag{2.2}$$

$$= f_4 - f_3 + f_2 - f_1 \tag{2.3}$$

The matrix for ∂_3 is in Fig 2.1 A. Similarly, one computes $\partial_2(f_2)$ by applying the formula directly $\partial_2(f_2) = \partial_2([v_0, v_1, v_3]) = [v_1, v_3] - [v_0, v_3] + [v_0, v_1] = e_5 - e_3 + e_1$. We highlight the values of this operation in a grey box in matrix ∂_2 (Fig 2.1 A, bottom). Finally, the matrix for ∂_1 can easily be assembled, since the formula for the boundary of edges is just $v_e - v_i$ where v_e is the endpoint and v_i is the initial point of the directed edge. The interested reader can confirm that multiplying $\partial_2\partial_3 = 0$, and $\partial_1\partial_2 = 0$. It turns out that since the tetrahedron is topologically equivalent to the solid $3d$ ball, its Betti numbers are trivial: $\beta_0 = 1, \beta_i = 0$ for all other dimensions i. Because of this, we introduce the example of Fig 2.1 B, where one can see there is one $1-$dimensional hole. We can directly compute the $1-$homology group to

confirm this geometric intuition. A basis for im∂_2 is just $\{\partial_2([v_0, v_1, v_2]), \partial_2([v_2, v_3, v_4])\}$, i.e. the image of the two triangles in the complex. The kernel of a linear map A (denoted kerA) is a subspace of the domain that gets mapped to zero. For a boundary map, this means all elements $x \in \ker\partial_n$ have the property that $\partial_n(x) = 0$. In one dimension, this has a very nice geometrical interpretation: a path that returns to the initial point (i.e. a cycle) will be in $\ker\partial_1$. Furthermore, by the algebraic property $\partial_1\partial_2 = 0$, the loops $e_1 + e_3 - e_2$ and $e_5 + e_7 - e_6$ will automatically be in $\ker\partial_1$. However, there is another member in the kernel, the loop $y = -e_3 + e_4 - e_5$ (Fig 2.1 B). Indeed, we can confirm y is in the kernel: $\partial_1(p) = \partial_1(-[v_1, v_2] + [v_1, v_3,] - [v_2, v_3]) = -(v_2 - v_1) + (v_3 - v_1) - (v_3 - v_2) = 0$, thus

$$H_1 = \ker\partial_1/\mathrm{im}\partial_2 \tag{2.4}$$

$$= \frac{\langle e_1 + e_3 - e_2, e_7 - e_6 + e_5, -e_3 + e_4 - e_5\rangle}{\langle e_1 + e_3 - e_2, e_7 - e_6 + e_5\rangle} \tag{2.5}$$

$$\sim \langle -e_3 + e_4 - e_5\rangle \tag{2.6}$$

which corresponds to the loop y in Fig 2.1 B. Thus, we've shown that we can algebraically extract the topological loops using this theory. Moreover, since a homology group is a quotient group, it consists of equivalence classes. For homology, the equivalence is the following: x is equivalent to y (denoted $x \sim y$) if its difference lies in im∂ (think of it as a *signed* set difference). For instance, the loop $x = e_1 + e_4 - e_5 - e_2$ (Fig 2.1, orange) is equivalent to y, since $x - y = e_1 + e_3 - e_2 = \partial_2([v_0, v_1, v_2])$, i.e. $x - y \in \mathrm{im}\partial_2$ and thus $x \sim y$. Since H_1 has only one generator, this implies that $\beta_1 = 1$. The curious reader could confirm that $\beta_0 = 1$ through a process similar to Gaussian elimination, respecting the integer coefficients of the matrices ∂_1 and noting that $\partial_0 = 0$.

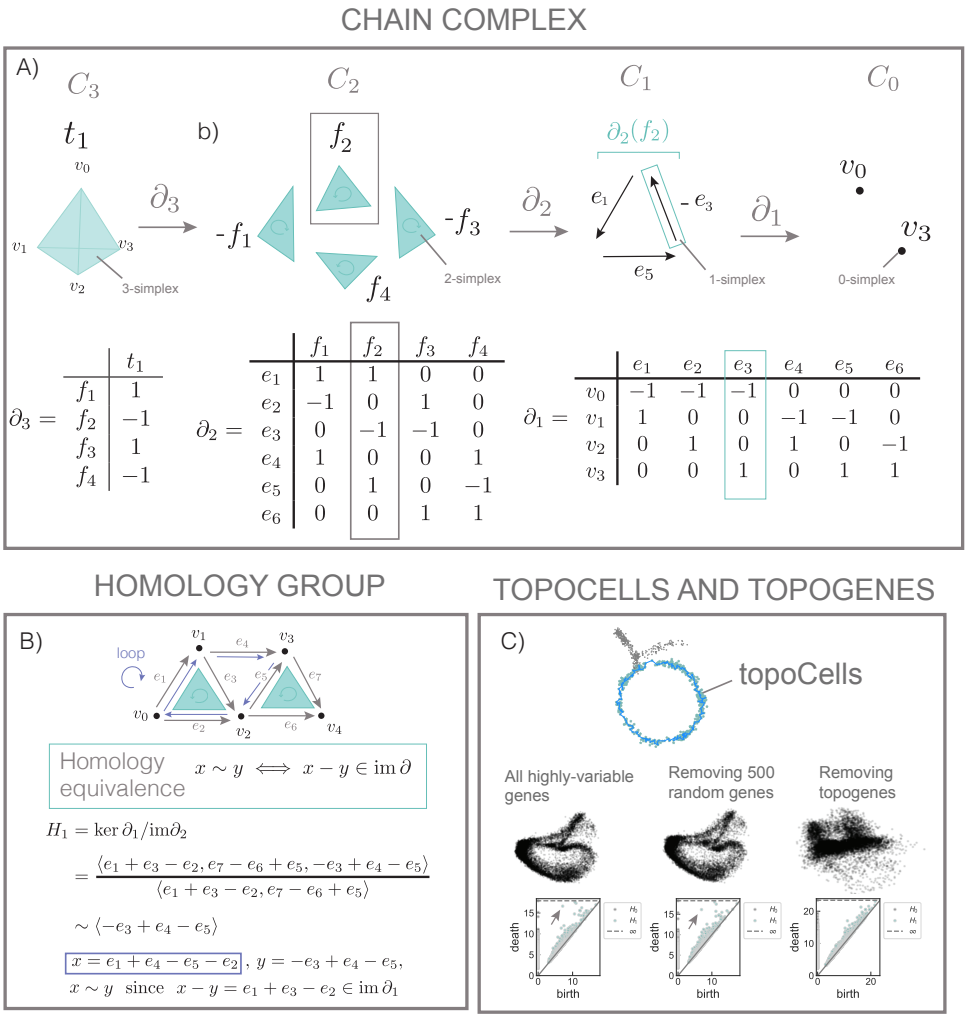

Figure 2.1: Introduction to simplicial homology. (See Box 1 for reference.)

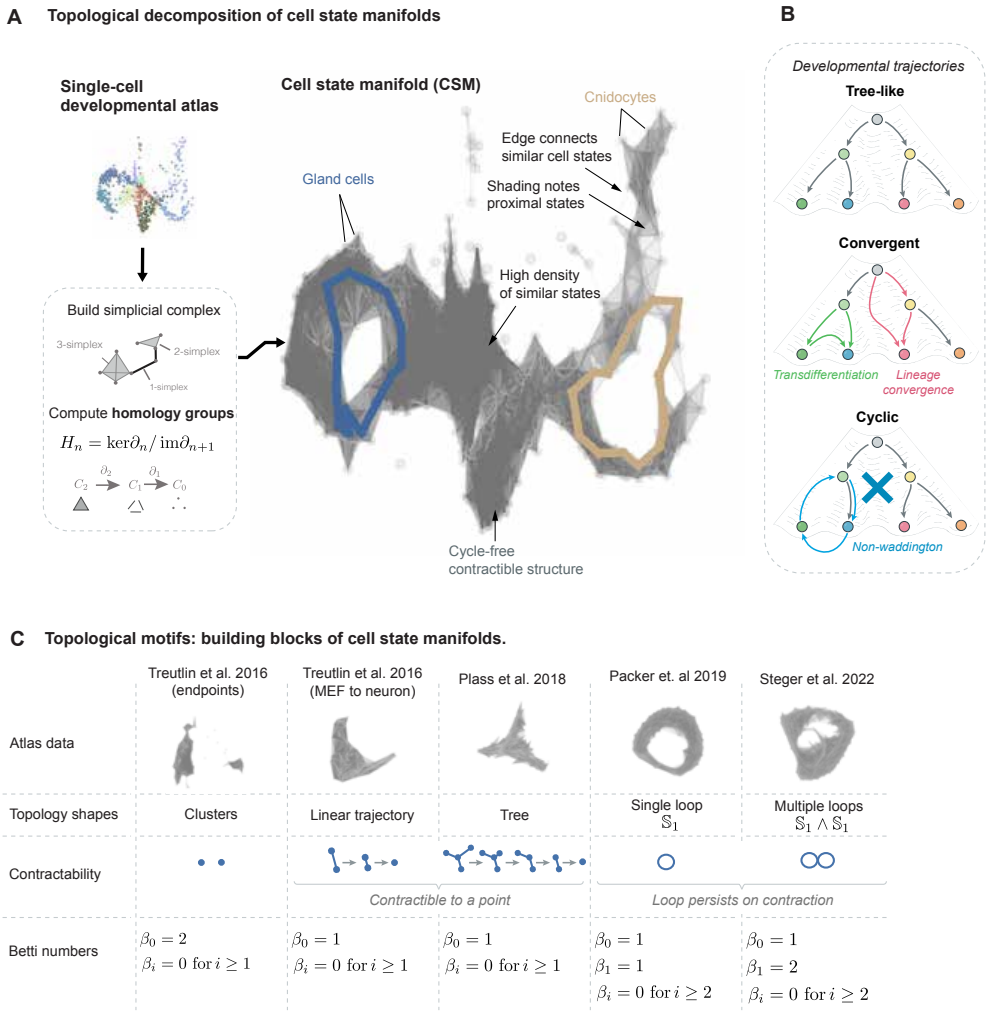

Figure 2.2: **Topological decomposition of cell state manifolds into fundamental building blocks.** (A) Cell state manifolds can be modeled using simplicial complexes constructed based on the proximity between single-cell transcriptome profiles. These manifolds can be decomposed into fundamental topological structures. These include contractible features such as clusters, linear trajectories, and trees, as well as non-contractible features like loops. The presence of loops can be identified through the homology group H_1. We visualize two representatives of the H_1 homology group in blue and golden. (B) Cell differentiation trajectories can be tree-like or contain loops. Topological loops can be qualitatively categorized as either convergent or cyclic. Crucially, cyclic trajectories represent dynamical processes that a Waddington landscape model of development cannot capture. (C) A comprehensive analysis revealed that cell state manifolds can be classified into topological motifs. These manifolds exhibit the topology of tree-like structures or wedge sums of circles, which can be quantified by 0 and 1-dimensional Betti numbers. We present examples of these topological motifs derived from real single-cell atlas data.

2.2 Results

A topological framework for single-cell developmental trajectory analysis

To investigate gene expression profiles forming topological loops along development we developed Trajectory Outlining for Topological Loops in Single-cell data (`totopos`). At the core of our approach is the construction of a Vietoris-Rips (VR) simplicial complex $K_\varepsilon = VR(X, \varepsilon)$ from a scRNA-seq data set X, where cells are connected based on transcriptional similarity up to a distance of ε (Box 2.2, Methods). The VR complex serves as a computational scaffold that captures the relationships between cell states through its chain complex (Box 2.2). We refer to this mathematical structure as the cell state manifold, representing all cell state transitions sampled in the timecourse experiment.

From this construction, we compute topological invariants called homology groups H_n, which encode n-dimensional holes in a cell state manifold (Box 1, Methods). Their ranks, or number of independent generators, are the Betti numbers $\beta_n = \mathrm{rk}\, H_n$, and quantify these holes. We focus on the zero- and one-dimensional homology groups (H_0, H_1) and Betti numbers (β_0, β_1), representing path-connected components and topological loops, respectively. Loops can be either convergent or cyclic (Figure 2.2 B). We leverage temporal information and biological knowledge to distinguish between these two classes of loops.

To robustly identify topological features from cell state manifolds, we employ persistent homology (PH), a method that examines topological features at increasing distance thresholds between cell states. To calculate PH, one first constructs a filtration, which is a sequence of nested simplicial complexes $\mathcal{K} = K_{\varepsilon_1} \subset K_{\varepsilon_2} \subset \ldots \subset K_{\varepsilon_l}$ that tracks the topology across different distance scales ε_i (Box 2, Methods). As the distance threshold varies, topological features like loops may appear or disappear. The lifetime or *persistence* of a homology class γ_i is measured as $\mathrm{pers}(\gamma) = \varepsilon_d - \varepsilon_b$, where ε_b and ε_d are the distance scales at which the feature is born and dies, respectively. These birth-death pairs are collected in a persistence diagram $\mathrm{Dgm}_n(X) = [b_i, d_i)_{i=1,..,k}$ that fully characterizes the n-dimensional topology of the data (Fig 2.7 A) (S. Y. Oudot, 2015). Crucially, PH enjoys a stability property: small perturbations to the input data result in small deviations between

persistence diagrams, and thus the corresponding detected topological features (Cohen-Steiner, Edelsbrunner, and Harer, 2007). We used Ripser (Bauer, 2021; Tralie, Saul, and Bar-On, 2018) for PH computation.

Let us provide some crucial definitions. Consider a single-cell dataset X containing C cells and G genes ($C \times G$ matrix), with a topological loop, i.e. with $\beta_1(X) = 1$. Let $I \subseteq \{1, 2, .., C\}$ and $J \subset \{1, 2, ..., G\}$. We define `topoCells` as a subset of cells that form a 1D loop:

$$\texttt{topoCells} = \{I \subseteq \{1, 2, ..., C\} \mid \beta_1(X[I]) = 1\}$$

,

where $X[I]$ denotes the matrix obtained by selecting the rows corresponding to that subset of cells.

Similarly, we define `topoGenes` as the set of genes J that, when removed, results in a trivial topology:

$$\texttt{topoGenes} = \{J \subseteq \{1, 2, ..., G\} \mid \beta_1(X[:, -J]) = 0\}$$

,

where $X[:, -J]$ represents the dataset with columns J removed. Note that both definitions can be more broadly defined for an $n-$dimensional Betti number. In practice, we use a set of highly variable genes and $\text{card}(J) = 500$ in our analyses. Importantly, both `topoCells` and `topoGenes` may not be unique. Furthermore, we perform Principal Component Analysis (PCA) with $n = 20$ components before computing PH to avoid the curse of dimensionality (Hiraoka et al., 2024). We explored the effect of dimension in topological inference in SI Note 2.6 and Figure S3.

An approach for identifying `topoCells` is to compute the cells in the neighborhood of a homology generator. Therefore, from this perspective, the key step for `topoCells` computation is identifying a representative for a target homology class. Ripser (Bauer, 2021) can return representatives of persistent cohomology classes, which look geometrically very different from homology representatives. In particular, taking the neighborhood around a cohomology class doesn't correspond to

`topoCells` (See Box 2.7). However, it gives crucial topological information: a cohomology class represents an obstruction to a homology class. Inspired by the cohomology algorithm of Dlotko (Dłotko, 2012), we developed an algorithm that uses this birth critical edge as a means to recover the corresponding homology class. Our algorithm proceeds as follows: Assume we have access to the birth and death time of a target persistent homology class. Build a spanning tree T from the 1-skeleton of $K = VR(X, \varepsilon_b)$ starting from one of the bounding points of the critical edge e. Adding e to T results in a cycle γ, and we claim that this cycle γ embeds as a representative of the homology class in K (SI).

To identify `topoGenes`, we developed a method that ranks genes based on their influence on a persistent homology class. In brief, our approach builds upon recent advances on topological optimization (Nigmetov and Morozov, 2022). To define the ranking score we computed gradients $\partial\mathcal{L}/\partial X$ of a loss function $\mathcal{L}$ that represents a perturbation to the input data that ablates the homology class (Methods). The score for each gene is then defined as the norm of its corresponding column in the gradient matrix $\partial\mathcal{L}/\partial X$. Intuitively, this method allowed us to identify subspaces of genes that contain the `topoCells`. Our key contribution is the observation that the simplices required to move the homology class from $(b, d) \mapsto (d, d)$ (i.e. the critical set as defined in (Nigmetov and Morozov, 2022)), are identical to the edges from the representative cocycle available from Ripser (SI). The intuition is that if we want to increase the birth time of a homology class we would need to expand the length of the critical birth edges in the corresponding cohomology generator (Fig 2.7 E). This finding results in a remarkable speedup for the gradient computation, and thus for the topological ranking.

Finally, we also define pseudotime methods for the convergent and cyclic topologies (Fig 2.2). The convergent pseudotime is defined as the geodesic distance from an initial cell state on the 1-skeleton of $VR(X, \varepsilon_b)$, where ε_b is the birth time of a target homology class. For the cyclic pseudotime we use the Toroidal Coordinates algorithm (Scoccola et al., 2023), which provides a map from the simplicial complex to the circle. In practice, we compute pseudotime coordinates on `topoCells`.

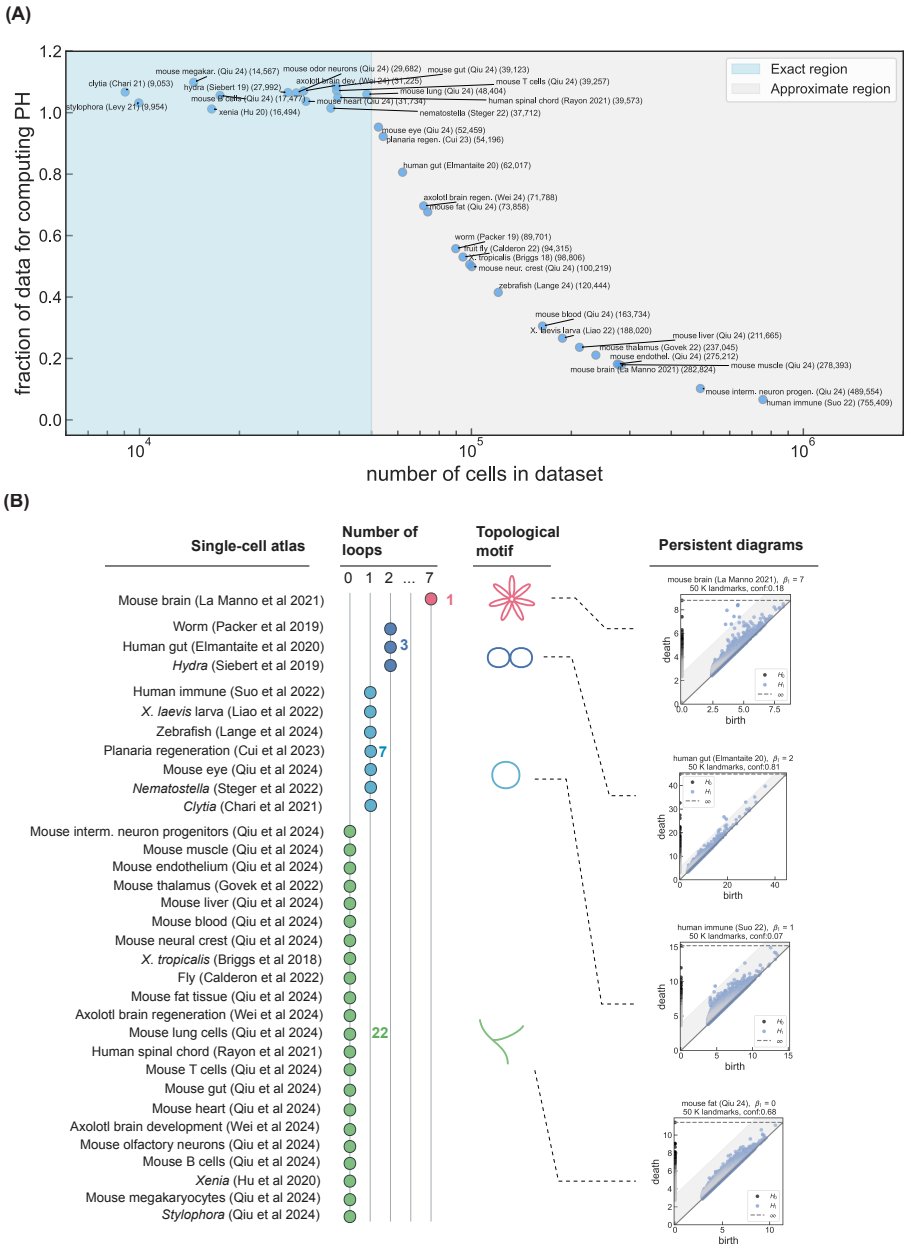

Figure 2.3: **Topological census of developmental cell state manifolds across metazoans can be classified into a small set of topological motifs.** (A) We performed a topological census by computing persistent homology on a broad set of single-cell developmental atlases across metazoans. Computational limitations allowed to calculate PH on at most 50,000 cells – we highlight the exact or confident region and the approximate region respectively. The approximate region was computed using furthest point sampling (Methods). We added jitter to datasets in the exact region for visualization purposes. (B) Left: Census of 33 published single-cell developmental atlases, sorted and color-coded by their usage of topological motifs, indicating the number of loops detected. Human and mouse datasets were analyzed as independent systems due to their scale. Right: Persistence diagrams for the selected datasets. Neighborhood threshold, is highlighted in light grey. Black arrows in persistence diagrams indicate prominent H_1 homology classes.

Topological analysis of cell cycle dynamics

The cell cycle is the most universal oscillatory process of biological systems. In their landmark study, (Kowalczyk et al., 2015) showed that cell cycle dynamics constitutes one of the primary sources of heterogeneity of the transcriptome. Subsequent studies have further investigated this phenomenon using single-cell genomics in different model organisms, and multiple computational approaches have been developed to identify the cell cycle signature. Of particular relevance, Schwabe et al. (Schwabe et al., 2020) elegantly demonstrated that proliferating HeLa cells manifest a cyclic signature in gene expression space that can be reconstructed using matrix factorization methods. From a topological perspective, their finding implies that even unsynchronized proliferating cells can form an H_1 homology group in gene expression space.

To validate our framework, we applied `totopos` to Schwabe's dataset, to see if we could identify the reported H_1 homology group. Employing our pipeline using $n = 20$ principal components (PCs) rendered a trivial topology. However, using a low-dimensional model ($n = 5$ PCs) we identified a prominent H_1 homology class that accurately captured the cell cycle trajectory (Fig S6 A), suggesting that the cell cycle signature is embedded in the most variable gene subspace, and spirals when considering other biological processes, consistent with previous studies (Schwabe et al., 2020; Kowalczyk et al., 2015). Moreover, using the topological gradients method, we identified `topoGenes` enriched for cell cycle functions including DNA replication, mitotic spindle assembly, and chromosome segregation. Pathway analysis of these `topoGenes` revealed Reactome terms related to cell division (P-value $< 10^{-6}$), confirming the biological relevance of our topological method. These findings demonstrate that `totopos` effectively captured the topological signature of cyclic biological processes, offering a rigorous framework for analyzing cyclic gene expression patterns. This application established a foundation for exploring more complex topological structures that may govern developmental trajectories and cell fate decisions.

Developmental cell state manifolds can be constructed using a small set of topological motifs

To systematically characterize the topology of development, we conducted a topological census by computing the first two Betti numbers across more than 15 developmental atlases spanning a diverse set of metazoans (Fig 2.3, SI Table 1). These topological invariants precisely measure manifold structure: β_0 counts the number of connected components and β_1 measures the number of loops. Our compendium comprises more than 4 million single-cell transcriptomes from model organisms (e.g. *C. elegans* (Packer et al., 2019), *M. musculus* (Qiu et al., 2024), and *D. rerio* (Lange et al., 2023)), four cnidarian species, mouse neuronal, and human immune (Suo et al., 2022a) lineages. Each dataset we analyzed contains densely sampled developmental trajectories with high temporal resolutions.

We first investigated the connectivity of cell state manifolds through H_0 persistent homology analysis. We developed a method to retrieve the data's most parsimonious β_0 by identifying the largest gap between consecutive persistence values (Methods). We benchmarked this approach by simulating clusters in high-dimensional spaces (SI). We found that most cell state manifolds are predominantly path-connected along development ($\beta_0 = 1$) (Fig 2.3). These results suggest that differentiation proceeds through continuous trajectories from progenitor to mature states in gene expression space.

We then assessed the presence of loops in developmental trajectories by analyzing 1-dimensional homology (H_1) across our curated cell atlas compendium. This analysis required different computational approaches based on dataset size, since PH computation requires more memory with both increasing number of points and homological dimension. For smaller datasets ($n < 50,000$ cells), we computed H_1 persistence diagrams exactly. For larger datasets, we employed farthest point sampling (FPS), which has computable error bounds through the persistence stability theorem (Chazal, Cohen-Steiner, et al., 2009). We provide a detailed discussion of this methodology in the SI.

Our analysis revealed that while most developmental atlases (67%) exhibited the expected tree-

like topology ($\beta_1 = 0$), and a substantial proportion (33%) contained one or more persistent loops ($\beta_1 \geq 1$) (Fig. 2.3A). This unexpected prevalence of non-tree topologies suggests that developmental manifolds employ a richer repertoire of topological structures than previously recognized. To the best of our knowledge, this is the first systematic and unbiased discovery of topological loops in developmental cell atlases reported in the literature.

Loops in developmental manifolds can be classified into two types: cyclic—where cells traverse the loop periodically—and convergent—where distinct developmental trajectories merge at a common fate (Figure 2.2 B). Crucially, by analyzing temporal and biological information, we found that almost all loops we identified were of the convergent type. The only instance of a cyclic loop was found in the seam cells of *C. elegans*.

Our systematic analysis of the topology of developmental manifolds revealed an organizing principle: cell state manifolds can be constructed using a few building blocks—clusters, linear trajectories, trees, and loops. We call these building blocks "topological motifs". Topology enables classifying these topological motifs into just three signatures: clusters ($\beta_0 \geq 1$), simply connected manifolds like linear trajectories and trees ($\beta_0 = 1, \beta_1 = 0$), and manifolds with loops ($\beta_1 \geq 1$). In the following sections, we examine how these motifs enable key developmental processes through detailed analyses of three case studies: human immune development, stem-like maintenance in *C. elegans*, and a conserved convergent loop in cnidarians.

Topological loop in early human immune development reveals crosstalk between myeloid and erythroid lineages

The human immune system emerges through an intricate developmental program originating from hematopoietic stem cells (HSCs) in the fetal liver and diverging into over a hundred specialized cell types distributed throughout the body (Laurenti and Göttgens, 2018). Recent advances in single-cell transcriptomics have revolutionized our understanding of hematopoietic lineage specification by revealing previously unrecognized heterogeneity within seemingly homogeneous populations and identifying novel transitional states along differentiation trajectories (Velten et al., 2017; Giladi

et al., 2018). Despite these advances, the topological organization of early human immune cell differentiation remains incompletely understood. We therefore applied our topological analysis to an integrated human cell atlas comprising over 800,000 cells across 9 tissues from post-conception weeks 4-17 (Suo et al., 2022b).

Using `totopos`, we identified several prominent H_1 classes that partially overlap and collectively form a tetrahedral shape when visualized in the first three principal components. Based on the results of our census, we focused on the most persistent H_1 class and extracted its representative loop using the critical edge method. We included all cells within a neighborhood radius equal to the birth filtration value of the class. The resulting `topoCells` form a loop topology with the majority of variation captured in two dimensions (Figure 2.4B).

Based on established cell type classifications (Suo et al., 2022b), two segments of the loop correspond to the megakaryocyte-erythroid lineage and the myeloid lineage, respectively (Figure 2.4A, B). These segments are connected by progenitor cell types with varying developmental potential, ranging from HSCs to common myeloid progenitors (CMPs) and megakaryocyte-erythroid progenitors (MEPs). This arrangement aligns with the classical hierarchical model of hematopoiesis, where HSC-derived cells progressively commit to distinct lineages that culminate in unipotent terminal cell types (Laurenti and Göttgens, 2018; Orkin and Zon, 2008).

Intriguingly, we identified a third segment that connects the myeloid and erythroid branches, completing the loop topology (Figure 2.4A, B). Adjacent to the myeloid segment lies a population of iron-recycling macrophages exhibiting high VCAM1 expression, which mediates cell-cell adhesion and provides survival and differentiation signals to erythroblasts (Chasis and Mohandas, 2008; Klei et al., 2017). This macrophage population extends toward the erythroid segment through erythroblastic island (EI) macrophages and a population of promyelocytes displaying transcriptomic profiles with markers for both myeloid lineage commitment and erythropoietic function (Figure 2.4C). Previous studies have demonstrated that EI macrophages establish direct receptor-ligand interactions with developing erythroid cells (Popescu et al., 2019; Chow et al., 2013), suggesting a

functional bridge between these lineages.

To identify genes responsible for generating this H_1 class, we computed topoGenes using the loop representative sampling strategy (Methods). Based on ablation testing, we identified 1000 topoGenes and applied orthogonal non-negative matrix factorization (ONMF) to cluster them into 20 functionally distinct gene programs (Supplementary Table 1). By computing topology-preserving circular coordinates, we established a natural ordering of `topoCells` and visualized the enrichment patterns of each gene program along the loop (Figure 2.4C). Notably, genes associated with erythroid maturation and hemoglobin/oxygen transport are active not only in the erythroid lineage but also in specific macrophages and promyelocytes (Figure 2.4C). Additionally, genes involved in cell cycle and proliferation are predominantly active in the erythroid lineage and a subset of progenitor cells, indicating that cell cycle variation is not the primary driver of this loop topology.

We complemented our topological analysis with SCENIC (Aibar et al., 2017) to identify transcription factor (TF) regulons enriched in different segments of the loop. Based on the regulon structures, we identified statistically significant ($p < 0.001$) regulatory relationships between TFs and gene programs (Figure 2.4D). This analysis revealed key regulators of erythroid development (GATA1, GFI1B), myeloid differentiation (CEBPB, CEBPD), and progenitor maintenance (ERG, GATA2).

To characterize differentiation dynamics within this topological framework, we computed RNA velocity fields using velocyto (La Manno et al., 2018) and scVelo (Bergen et al., 2020). The overall flow patterns largely recapitulate known biology, with vectors generally aligned with differentiation trajectories and tangential to the loop. However, multiple sources and sinks distributed along the loop and locally incoherent vectors highlight the underlying complexity and stochasticity of the system (Figure 2.4E).

For deeper investigation of the vector field, we implemented Helmholtz-Hodge decomposition of the RNA velocity ((Su, Tong, and Wei, 2024)), which provides a unique orthogonal decomposition

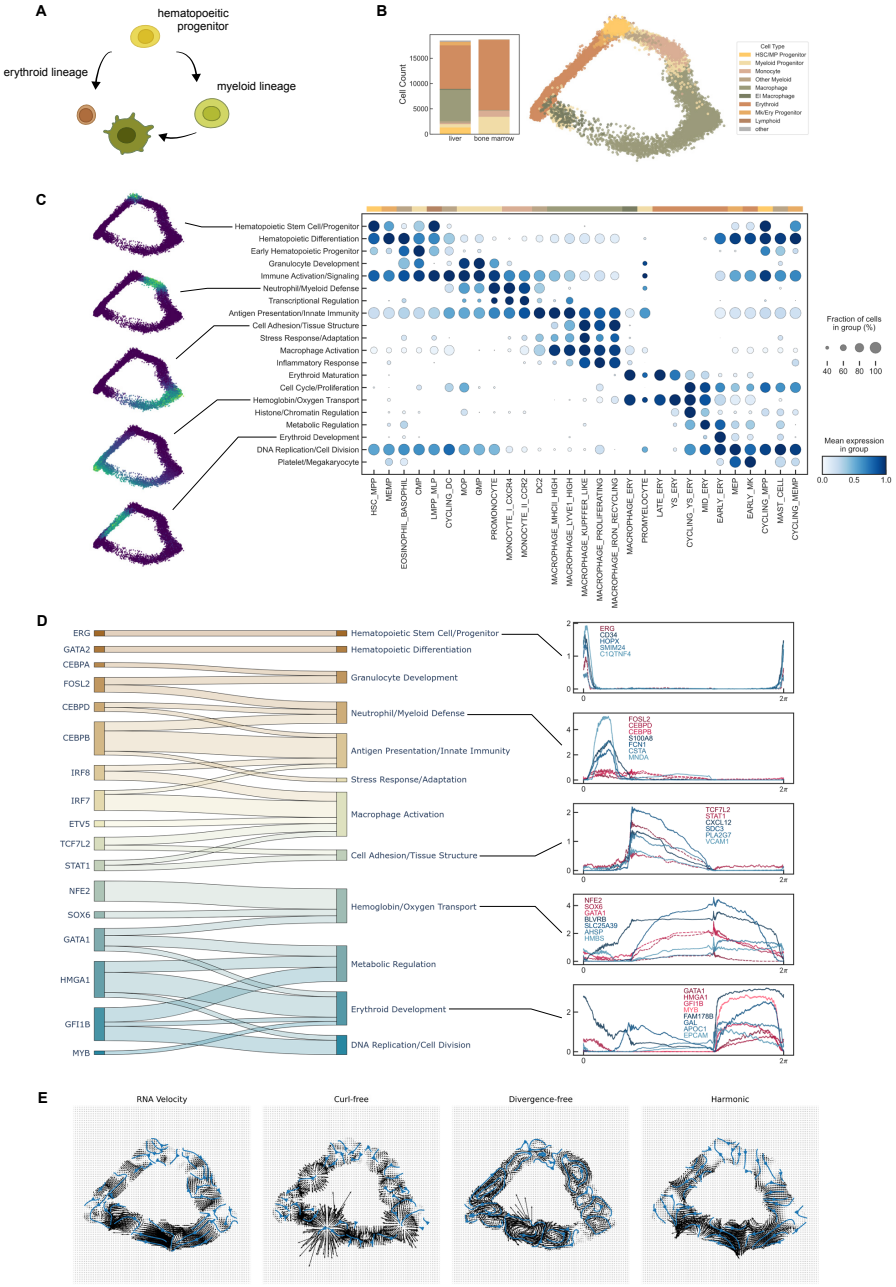

Figure 2.4: **Topological loop in early human immune development highlights macrophage-mediated pathway in erythropoiesis.** (A) Conceptual model of the hematopoietic loop connecting the myeloid and erythroid lineages through a previously overlooked pathway. (B) Cell type distribution of `topoCells` across liver and bone marrow tissues (left). `topoCells` projected onto principal components reveal a loop topology. (C) Gene program enrichment along the loop. Left: `topoCells` colored by enrichment of key gene programs, revealing localization of transcriptional modules along the loop. Right: dotplot of gene program activity across the circular coordinates, with cells grouped by original labels as in (Suo et al., 2022b). (D) Left: transcription factor regulatory network connecting TFs (left column) to gene programs (right column). Only associations with a q-value ≤ 0.001 are shown. Right: expression of key regulators (reds) and top genes of gene programs (blues) in `topoCells` along circular coordinates. (E) Helmholtz-Hodge decomposition of RNA velocity: total velocity field, curl-free (gradient), divergence-free (rotational), and harmonic components. The predominant gradient component (Waddington ratio ≈ 2) supports classical developmental modeling despite the circular topology.

of the vector field into gradient, curl, and harmonic components (Figure 2.4E). The Waddington ratio, measuring the relative magnitudes of gradient and curl components, quantifies how closely a developmental system adheres to Waddington's classical model of development as a gradient descent process (C. H. Waddington, 2014). We found that the Waddington ratio for our `topoCells` is approximately 2, indicating that gradient flows predominate despite the circular topology, and suggesting that the system broadly conforms to Waddington's epigenetic landscape model.

A stem-like fate maintenance circuit drives an H_1 homology class during the embryonic development of *C. elegans*

C. elegans exhibits remarkable developmental invariance, with its embryogenesis mapped cell-by-cell in classical studies by Sulston and Horvitz (Sulston et al., 1983). The recently developed *C. elegans* developmental atlas (Packer et al., 2019) provides single-cell resolution data covering embryogenesis from early cleavage events up to the onset of larval L1 stage (Fig. 2.5 A). Leveraging this dataset, we applied persistent homology and identified a prominent 1-dimensional (H_1) homology class forming a loop structure. Cell-type annotations revealed this loop predominantly comprises seam cells, a group of lateral hypodermal stem-like cells, and hypodermal cells (Fig. 2.5 B). The identified seam cell loop persists from early gastrulation (≈200 min after cleavage) until the early L1 larval stage (720 min). Furthermore, persistent homology analysis of cell cycle genes alone showed no prominent loop structure, indicating the identified loop is not solely driven by cell-cycle-related expression dynamics. Visualizing the dynamics of the cell state manifold indicated a cyclic loop topology, whereby the cell system returns to a gastrula-stage cell state around the 12 hours of embryonic development (Fig 2.5). Such a cyclic trajectory challenges the traditional tree-based developmental model.

To elucidate the genetic drivers of the loop structure we applied `topoGene` analysis using the Laplacian eigenvector (LE) method (Methods). In brief, the LE method has the purpose of retrieving oscillatory or transiently active genes and thus more consistent with a cyclic topology. We confirmed that the LE method also revealed `topoGenes` by ablation analysis (Fig S12). We hypothesized

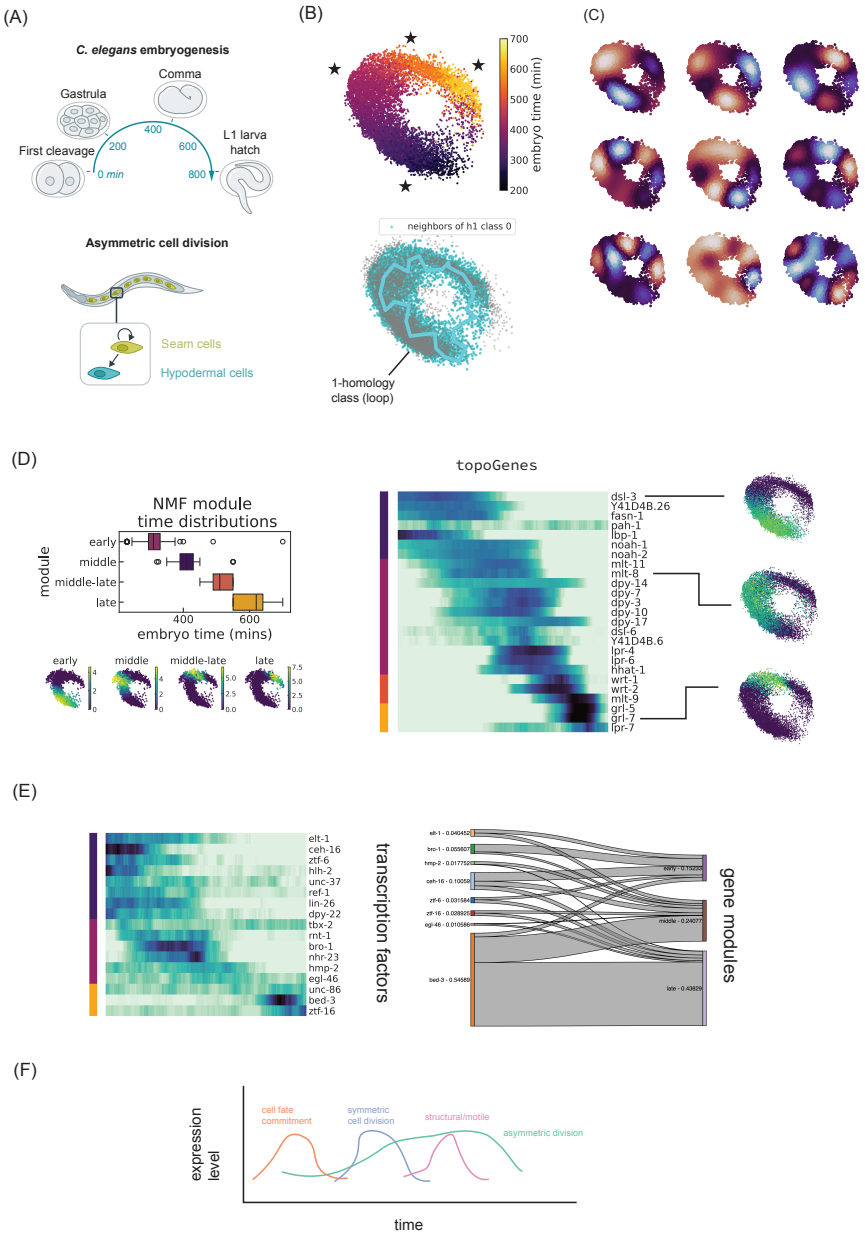

Figure 2.5: **Stemness maintenance is encoded as a cycle in transcriptome space in *C.elegans*' seam cells.** (A) Section of the life cycle profiled by Packer et al. (Packer et al., 2019) (top). Seam cells can divide asymmetrically post-embryogenesis. (B) Seam cells colored by embryo time (minutes) projected onto the first 2 principal components (top), and corresponding homology class (bottom, blue line). (C) Harmonic modes (Laplacian eigenvectors) are used to compute the topology-generating genes (`topogenes`). (D) Left: NMF found clusters of coexpressed genes concordant with the embryo time. Right: `topoGenes` ordered by the first harmonic mode and clustered by NMF. (E) Subset of transcription factors identified using a linear model for predicting chronological `topogene` gene expression.

that the topoGenes played essential roles in regulating and maintaining these cyclic cellular transitions. To investigate this notion, we clustered topoGenes chronologically and performed functional enrichment analysis in each module (Fig 2.5). Our analysis revealed that topoGenes are significantly associated with Notch signaling, and structural and motile processes including molting and cuticle development. Notably, genes such as *dsl-3*, a ligand implicated in restricting developmental plasticity, were prominently expressed during gastrulation, suggesting their role in initial cell fate commitment. In contrast, molting genes (e.g. *mlt-8*, *noah-1*) and structural collagens (e.g. *dpy-7, sqt-3*) were prominently active during elongation and late embryogenesis, indicating their involvement in essential structural hypodermal functions.

Lastly, our computational pipeline allowed the identification of 11 transcription factors (TFs) enriched among the 476 topoGenes. Key TFs included *elt-1* and *elt-3*, known GATA-like regulators of hypodermal differentiation and seam cell division symmetry (Gilleard and McGhee, 2001; Brabin, Appleford, and Woollard, 2011). To further elucidate gene regulatory control, we used a linear model (Methods) that revealed distinct temporal expression waves, with early TFs (e.g. *unc-37, elt-1*) associated with asymmetric cell division control (Horst et al., 2019; Pflugrad et al., 1997; Calvo, 2001) , middle-expressed TFs including *nhr-23* controlling structural genes (Kouns et al., 2011; Meli et al., 2010), and late-expressed TFs (*ztf-16 unc-86*) implicated in neuron fate and asymmetric division (Fig 2.5). These findings suggests that the expression of asymmetric components could be predominantly present in seam cells, and only overthrown by symmetric division TFs such as *elt-1*, and *rnt-1*, consistent with previous findings (Horst et al., 2019). Thus, we propose three core regulatory modules underlying seam cell topology: (1) fate commitment and maintenance, (2) structural and motile functions, and (3) asymmetric division control (Fig. 2.5). Further experimental characterization may reveal additional insights into the precise regulatory logic controlling this cyclic developmental trajectory.

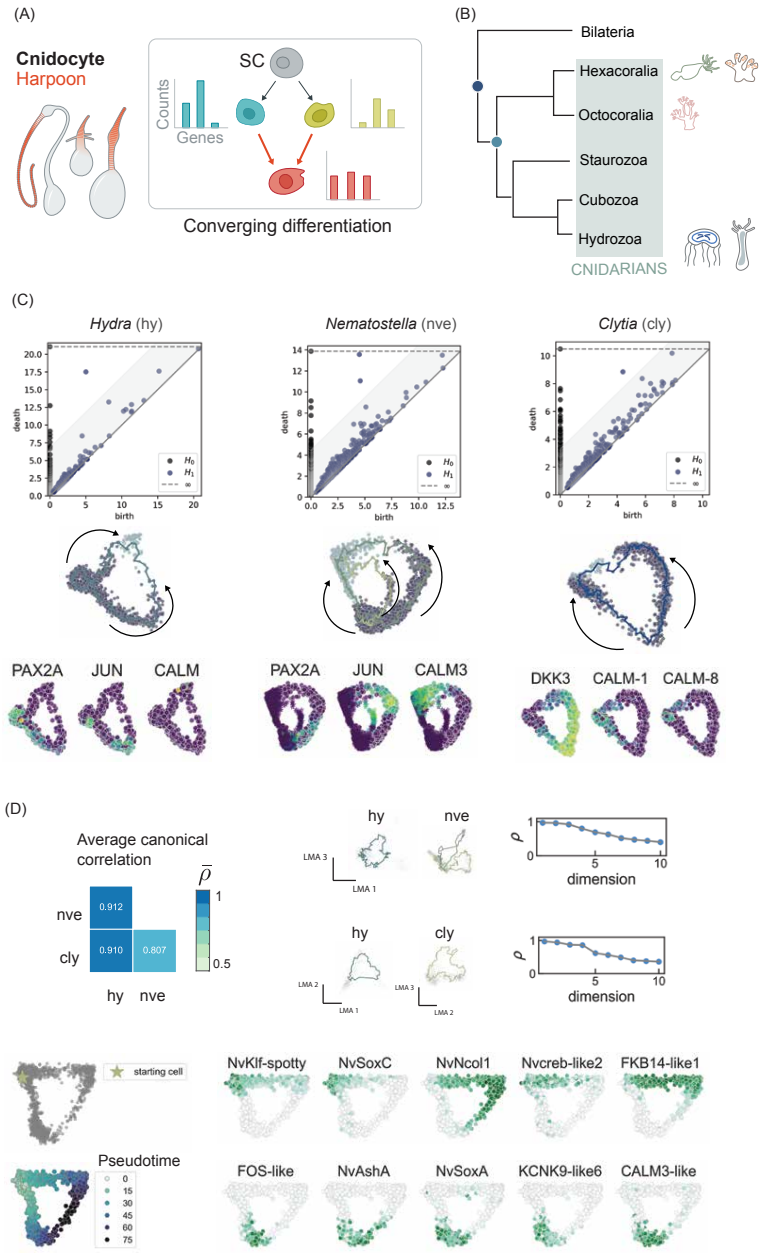

Figure 2.6: **A conserved topological structure in cnidocytes reflects 500 Million years of evolutionary stability.** (A) Schematic diagram of cnidocytes and their development. Cnidocytes are a specialized cell type that can display adhesive, stinging, or ensnaring functions. We found that cnidocyte development is structured as convergent loops, in which a cnidoblast can take alternative paths to generate functional diversity, but all cnidocytes converge in their molecular profile as they all have the same mechanism of ejection. (B) Schematic phylogenetic tree of cnidarians and species analyzed. The split between major cnidarian clades occurred more than 500 Ma (Park et al., 2012). We analyzed the transcriptomes of *Nematostella vectensis*, *Stylophora pistillata* (Hexacoralia), *Xenia sp.* (Octocoralia), *Clytia hemisphaerica*, and *Hydra vulgaris* (Hydrozoa). (C) Persistent diagrams and developmental trajectories of cnidocytes of *Hydra*, *Nematostella*, and *Clytia*. (D) Comparative transcriptomic analysis using linear manifold analysis.

A conserved loop in the clade-defining cell type of cnidarians

Cnidarian organisms are a special clade within the animal kingdom for their capacity for full-body regeneration, and the presence of a primitive neural system. Furthermore, cnidarians hold a key evolutionary position by being the sister group to bilateria. Cnidocytes are the clade-defining cell type of cnidarians, and recent research (Steger et al., 2022) has shown evidence pointing to its origin as a repurposing of a protoneural lineage (Steger et al., 2022). Cnidocytes have a remarkable variety in morphology, but broadly, they can be classified as piercing (nematocytes), ensnaring (spirocyte) or adhesive (ptychocytes) subtypes (Babonis et al., 2023), and they all act through the same mechanism: the activation of chemo- and mechano-sensitive sensors that trigger a discharge of a harpoon-like structure (Fig 2.6 A). These functions represent key cellular innovations that arguably contributed to the evolutionary success of this taxonomic group.

Because of these remarkable biological properties, there has been wide interest in investigating the single-cell transcriptome repertoires of cnidarian species (Fig 2.6 B) (Siebert et al., 2019; Steger et al., 2022; Chari et al., 2021; Sebé-Pedrós, Saudemont, et al., 2018; Hu et al., 2020; S. Levy et al., 2021; Link et al., 2023). Based on recency, sequencing depth and sample size, we analyzed the atlases of *Hydra vulgaris* (Siebert et al., 2019), *Nematostella vectensis* (Steger et al., 2022), *Clytia haemisphaerica* (Chari et al., 2021), *Xenia sp.* (Hu et al., 2020), and *Stylophora pistillata* (S. Levy et al., 2021). By exploiting cell type information, we found that three out of five of the cnidarian manifolds analyzed harbored a loop in their cnidocytes despite their divergence more than 500 million years ago (Park et al., 2012) (Fig 2.6 C, Fig S S14). Specifically, we found one loop in the cnidocytes of *H. vulgaris*, *C. hemisphaerica*, and two loops in *N. vectensis* despite their divergence more than 500 million years ago (Park et al., 2012). We also found a loop in the cnidocytes of *Xenia* but it was only apparent when computing PH on five principal components (Fig S14).

A possible explanation for the difference in Betti numbers of *Hydra*, *Nematostella* and *Clytia* is that all three species have piercing venomous cell subtype, but only *Nematostella* has the ensnaring subtype (Babonis et al., 2023). We found that indeed one of the trajectories in *Nematostella* corresponds to the the ensnaring cell type, consistent with the notion that the ensnaring subtype

emerged as a novel subtype by repurposing machinery of the piercing ancestor (Babonis et al., 2023). Furthermore, we used the cell type information from the original studies as well as conserved gene markers for early (e.g. PAX2A) and late cell states (e.g. Calmodulin) to determine that the loops were of the convergent type (Fig 2.6).

Next, we used the topological information to conduct a comparative transcriptomics analysis. A manifold can be characterized by its local resemblance to euclidean space. In consequence, a cell state manifold should be in principle locally characterized by its principal components (Budninskiy et al., 2019). Under this assumption, one could aim to find shared features between cnidocyte cells of two organisms by calculating the intersection between their invariant subspaces. A recent result from matrix factorization theory (Sørensen, Kanatsoulis, and Sidiropoulos, 2021) showed that performing Canonical Correlation Analysis(CCA) corresponds to finding the intersection between the linear spaces corresponding to the range of data. We exploited these notions to develop Locally Linear Manifold Alignment (LMA), a method that enables finding shared gene subspace by aligning cell type tangent spaces (Methods). LMA works by first finding corresponding subsets of corresponding cells in two datasets using Mutual Nearest Neighbors (MNN) followed by CCA. Importantly, we used the principal components as a tangent space approximation (Budninskiy et al., 2019), a method first reported in (Brown, Bray, and Pachter, 2018) which can also be viewed as a mechanism to avoid overfitting. We applied LMA to all the pairwise combinations between *Clytia*, *Nematostella*, and *Hydra* (from here on cly, nve, hy) to perform pairwise comparative analysis as their datasets contained more than 1000 cnidocyte cells.

In Fig 2.6 we report the average canonical correlations between the analyzed cnidarian species. First, since LMA is designed to find a linear correspondence between datasets, we first tested if LMA preserved the topological information. We found that the pairs Hy-Cly and Hy-Nve preserved the topological structure after alignment, but Nve-Cly did not (Fig 2.6). This result could reflect complex geometrical relationships that cannot be accounted for by the linear objective of LMA. Moreover, this result implies that, we can treat the cnidocyte systems of Hy-Cly and Hy-Nve equivalently, under the LMA map, up to linear approximation error.

To compare the cell states across species, we computed the genes with high absolute weights in both organisms, which we called the LMA genes. Specifically, we focused on transcriptional regulators to compare regulatory states across the loop. For the Hy-Nve comparison we found a cnidocyte development TF signature previously reported in the literature such as the of Kruppel-like zinc Fingers PRDM-6 and PRDM13, CREB, JUN, and AshA (S. Levy et al., 2021; Siebert et al., 2019; Steger et al., 2022). In particular the expression of achaete-scute ortholog AshA on mature cnidocyte states, which has been previously reported as a possible defining factor of neuronal fate, points out the importance of neuron-like machinery recruited for cnidocyte function. Together these results suggest that our analysis can extract interpretable factors determining the topology and in turn cell state control across species.

Another interesting application of LMA of our topological is that it enables to compute topology-preserving pseudotime. This helps to parametrize the geometry of the cell state manifold and investigate broad gene expression patterns at different locations along the loop. For example, a natural question would be to ask what are the differentially expressed genes along the different trajectories of the loop. Given that we hypothesized that the geometry of the loop is convergent, we defined the convergent loop pseudotime as the geodesic distance from a prototypical progenitor cell, an analog of the classic definition. From the pseudotime, one can easily compute the trajectories by clustering cells at intermediate timepoints (Methods). We used Nematostella dataset to showcase this analysis.We show results of genes overexpressed in both trajectories in Fig 2.6. In summary, LMA enabled the comparative analysis of single-cell transcriptomes at the local level using topology as a guiding principle.

Box 2. Persistent homology

In this section, we provide key terms for the persistent homology.

- Filtration : A filtration is a nested sequence of simplicial complexes $\mathcal{K} = K_1 \hookrightarrow K_2 \hookrightarrow \ldots \hookrightarrow K_m$, i.e. at each index i, the simplicial complex K_i is a subcomplex of K_{i+1}.

- Simplex-wise filtration. A filtration with the property that, at each step we add a single simplex is called a simplex-wise filtration.

- Simplicial map: A map between simplicial complexes $f : K \to L$ with the property that a simplex σ in K is mapped to a simplex $f(\sigma)$ in L, is called a simplicial map.

- Induced maps: Given a simplicial map between abstract simplicial complexes $f : K \to L$, these maps extend linearly to chains, i.e. for a chain $c = \sum_i a_i \sigma_i$ the induced map is $f(c) = \sum_i a_i f(\sigma_i)$. Note that f sends cycles to cycles and boundaries to boundaries. Thus, f also induces a map on homology groups: a homology class $[\gamma] \in H_n(K)$ is mapped to $f_*([\gamma]) = [f(\gamma)]$ in $H_n(L)$. f_* is also called the pushforward of f.

- Persistence module: Given a filtration of Vietoris Rips complexes $\mathcal{K}$, one can study the sequence $H_n(K_1) \to H_n(K_2) \to ...H_n(K_m)$, called the persistence module.

- Persistent homology groups: Given a VR complex filtration, the homology classes that persist from index i to index $j > i$ are called the persistent homology groups, and are defined as the image of the pushforward of the inclusion map $H_n^{i,j} = \mathrm{im}\iota_*^{i,j}$. If a homology class is born at index b and dies–i.e. merges with a previous class– at index d we say that its persistence or lifetime is $\mathrm{pers}([\gamma]) = d - b$.

- Persistence diagram. At a given homological dimension n, the set of births and deaths of all homology class in the filtration is called the persistence diagram $\mathrm{Dgm}_n = \{[b_i, d_i)\}_{i=1,...,k}$. The persistence diagram completely characterizes the persistence module (up to reordering of the barcodes) (Gabriel, 1972; S. Y. Oudot, 2015). Persistence diagrams can be visualized as a 2D scatter plot on the plane, where points lie above the diagonal.

Box 2 (continuation). Persistent (co)homology using the critical edge method

The goal for this section is to illustrate the ideas for computing the 1−homology groups using our approach, the critical edge method. Particularly, we want to focus on the intuition. The technique is simple, but it requires some machinery. Let us begin by introducing cohomology.

Introducing cohomology

We define the n-cochain group as the dual of the n-chain group, i.e. $C^n = \mathrm{Hom}(C_n, G)$, or the space of functions that are group homomorphisms from the n-chain group to a group G, $\phi : C_n \to G$. In our case $G = \mathbb{Z}_2$, and we can thus think of the maps as indicator functions of the corresponding chains. In other words, an n-cochain will be a list of size equal to $\dim C_n$ with each entry being 1 if the corresponding n-chain is present, and 0 otherwise. This induces a natural correspondence between chains and cochains under $\mathbb{Z}_2$ coefficients. It is worth noting that, by using a more expressive group such as $\mathbb{Z}$, cohomology groups are endowed with richer properties than homology groups, although this extension is beyond the scope of our discussion. The cohomology groups are connected by coboundary maps $d_n : C^n \to C^{n+1}$. For $\mathbb{Z}_2$ coefficients, the n-coboundary map is the transpose of the $(n+1)$-boundary map, i.e. $d_n = \partial_{n+1}^T$.

The cohomology groups are then defined as:

$$H^n = \ker d_n / \mathrm{im} d_{n-1} = \mathrm{coker}\, \partial_{n+1} / \mathrm{coim} \partial_n \tag{2.7}$$

with the property that $d \circ d = 0$, i.e. $\mathrm{im}\, d_{n-1} \subset \ker d_n$, or equivalently $\mathrm{coim}\, \partial_n \subset \mathrm{coker}\, \partial_{n+1}$.

Constructing a cohomology class on the annulus

Let's construct a cohomology class on the annulus from its definition: we need to find a 1-cochain ψ that will be in the cokernel of ∂_2. This will occur if the number of times each ψ takes on the value of 1 on the boundary of each 2 simplex is either 0 or 2, since $d_1(\psi)(\sigma) = \psi \partial_2(\sigma) = 0 \, \forall \sigma \in C_2$ (Hatcher, 2001). The 1-cochain in Fig 2.7 has precisely that property. Furthermore, as you can verify, no 0-cochain solves $d\phi = \psi$, and hence

$\psi \notin \operatorname{im} d_0 \implies [\phi] \in H^1(X)$, i.e. ψ is a representative of the 1-cohomology group.

Persistence cohomology

A groundbreaking advancement in topological data analysis occurred when researchers demonstrated that the persistence cohomology yielded the same persistence diagrams as traditional persistent homology with an increase of more than an order of magnitude in speedup (Silva, Morozov, and Vejdemo-Johansson, 2011). Furthermore, there has been great effort for making computational tools widely accessible. For example, computing the persistent cohomology of a point cloud using pyRipser (Tralie, Saul, and Bar-On, 2018) is just `ph = ripser (X)`, where `X` is a numpy array and `ph` contains the persistence diagrams and corresponding cohomology classes up to a certain dimension. We visualize the cohomology class with largest lifetime of data sampled from an annulus in Fig. 2.7.

Despite exciting progress in PH computation (see e.g. (Bauer, 2021; Scoccola et al., 2023; Bauer et al., 2024)), challenges remain in scaling these algorithms to datasets exceeding 10^6 points and computing higher dimensional homology groups.

The critical edge method in a bumpersticker

In the critical edge method, we exploited persistent cohomology to retrieve corresponding homology classes. The method is fairly straightforward, here we write it in pseudocode using `totopos`:

```
import anndata as ad
from ripser import ripser
import totopos.genes as tpg
import totopos.cells as tpc

# Read sc data
```

```
adata = ad.read_h5ad("path/to/sc.h5ad")


# Compute persistent homology to determine if there are loops
ph = ripser(adata.obsm["pcs"])


# Compute topoCells
topological_loops = tpc.critical_edge_method(
    adata.obsm["pcs"], ph, n_loops=1
)


# Compute topoGenes
grads, tpg_scores = tpg.topological_scores_perturbation_torch_ripser(
    adata, ph, n_pcs = 20, ix_top_class = 1
)
```

2.3 Discussion

Understanding how a single cell develops into a complex organism remains one of biology's fundamental questions. While recent advances in single-cell genomics have provided unprecedented views of development, extracting fundamental principles from these massive datasets requires new analytical approaches. Accurate mapping of differentiation trajectories is particularly crucial for understanding cell states across development, as traditional methods may impose artificial constraints or miss important biological features. This is particularly important as large-scale efforts–like the Human Developmental Cell Atlas (Haniffa et al., 2021)–aim to create comprehensive maps of human development, and hold the potential of helping treat and prevent disease (Liberali and Schier, 2024).

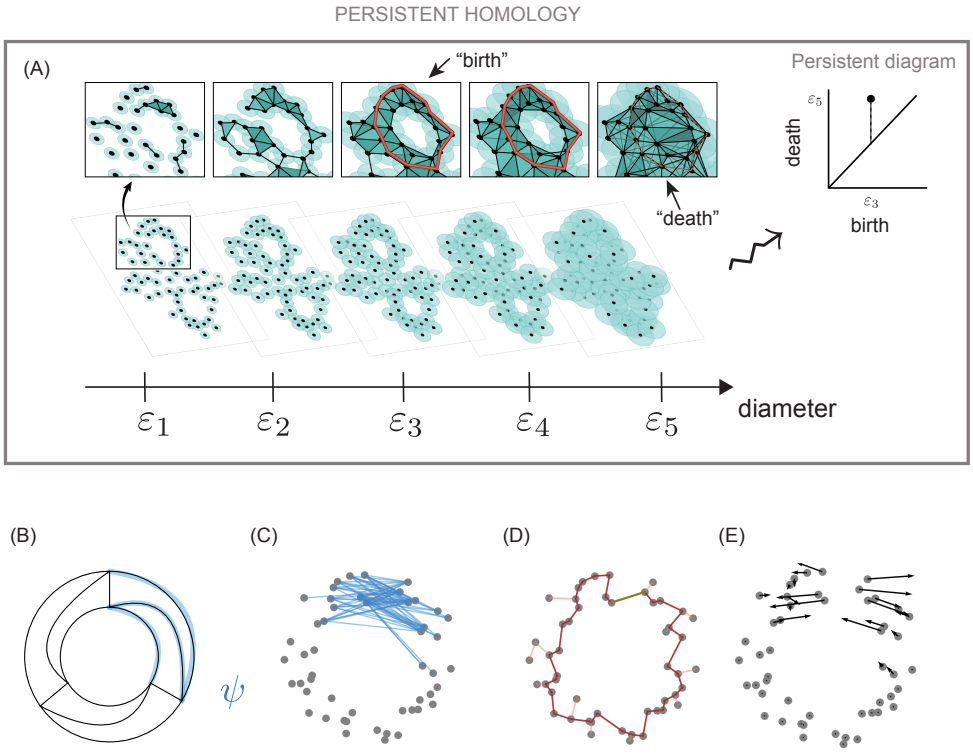

Figure 2.7: **Persistent (co)homology, the critical edge method, and topological gradients.** (A) Cartoon of the persistent homology algorithm. (B) A 1 cohomology class ψ on the annulus. (C) A 1 cohomology class computed using pyRipser, on data sampled from the annulus. (D) Visualization of the critical edge method. The CritEdge method constructs a minimum spanning tree T (orange) using the 1-skeleton of a Vietoris Rips complex at distance scale b_e the birth distance of the homology class. Adding the birth edge b (olive) to the MST results in a simplicial complex containing a loop, which is a representative of the H_1 homology class (red). (E) Visualization of the topological gradients calculated on the same data. Note that perturbing the data according to the negative of the gradient would result in increasing the birth time of the corresponding homology class.

Here, we show that topological analysis of cell state landscapes provides a powerful framework for understanding developmental trajectories. By analyzing homology groups and Betti numbers across more than 10 developmental cell atlases, we propose that cell state manifolds can be classified into simple building blocks: clusters, linear trajectories, trees, and topologies containing loops. Most importantly, we demonstrate that unbiased discovery of these loops reveals previously unrecognized biological processes involving convergent differentiation, stem cell maintenance, and tissue regeneration.

To systematically analyze complex developmental trajectories, we developed `totopos`, a computational framework built on rigorous topological foundations. At its core, `totopos` uses persistent homology (PH) to identify topological structures such as loops from noisy single-cell data, a key advantage over existing methods that can miss or distort important biological features. Moreover, `totopos` can also help determine when tree-based trajectory inference methods are appropriate, providing a principled approach to developmental trajectory analysis.

In this work, we provide a conceptual framework that bridges biology and topology. For instance, we defined two key concepts: 1) `topoCells`, the cells in the neighborhood of a loop trajectory, and 2) `topoGenes`, the subset of the transcriptome that drives the topological loop. The scale of current single-cell genomic datasets required algorithmic innovations to compute `topoCells` and `topoGenes` efficiently. For identifying `topoCells`, we developed the critical edge method, which is to the best of our knowledge the fastest method to compute persistent H_1 homology representatives. To identify `topoGenes` we extended the work of Nigmetov and Morozov (Nigmetov and Morozov, 2022) for computing gradients to trivialize a homology class in the context of topological optimization. Our computations of the gradients of the topological loss function are essentially "free" with a precomputed persistence diagram. In the restricted case of mapping a homology class from $[b, d) \mapsto [d, d)$ we also achieve state-of-the art performance. The key for our advances is exploiting duality (Silva, Morozov, and Vejdemo-Johansson, 2011) by leveraging information from the cohomology generators supplied by Ripser (Bauer, 2021), the fastest method to date to compute persistent diagrams. As a whole, the advances we present enable the topological analysis

of datasets on the order of 10^6 single-cells. Given the broad applicability of our methods, we anticipate they will catalyze discoveries beyond developmental biology.

Our analysis using `totopos` revealed that convergent loops are a recurring motif across animal development. We identified four prominent loops in human immune cell development, including one in the myeloid-erythrocyte system (Suo et al., 2022a). In *Hydra vulgaris*, we found that a convergent loop in gland cells that may enable transdifferentiation during regeneration. Most strikingly, we found a convergent loop in the characteristic cell type of cnidarians, the cnidocytes, that could be conserved across $\approx$ 500 million years of evolutionary time. While most loops we found were convergent, *C. elegans* seam cells displayed a cyclic loop (Fig 2.5). This raises the question of whether the cyclic topologies are unique to seam cells, or if this is a more general feature of stem-like cell states.

Our findings provide a unified theoretical framework for understanding cell state control during development. For over 60 years, Waddington's landscape model (C. Waddington, 1957) has been instrumental in our understanding of development. This metaphor suggests that developmental trajectories should form a branching tree-like structure. Our systematic analysis supports this view: two-thirds of the cell state manifolds we examined are topologically equivalent to trees, providing quantitative evidence for the pervasiveness of Waddington's conceptual model.

Recent work by Rand et al. (Rand et al., 2021) has formalized Waddington's metaphor mathematically, proposing that cells move through gene expression space following gradient flows. At first glance, our discovery of loops in cell state manifolds might seem to contradict this model, since gradient flows cannot harbor cyclic trajectories. However, our findings reveal a more nuanced picture: while we did find cyclic loops in the seam cells, most of the loops we discovered were convergent. Importantly, we mathematically prove that gradient flows can indeed create such convergent trajectories (SI Note 2.5, SI Figure S21), reconciling our topological discoveries with the dynamical systems framework proposed in (Rand et al., 2021). In other words, the topological motif perspective, in which cell state manifolds are constructed from trees and convergent loops is

consistent with the dynamics of a gradient field, i.e. a flow in a landscape. As the authors suggest in (Rand et al., 2021), the presence of cyclic trajectories can be taken into account under the more general notion of a Morse-Smale dynamical system.

Our study doesn't come without technical limitations. First, our computational resources restricted persistent homology computation to a maximum of 50, 000 single-cells per run. We approached this limitation by computing using techniques with reasonable error bounds (Cavanna, Jahanseir, and Sheehy, 2015). A different approach to handle large datasets is to compute PH on subsets of data, by e.g. partitioning by lineage, tissue or cell type. Recent advances in theoretical and computational topology are providing faster and more efficient methods (Bauer, 2021; Scoccola et al., 2023), and more insights can be drawn from the compendium we analyzed once new computational techniques are available.

Furthermore, in this work, we focused on the most prominent topological features. We defined heuristic methods to identify prominence. For β_1, we defined a cutoff for topological noise based on local calculation of loops in the manifold: prominent loops were defined as those whose lifetime exceeded twice the maximum lifetime of loops in neighborhoods of 350 cells. This heuristic provides an intuitive way to study loops, but severely limits the richness of topological information present in persistent diagrams and could underestimate the number persistent homology classes. For instance, for the *Xenopus tropicalis* dataset, we calculated that the ratio between the most prominent class and the neighborhood threshold was 1.93 and was therefore considered not significant using our approach. In this sense, we encourage researchers interested in a particular organism to use our tools to explore the topological features of their data and further focus on particular tissues or cell types for more detailed inquiry.

Another limitation of our approach is that persistent homology can be sensitive to outliers. Multiparameter persistent homology (MPH), generalizes PH by creating bifiltrations (for example by distance and density) and is more robust for outliers. MPH is an exciting new field with multiple interesting theoretical and computational problems (Botnan and Lesnick, 2023), and has recently

been applied for cell type classification (Benjamin et al., 2024). We anticipate that the resolution of topological identification will increase with the use of MPH, and will be an exciting avenue for future research.

2.4 Methods

Single-cell RNA-seq data pre-processing

Single-cell RNA seq count matrices were pre-processed using a standard pipeline. First, we filtered out cells with less than 500 detected genes and 1,000 UMIs. To minimize variability in the total number of reads per cell the expression values were normalized to get values roughly equivalent to those expected in a single cell, using the following equation:

$$\tilde{X_{ij}} = \ln\left(\frac{X_{ij} \times 10^4}{\sum_j^m (X_{ij})} + 1\right). \tag{2.8}$$

In addition, we identified the highly variable genes using the coefficient of variation method. Finally, we reduced the dimensionality of the transcriptome by projecting the transcriptome to the first 20 principal components.

Topological census of Developmental Cell state Manifolds

To characterize the global topological structure of developmental cell state manifolds, we performed a topological census using persistent homology. Persistent homology was computed using the Ripser algorithm on normalized scRNA-seq count matrices, generating persistence diagrams for the zero- and one-dimensional homology groups (for details see Supplementary Methods).

For datasets exceeding 50,000 cells, we employed Farthest Point Sampling (FPS) to subsample representative points while preserving the topological structure of the manifold, leveraging the bottleneck stability theorem to bound deviations in persistence diagrams (see SI for details). To identify the most prominent topological features, we developed a heuristic to distinguish signal from topological noise: a loop was considered significant if its persistence exceeded twice the maximum lifetime of any H1 class in local neighborhoods of 350 cells. The resulting Betti numbers were

used to classify each cell state manifold into one of several canonical topological motifs: clusters ($\beta_0 > 1, \beta_1 = 0$), tree-like structures ($\beta_0 = 1, \beta_1 = 0$), and manifolds with nontrivial loops ($\beta_0 = 1, \beta_1 \geq 1$).

We used the efficient implementation of Ripser (Bauer, 2021) in python, pyRipser (Tralie, Saul, and Bar-On, 2018) (version 0.6.4), to perform our persistent diagram calculations. All computations were performed on a workstation with an AMD Ryzen Threadripper 3960X with 256 GB of RAM, and using Python 3.11.

Identifying `topoCells` using the critical edge method

Let X be a developmental cell atlas, with a topological loop $\beta_1(X) = 1$, and assume one has access to its H_1 persistence diagram Dgm_1. Assume the most prominent homology class $[\gamma]$ is born at scale b, and let e be its birth edge. The critical edge method then proceeds as follows:

1. Build a Vietoris Rips complex $K = \mathrm{VR}(X, b)$ with b the birth scale of γ.
2. Construct a minimum spanning tree T on the 1-skeleton of K starting from a bounding point of e. such that T does not contain e.
3. Scan for a loop α on $L = T \cup \{e\}$.

Note that in the second step, since all edges will have a distance smaller than e, the tree T will not contain e by the greedyness of the algorithm. Hence, adding e to T will necessarily induce a loop in the resulting simplicial complex.

Following this procedure, the loop α found by the algorithm is a representative of the homology class $[\gamma]$. We prove the algorithm in the SI.

Finding `topoGenes` using topological gradients

In this section we define a ranking procedure that reflects the importance of each gene in X in a target homology class γ. Our aim was to find the topology-driving genes `topoGenes` using

this ranking. Our approach uses topological optimization building upon the work of Nigmetov and Morozov [(Nigmetov and Morozov, 2022)] and exploiting computational efficiency of Ripser. Namely, we define a "loss" function that aims to reflect perturbations in the simplicial complex that would be necessary to trivialize a persistent homology class, i.e. mapping it from a point $[b, d)$ in Dgm_1 to a the diagonal $[d, d)$. Recall that one can define the Vietoris-Rips filtration as the sublevel set of the diameter function $\operatorname{diam}\sigma = \sup\{d(x, y) : x, y \in \sigma\}$, that is $K_\varepsilon = \{\sigma \subseteq X \,|\operatorname{diam}\sigma \leq \varepsilon\}$. In particular, the general form of this loss function is:

$$\mathcal{L} = \sum_{\rho \in C} [f(\rho) - f'(\rho)]^2,$$

where $f(\rho)$ is the largest distance of an edge in the original filtration, and $f'(\rho)$ is a target distance. Our goal $f'(\rho)$ will be set to the death value d. The set of simplices C one needs to evaluate in the loss function is called the critical set, as defined in (Nigmetov and Morozov, 2022). It turns out that under our restricted set up C contains the edges associated to the cohomology class generator. We provide a proof of this result in the SI. The intuition is that increasing the birth time of γ can be achieved by increasing the edge lengths of the corresponding cohomology generator. Our approach becomes fast by exploiting the cohomology computation from Ripser.

To get the ranking, we estimate the gradients of the aforementioned loss function w.r.t. the input data via backpropagation. More precisely, given a scRNA-seq dataset X wtih a topological loop $\beta_1(X) = 1$, we define the ranking score for gene j to be the norm of the j-th column of the gradient matrix.

$$S_j = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial \mathcal{L}}{\partial X}\right)_{ij}^2}.$$

We can then define `topoGenes` as the genes with the largest ranking score. Intuitively, the ranking score encodes the relative contribution for destroying a persistence homology class by moving input

data points. Ranking scores will be non-zero for the subspace of the cell-state manifold where the homology class is geometrically embedded.

Alternative Approach for Identification of `topoGenes` via Iterative Gradient Optimization

Building upon the previously described topological gradient optimization method, we introduce an iterative sampling-based strategy designed to efficiently identify critical genes influencing topological structure within the data manifold.

Our method involves repeatedly sampling subsets of cells from the principal component embedding. Specifically, we performed $n_{reps} = 100$ independent iterations, each time randomly selecting a small subset of cells (up to 10% of the total dataset). Within each subset, cells were sorted according to an established pseudotime or cell-cycle ordering. Then, we constructed closed loops by appending the first cell in the ordered subset to its end, effectively creating a cyclic structure or a representative of the target homology class. The topological loss for each loop was computed by summing Euclidean distances between consecutive cells along this trajectory:

$$\mathcal{L}_{topo} = \sum_{i=1}^{m} ||x_{i+1} - x_i||,$$

where $x_{m+1} = x_1$ ensures the closure of the loop.

We accumulated the loss values across iterations to obtain a total loss:

$$\mathcal{L}_{total} = \sum_{rep=1}^{n_{reps}} \mathcal{L}_{topo}^{(rep)}.$$

Gradients of the cumulative loss with respect to the original high-dimensional gene expression data were computed via backpropagation. The resulting gradients directly measured how perturbations in individual gene expressions influenced the homology class. Genes were then ranked based on the magnitude of their gradient norms across cells, defined by:

$$S_j = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial \mathcal{L}_{total}}{\partial X} \right)_{ij}^2}.$$

Genes exhibiting the highest gradient-based scores were classified as `topoGenes`, indicating their essential role in shaping and preserving topological features within the cellular data manifold.

Finite difference method to identify persistent $0-$homology

A parsimonious approach to identify the most prominent $0-$homology features of a dataset is to select the number of connected components that persist significantly longer than others, indicated by a clear separation in their persistence lifetimes.

The $0-$th homology persistence diagram has the special characteristic that all homology features are born at the start of the filtration. Thus, the lifetime equals the death time for all persistent $0-$homology features. This induces a natural ordering between these features.

Let $\{l_i\}_{i=1,\dots,n}$ denote the sequence of lifetimes of the $0-$homology features, ordered in descending order such that $l_i > l_{i+1}$ for all $i \in \{1, \dots, n-1\}$. To identify significant gaps between consecutive lifetimes, we compute the first-order differences:

$$d_i = l_i - l_{i+1} > 0 \quad \text{for } i \in \{1, \dots, n-1\}. \tag{2.9}$$

These differences d_i form a sequence $\{d_i\}_{i=1,\dots,n-1}$ that quantifies the gaps between consecutive ordered lifetimes. To identify the most significant gap, which would separate the prominent persistent features from the rest, we compute the second-order differences:

$$\Delta d_i = d_i - d_{i-1} \quad \text{for } i \in \{2, \dots, n-1\}. \tag{2.10}$$

The optimal number of significant $0-$homology features, k^*, can then be identified as:

$$k = \underset{i \in \{2,\dots,n-1\}}{\operatorname{argmax}} \Delta d_i. \tag{2.11}$$

This approach identifies the index where the rate of change in consecutive lifetime differences is maximized, effectively detecting the "elbow point" in the lifetime distribution. The first k persistent $0-$homology features are then selected as the significant connected components of the dataset. In cases where multiple maxima exist in $\{\Delta d_i\}$, we select the smallest index to favor parsimony.

Bootstrap permutation test to quantify significance of topological signatures

We used a bootstrap permutation test using a bifurcating tree as a null topology to quantify if the persistent homology signatures could be explained by random chance. To do this we employed difference of maximal lifetime between a test dataset and a null hypothesis dataset. Let the lifetime of a persistent homology class be $l_{i,X}^{H_n} = \varepsilon_{i_{\text{death}}}^{H_n} - \varepsilon_{i_{\text{birth}}}^{H_n}$ where i is the index of the homology class in the filtration.

Our statistical test is based on the following hypotheses:

1. H_0: $\max\{l_{X_{\text{null}}}^{H_n}\} = \max\{l_{X_{\text{test}}}^{H_n}\}$
2. H_1: $\max\{l_{X_{\text{null}}}^{H_n}\} < \max\{l_{X_{\text{test}}}^{H_n}\}$

We thus define our test statistic as follows:

$$\hat{\theta} = \max\{l_{X_{\text{test}}}^{H_n}\} - \max\{l_{X_{\text{null}}}^{H_n}\}. \tag{2.12}$$

To simulate the null hypothesis, we concatenate the null and test datasets, shuffle, partition, and compute the test statistic $\hat{\theta}^{(b)}$ for each bootstrap replicate. We performed $B = 10^4$ bootstrap replicates of this test and report the P-value as the fraction of simulations in which $\hat{\theta}^{(b)}$ is more extreme than the test statistic $\hat{\theta}$. In order for the datasets to be in the same scale, we use the singular values of the test data to scale the principal components of the null dataset prior to all computations.

Additionally, we employed the same test by replacing the maximum, considering the k-th order statistic. This enabled generalizing our test for the case when there was more than one persistent feature.

Perturbation-based validation of topology-driving genes

To assess whether the inferred `topoGenes` drive topological features in data, we applied our bootstrap-based hypothesis testing framework. We asked if removing the `topoGenes` from the data

resulted in a significant decrease of the lifetime of the most prominent homology class. For each test, we constructed X_{null} as the original dataset consisting of `topoCells` for a single loop and X_{test} as the dataset with the `topoGenes` set removed. Both datasets were centered and processed using singular value decomposition, retaining the top 20 principal components. To ensure comparable scales, we projected both datasets using the singular values from X_{null}. Analyses were performed with bootstrap sample sizes set to 85 % of the dataset size and 1,000 bootstrap iterations.

Laplacian eigenvector method for `topoGene` ranking

We exploited homology generators to identify genes expressed transiently along complex topological features. Our methodology is based on the property that Laplacian eigenvectors encode geometrical properties of a manifold (B. Levy, 2006). Furthermore, eigenvectors can be interpreted as vibration modes, with increasing eigenvalue corresponding to an increasing spatial frequency. We thus extracted transient genes by asking which genes had the highest mutual information with respect to the first nonzero eigenvectors of the Laplacian. We used this method to find the `topoGenes` of the *C. elegans*' seam cell loop.

Pseudotime methods for closed-loop trajectories

Cells may traverse the cell state manifold in a way that may not reflect temporal progression. To address this, and leveraging topological information, we developed pseudotime methods that reconstruct natural coordinates parametrizing of the cell state manifold. Furthermore, we designed different approaches for the convergent and cyclic systems.

The convergent trajectory method calculates a pseudotime based on geodesic distances in a kNN graph constructed from `topoCells`. Given a specified initial cell state, the method assigns pseudotime values according to the geodesic distance to this reference point. This approach is suited for convergent systems where cells traverse multiple paths to reach a common end state. We leveraged this framework to partition `topoGenes` into pervasive and exclusive sets, corresponding to genes expressed throughout the loop versus those restricted to specific trajectories. We used Dijkstra's method implemented in `scipy` (Virtanen et al., 2020) and $k = 10$ neighbors for our

experiments.

For cyclic trajectories, we implemented the circular coordinates algorithm as implemented in Dreimac (Silva, Morozov, and Vejdemo-Johansson, 2011; Scoccola et al., 2023). In brief, this algorithm constructs a circular parametrization of the data that captures the underlying periodic structure by exploiting information from persistent cohomology computation. By specifying an initial cell state, the method generates an ordering of `topoCells` that optimally decodes the biological progression of cells around the loop trajectory.

Locally linear Manifold Alignment for comparative transcriptomics analysis

Locally Linear Manifold Alignment aims to retrieve a shared features between two cell state manifolds by aligning tangent spaces of two cell-state manifolds defining locality at the cell type level. The method starts by generating a set of one-to-one highly-variable orthologs to capture genes that contribute to cellular heterogeneity, which we explain in detail in the section below. We perform this HV ortholog selection to select approximately 3000 orthologs. Then, count matrices X_a, X_b are generated for two cell state manifolds with columns specified by the orthologs. Since not all cell states from one species may have a corresponding representative in the other species, a Mutual Nearest Neighbors search is performed to retrieve sets of "homologous" cells. After this step we have two data sets $\tilde{X_a}, \tilde{X_b}$ on the same ortholog space. We find a shared subspace using Canonical Correlation Analysis (CCA) using well-known algorithms. Geometrically, CCA finds a shared subspace where two datasets are well aligned. We perform PCA with n=50 principal components as a preprocessing step to retrieve the local tangent space, as defined in (Budninskiy et al., 2019). The approach of performing PCA before CCA was first defined in (Brown, Bray, and Pachter, 2018), and helps to reduce overfitting, particularly in the case where the number of orthologs exceeds the number of cells to analyze. We implemented LMA in pytorch enabling efficient matrix computations on the GPU.

Generating One-to-One Highly-Variable Orthologs for LMA

We developed a heuristic to generate one-to-one highly-variable orthologs for LMA using the coefficient of variation (CV) as a filtering statistic. This approach leverages precomputed orthologs obtained through methods like, e.g., a bidirectional BLAST search.

Under a Binomial sampling model, one can derive that the coefficient of variation $CV = \sigma/\mu$ per gene is given by:

$$\log(CV) \approx -\frac{1}{2}\log(\mu) + \epsilon,$$

where σ is the standard deviation and μ is the mean expression level of the gene using a Poisson approximation. We computed the CV for each gene within each species and ranked them based on the distance between the observed CV and the expected value under Binomial sampling. This ranking prioritizes genes with the most variable expression relative to their mean.

To identify one-to-one orthologous highly-variable genes (HVGs), we applied the following procedure:

1. **Initial Filtering and Ranking:** We first ranked genes in species A and B based on their CV. For each gene set, we selected the top k genes as candidate HVGs.

2. **Ortholog Retrieval and Intersection:** For each gene in the HVG list of species A, we retrieved all corresponding orthologs in species B. This initial mapping often resulted in a one-to-many relationship. To address this, we calculated the intersection between HVGs in species B and the orthologs derived from species A, ensuring that only highly-variable orthologs were considered.

3. **Bidirectional retrieval** From the highly-variable orthologs of species A, we retrieve highly variable orthologs from species B.

4. **Refinement** To achieve one-to-one ortholog mapping, we performed a filtering step. For each candidate HVG ortholog in species A:

    - We retrieved its orthologs in species B.
    - These orthologs were then ranked by their mean expression levels.
    - If multiple orthologs were present, the highest-ranking gene that was not already matched was selected, ensuring a unique one-to-one relationship.

This heuristic yields a set of one-to-one ortholog genes that can be used for comparative transcriptomics analysis taking into account data statistics.

SCENIC Analysis for Transcription Factor Regulon Identification

We employed the Single-Cell Regulatory Network Inference and Clustering (SCENIC) pipeline to identify transcription factor (TF) regulons enriched in different segments of the topological loop in the human immune development dataset. Following the protocol described in (Aibar et al., 2017), with modifications specified below, we conducted regulatory network inference and TF regulon activity analysis.

First, we created a dedicated conda environment without specifying the Python version and installed SCENIC by cloning the `pyscenic` GitHub repository directly to circumvent potential compatibility issues encountered when using pip-based installation methods. Minor adjustments to the source code of the `arboreto` and `dask_expr` packages were implemented to ensure compatibility and robustness during distributed computations.

Gene regulatory networks (GRNs) were inferred using the GRNBoost2 method from the `arboreto` package on single-cell RNA-seq data stored in loom format, based on the list of human TFs from (Lambert et al., 2018). The GRN output was used to generate candidate regulons, employing motif rankings based on the human genome hg38, considering a 10 kb upstream and downstream window (`hg38_10kbp_up_10kbp_down_full_tx_v10_clust.genes_vs_motifs. rankings.feather`)

and motif annotations `motifs-v10nr_clust-nr.hgnc-m0.001-o0.0.tbl`. Cellular enrichment of candidate regulons was then computed to obtain the regulon activities across cells. The commands were run with a distributed computing framework utilizing the `dask_multiprocessing` mode with 24 parallel workers.

Inference of Regulatory Relationships Between TFs and Gene Programs

To infer regulatory relationships between transcription factors (TFs) and gene programs, we performed downstream analysis using SCENIC. Target genes associated with each TF regulon were grouped and ranked. Independently, gene program-specific gene lists were compiled. The statistical significance of overlaps between TF regulon target genes and gene programs was evaluated using Fisher's exact tests, employing a background set consisting of all genes detected across the dataset. Multiple testing corrections were applied using the Benjamini-Hochberg procedure to control for false discovery rate (FDR).

2.5 Proofs

Theorem. Critical edge method yields a representative of H_1.

Let $K = \mathrm{VR}_{\varepsilon_b}$ be the Vietoris–Rips complex at scale ε_b constructed from a metric space (X, d). Assume one has access to the H_1 persistent homology of X encoded in its persistent diagram $\mathrm{Dgm}_1(X)$. Let e be the critical (positive) edge at which a persistent homology class $[\gamma]$ is born at ε_b, and assume this class has multiplicity one. Let T be a minimum spanning tree (MST) of the $1-$skeleton of K, constructed via starting from one of the bounding points of e. Let $L = T \cup \{e\}$. Then the loop α formed by adding e to T is a representative of the homology class $[\gamma]$ in $H_1(K)$.

Proof:

Since T is a tree spanning all vertices of K, it is connected and acyclic. Adding the edge e to T creates a unique cycle α in the graph L, i.e. $\alpha \in \ker\partial_1^L$. By construction, a complex L is a subcomplex of K because all edges in L are also in K. The inclusion map $i : L \hookrightarrow K$ is simplicial and induces a homomorphism on homology groups $i_* : H_1(L) \to H_1(K)$, so automatically $i_*([\alpha]) \in H_1(K)$.

Next, we need to show that $[\alpha]$ is not a trivial homology class in K, i.e. $\alpha \notin \mathrm{im}\partial_2^K$. Suppose, for contradiction, that $\alpha = \partial_2^K(c)$ for some 2-chain $c \in C_2(K)$. This would imply that α bounds a collection of 2-simplices in K, and thus α represents the trivial class in $H_1(K)$. However, this is a contradiction given that the $1-$chain α contains the edge e, which is the edge added exactly at ε_b, and, at that scale, e does not have cobounding triangles since it is the largest edge in the filtration. Since α is a cycle in K and $\alpha \notin \mathrm{im}\partial_2^K$, it is a representative of a nontrivial homology class in $H_1(K)$.

The last argument also implies that α is not a representative of homology classes born before ε_b, since it contains an edge that was not present in previous steps in the filtration. Furthermore, we know that a single class is born at that scale by the assumption of multiplicity one, i.e. β_1 increases by one, when adding e. Therefore, α must be a representative of $[\gamma]$ given that $[\gamma]$ is the single homology class born at ε_b. ■

Theorem. Perturbation method only requires the simplices of a cohomology generator

Suppose there is a persistent homology class α with birth time b and death time d with positive simplex σ and negative simplex τ. The perturbation method works by mapping a PH class from $(b, d] \mapsto (d, d]$ in the persistence diagram. Then, the simplices needed to modify to perform this perturbation in the diagram are precisely the simplices forming the representative of the corresponding cohomology generator available from Ripser.

Proof:

By Theorem 11 in Nigmetov and Morozov, 2022 increasing the birth time to d of a single persistent homology class amounts to changing the values of the simplices ρ satisfying two conditions:

1. $f(\rho) \in [b, d]$
2. $V^{\perp}[\rho, \sigma] \neq 0$

But by section 3.4 Silva, Morozov, and Vejdemo-Johansson, 2011 we have that if $\text{low}R^{\perp}[:, \sigma] = \tau$ then the simplices in $V^{\perp}[:, \sigma]$ are the cohomology generator, so the second condition is met. By duality, the cohomology class corresponding to α are born at d and die at b, i.e. the conditions are switched, by contravariance of the cohomology functor. Therefore the simplices in $V^{\perp}[:, \sigma]$ satisfy the first condition automatically. ■

Theorem. There exist gradient fields with convergent trajectories.

Proof: We provide a constructive proof.

Let M be a compact, 2-dimensional, oriented manifold with boundary and let $f : M \to R$ be a Morse function with a subset of critical points $\{r, s_1, s_2, a\}$ with indices $\{2, 1, 1, 0\}$ with critical values $v_r, v_{s_1}, v_{s_2}, v_a$. Consider the gradient flow $dx/dt = -\nabla f$. In this set up, s_1 and s_2 represent saddle points, and r a are unstable and stable fixed points respectively. In other words, a is an attractor, and r is a repellor. We visualize the system in Figure S21.

Assume that there is a parametrization of the unstable manifold of the repellor r where each basis vector coincides with one stable manifold of the saddles s_1, s_2. Furthermore, assume that there is a parametrization of the stable manifold of the attractor a where each basis vector coincides with the unstable manifold of the saddle points.

Then, there exist convergent trajectories γ_1, γ_2 with the following properties:

- $\lim_{t\to-\infty}\gamma_i(t) = r$.
- $\gamma_1(t) \neq \gamma_2(t)$ for some $t \in (T_1, T_2)$.
- $\lim_{t\to\infty}\gamma_i(t) = a$.

2.6 Supplementary Notes

High dimensions can mask topological structure in persistent diagrams

One important example of the curse of dimensionality is that Euclidean distances become meaningless in high-dimensions. Since Vietoris-Rips complexes are defined solely on pairwise distances, this effect becomes significant for the interpretation of persistence diagrams. With the goal of mitigating this effect, one crucial step in our topological analysis is projection of single-cell data using PCA. In (Hiraoka et al., 2024), Hiraoka et al. showed that in high dimensions, the topological signal in persistence diagrams become unreliable under a Gaussian noise model, and showed that PCA remediates this effect.

To systematically investigate this phenomenon, we designed an experiment using the following probabilistic model:

$$t \sim \mathrm{Unif}[0, 2\pi) \tag{2.13}$$

$$C = (r\cos t, r\sin t, 0, \ldots, 0) \in \mathbb{R}^d \tag{2.14}$$

$$N \sim \mathrm{MVN}(0, \eta^2 I_d) \tag{2.15}$$

$$X = C + N \tag{2.16}$$

In this model, points are sampled uniformly from a circle of radius r in the first two coordinates, embedded in $\mathbb{R}^d$, and subjected to isotropic Gaussian noise with variance η^2 in each coordinate. This model allows us to directly observe the effect of dimension on the recovery of the manifold's topology, even when the intrinsic signal structure remains unchanged.

For large dimensions $d \gg r$, the signal becomes dominated by noise. This is evident from considering the random Gaussian vector $N \sim \text{MVN}(0, \eta^2 I_d)$. Its expected squared norm is given by:

$$\mathbb{E}[\ \|N\|^2\ ] = \sum_{i=1}^{d} \mathbb{E}[N_i^2] = \sum_{i=1}^{d} \mathbb{V}[N_i] = \sum_{i=1}^{d} \eta^2 = \eta^2 d$$

since the N_i are i.i.d. Gaussians. This result implies that under the given probabilistic model, the norm of X scales linearly with the dimension d. Consequently, in high-dimensional settings, distances become increasingly influenced by noise, thereby diminishing the persistence of the homology class corresponding to the circle.

We performed simulations to test the effect of dimension on the noisy circle model and the *C.elegans*' seam cell dataset. We used $r = 10, \eta = 1$, and $n = 2,766$ points (the number of seam cells). Furthermore, we explored computing PH on dimensions $d = \{20, 100, 500, 1000, n_{\text{genes}} = 2501\}$

The results in Fig S3 reveal several key insights. First, with increasing dimension the homology classes shift to birth times involving higher birth distances, which makes sense as the volume in higher dimensional spaces increases. Furthermore, we found that in both cases, dimension alone can decrease the topological signal, demonstrated by the decrease in the lifetime of the most prominent homology class. This is interesting since the two datasets have different noise structure: constant for the noisy circle, and exponentially decaying in the case of *C. elegans* under the change of basis in PCA. Moreover, we found that in the case of the seam cells, the signal is present even in the original dimension when using all highly variable genes ($n = 2,501$). These results suggest that PCA can indeed ameliorate the loss of topological signal by the curse of dimensionality.

Error bounds for topological inference of large datasets

In this note, we provide an error estimate of the topological census for large datasets. Concretely, we will show that the Hausdorff distance d_H provides an error bound on the distance between the persistence diagrams of a sample and the entire dataset. This implies that, even if we cannot compute the persistent diagram of a large dataset, we can use the persistent diagram of a sample as a proxy of its topological signature, and provide an estimate of how uncertain we are of this approximation. At the end, we also discuss other possible strategies and their limitations.

Introduction

Determining the Betti numbers of developmental cell state manifolds (and in fact of any real dataset) is fundamentally an inference problem. In essence, our approach has at least one limitation. First, one of our assumptions is that we start with a representative sample X from an underlying developmental manifold $\mathcal{M}$. Then, we can compute a persistent diagram of X (which contains all its topological information (S. Y. Oudot, 2015)) to approximate the topology of $\mathcal{M}$. For large datasets (given our computational infrastructure, sets of $n > 50,000$ single-cell transcriptomes), the topological inference becomes more challenging, since the computation is dependent on the number of simplices, which can grow exponentially as a function of the number of points. Thus, we can only compute the persistent diagram using a subsample $\tilde{X} \subseteq X$ to approximate the topology of $\mathcal{M}$. Despite this endeavor might seem daunting, there is one theorem that puts us on firm grounds: the stability theorem for persistent homology (Chazal, Cohen-Steiner, et al., 2009; Chazal, Silva, and S. Oudot, 2014).

The stability theorem states that the distance between persistent diagrams of two datasets X, Y its bounded above by the Gromov-Hausdorff distance:

$$d_b\left(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)\right) \leq d_{GH}(X, Y) \leq d_H(X, Y). \tag{2.17}$$

On the left d_B is the bottleneck distance, which measures the longest edge between the best matching between persistent diagrams:

$$d_b\left(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)\right) = \inf_{\gamma} \sup_{p \in \mathrm{Dgm}(X)} \|p - \gamma(p)\|_\infty,$$

where the infimum is taken over all bijections $\gamma : \mathrm{Dgm}(X) \to \mathrm{Dgm}(Y) \cup \Delta$, and Δ denotes the diagonal set consisting of points of the form (a, a) in $\mathbb{R}^{2+}$.

On the right hand side of (2.17), d_{GH} and d_H are the Gromov-Hausdorff and the Hausdorff distance, respectively. The Hausdorff distance is a metric on subsets of a metric space. The Gromov-Hausdorff extends this notion by measuring the distance between two metric spaces, under isometric embedings onto a common space.

It turns out that this result mitigates uncertainty on both of our assumptions for topological inference. First, it tells us that small perturbations on the data result in small perturbations of the persistence diagrams. This suggests that we can hope to get similar results, if we were to perform our calculations with biological replicates of a given single-cell atlas. Secondly, despite we can't compute the persistence diagram of the entire dataset $\mathrm{Dgm}(X)$, we can very precisely compute an upper bound of its bottleneck distance with respect to the subsample's diagram $\mathrm{Dgm}(\tilde{X})$ using the Hausdorff distance.

Assessing sampling strategies on synthetic data

To evaluate sampling strategies for topological inference we compared three methods:

1. Farthest point sampling (FPS)
2. Closest representatives to k-Means centroids
3. Uniform random sampling

FPS offers a known two-approximation of the optimal subsample in terms of the Hausdorff distance d_H to the full dataset (Sheehy, 2020), and its algorithm computes d_H in the process. We benchmarked these methods on a synthetic dataset—a noisy circle (a circle with added Gaussian noise)—by quantifying the Hausdorff distance d_H and Bottleneck distance d_B of the subsample to the full dataset. In Figure S4 we provide a visualization of our results. Our findings were the following:

1. The d_H decreases with the number of samples on all three strategies, as expected.
2. FPS performs best, with k-Means being competitively close in d_H for certain cases.
3. We numerically confirmed that d_H provides an upper bound for d_B.

Based on these results, we adopted FPS as our default sampling strategy for the topological census in datasets exceeding 50,000 cells. For each such dataset, we report the Hausdorff distance of the FPS subsample as a conservative upper bound on the bottleneck distance between the inferred persistence diagram and that of the full dataset. This quantifies the approximation error introduced by subsampling in our topological analysis.

Assessing the effect of greedy furthest point sampling on real data

We evaluated the robustness of topological inference under FPS, by computing persistent homology (PH) on subsamples of the *C. elegans* atlas. We used fractions of data corresponding to $\{1, 5, 10, 25\}$ per cent of the data, spanning the approximation regime in our census. The subsampled cells are highlighted in the top row Figure S5.

Our analysis consisted on the following procedure:

- We used the critical edge method to retrieve the two most persistent loops.
- We then computed the `topoCells` by taking the neighborhood at a birth distance away from the zero skeleton of the H_1 homology representative.

- Compute the Jaccard score with respect to a reference `topoCell` set to quantify the "accuracy" of retrieval at sparser samples.
- Perform a null distribution by sampling random sets of the same cardinality as the `topoCells` for comparing the computed Jaccard score.

Our results for the second most persistent loop are visualized in Figure S5, but a similar analysis holds for the most persistent class, and can be computed using the code provided in the Github reproducibility notebooks. We used the topology from the data at 25% sampling level as reference. We highlight the most relevant findings from our analysis.

First, by looking at the loop representative, we can see that, even at a 5% sampling level, the critical edge method is still able to retrieve the geometrically correct topological loop, and at 1% sampling level, the loop is significantly degraded, not spanning the same path of cell states as the true set. Second, we can see that the persistent diagrams have a similar structure up to the 5% sampling level too (Figure S5, middle row). Third, by visualizing the `topoCells`, we can see that, even at a 1% sampling level, we can retrieve some relevant cell states, but the quality is almost identical in distribution for the 5 and 10% sampled data. The latter result is confirmed by the fact that the Jaccard scores from the inferred `topoCells` are far than we would get at random (inset distributions Fig S5). Together, these results suggest that FPS can effectively capture biologically relevant topological structure up to the 5% sampling regime, which underscores the relevance of census results.

Benchmarking our topological approach using a simulated cyclic scRNA-seq dataset

To assess the efficacy of our framework, we conducted control experiments using simulated scRNA-seq datasets with ground truth topology using dyngen (Cannoodt et al., 2021). This software package utilizes the Gillespie algorithm and real data statistics to simulate the acquisition of scRNA-seq data with a user-specified gene regulatory program. We designed a GRN consisting of 100 transcription factors, 1,000 target genes, and 500 housekeeping (HK) genes. Its wiring diagram is visualized in

(Fig S S1 A). The simulated dataset exhibited Poisson statistics, characteristic of real scRNA-seq data, ensuring a realistic case study (Fig S S1 B).

Applying our pipeline, we constructed the cell state manifold from the simulated data and computed persistent homology to identify prominent topological features (Methods). The persistence diagram revealed that the $0-$homology classes could not be well separated, indicating a large connected component subject to noise. Furthermore, the one-dimensional persistence diagram revealed a prominent H_1 homology class, indicating the presence of a loop (Fig S1 E. orange dots).

To evaluate the statistical robustness of our approach, we developed a permutation test to provide an uncertainty estimate for our results (Methods). In brief, we asked if the topological feature of a dataset could be explained by chance. To answer this question, we set out to test the null hypothesis that the difference between the lifetime of the maximal H_1 feature of a cyclic dataset and a tree-like dataset was null, versus the alternative of the maximal H_1 feature being more prominent in the cyclic dataset. Interestingly, we found that the difference between the simulated cyclic data and the tree dataset was significant (P-value $< 10^{-4}$). In the SI we show a systematic evaluation of this approach using both positive and negative controls (Fig S2). Together, these results demonstrate that our approach can robustly detect the topological signature corresponding to cyclic gene expression, even in the presence of noise inherent to single-cell data.

Next, we sought to identify the specific genes contributing to the detected topological loop, or the `topoGenes`. While persistent homology detects the presence of topological features, distinguishing the genes responsible for these features requires additional analysis. We leverage a topological optimization approach to rank genes based on their contribution to the prominent homology class. We identify topology-driving genes by computing the norm of the gradient of a persistence-based loss function with respect to gene expression coordinates, effectively measuring each gene's contribution to the prominent topological features in the cell state manifold.

Applying this approach to our simulated data, we successfully identified the transcription factors and target genes contributing to the topological loop, while excluding housekeeping genes that did

not influence the topology (Figure S1D). These results suggest that our approach can both identify the topological structure and its underlying gene expression signature of a cell state manifold.

Benchmarks of topological permutation test using synthetic data

In order to evaluate the effectiveness of our topological analysis, we conducted simulations of scRNAseq datasets incorporating predetermined data topologies and subjecting them to our topological statistical test. For our simulation, we used dyngen, a method that uses the Gillespie algorithm and real data statistics (such as capture rates and library sizes) to mimic the acquisition process of scRNAseq data.

To establish a baseline, we a null hypothesis dataset with a simple bifurcation tree topology. To assess the performance of our method, we performed a positive control experiment with the cyclic gene regulatory topology, and found a significant difference compared to the max H1 lifetimes of the control bifurcation dataset (P-value $< 10^{-4}$). We also performed negative control experiments featuring trifurcation, linear trajectory and binary tree topologies. Our statistical test revealed no significant differences in these datasets: trifurcation P-value = 0.4, linear trajectory P-value = 0.21, binary tree P-value = 0.54.

Finally, to verify that the test had low false discovery rate, we asked if the second most prominent H1 feature of the the dyngen cyclic topology was significant. For this case, we found that a P-value = 0.24, indicating that the test identifies a single significant topological feature as expected.

For all our experiments, we used 10^4 permutation replicates and 20 principal components for all of our experiments. These parameters were consistently applied across all experiments to ensure consistency and reliability of our results.

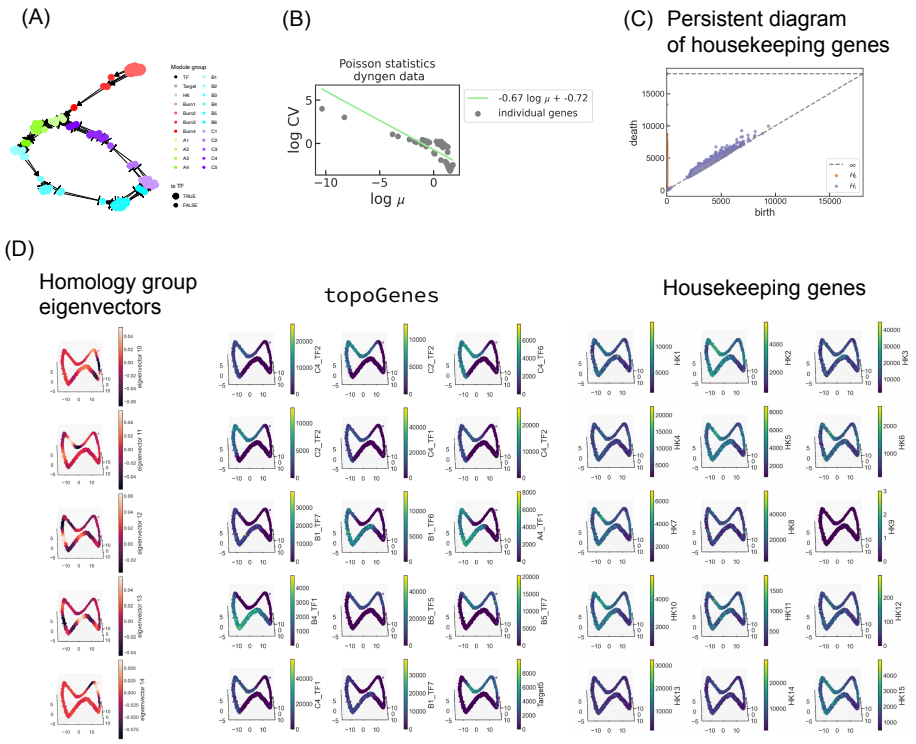

Figure S1: **Benchmark using a known gene regulatory network.** (A) Gene regulatory network used for benchmark experiments. (B) Dyngen data has Poisson statistics. Observed slope of $\log \mu$ vs logCV is -0.67 which is close to the predicted -0.5 of Poisson statistics. (C) Persistence diagram using only housekeeping genes. Note that no salient persistent $1-$homology classes are present. (D) Left: Eigenvectors of the $0-$Laplacian of homology generator. Middle: `topoGenes` with the highest mutual information for the Laplacian eigenvectors on the left. Please note that the gene expression patterns are transient. Right: Examples of housekeeping genes; note that their expression is spurious or constant.

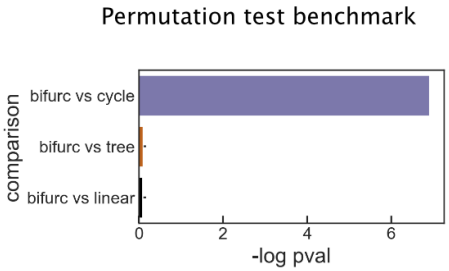

Figure S2: **Benchmarks of topological statistic tests using synthetic data.** We performed 3 control experiments to evaluate the efficacy of our statistical test. We used a simple bifurcation tree as a null hypothesis dataset for all experiments. For a positive control, we tested a cyclic dataset as a test dataset and found that the difference between maximal lifetime of H_1 classes was significant (P-value $< 10^{-4}$). In contrast a linear, another binary tree, and a trifurcation datasets where all deemed to have a non-significant difference between the maximal H_1 classes (P-values = 0.21, 0.54, 0.4 respectively).

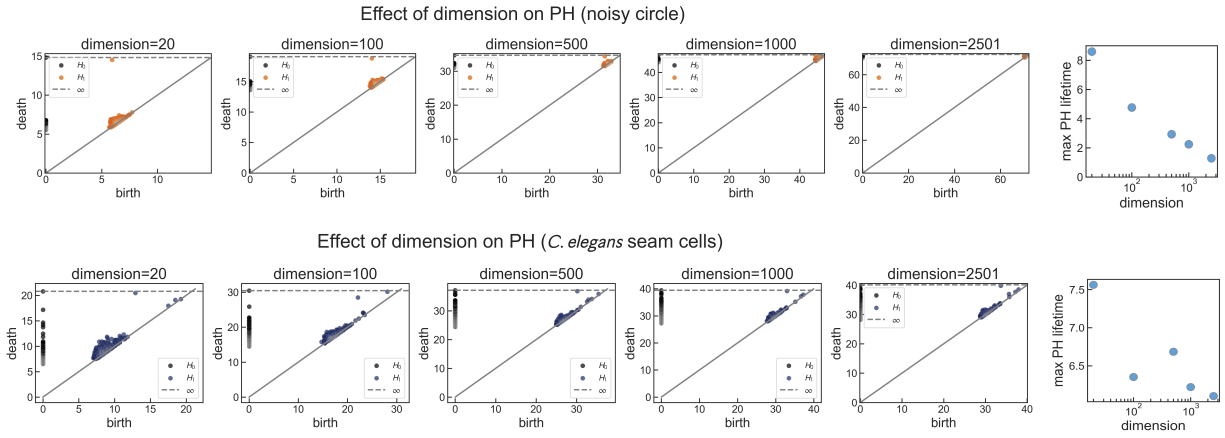

Figure S3: **Increasing dimension can diminish topological signal.** We studied the effect of increasing dimension on persistent homology calculation using a probabilistic model of a circle embedded in high dimensions (top), and using real data with increasing principal components (bottom). In the right we plot the lifetime of the most persistent homology class for each dataset. Note that the maximum lifetime (topological signal) decreases with increasing dimension.

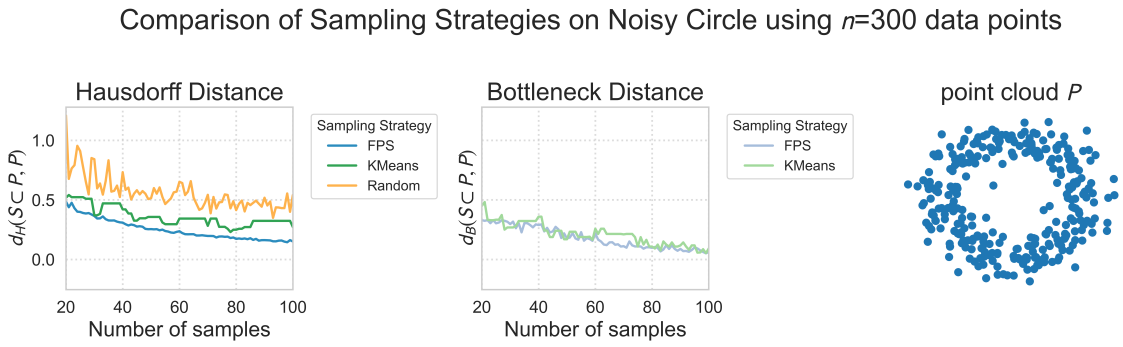

Figure S4: **Numerical comparison of sampling strategies for PH computation.** Quantitative comparison of sampling quality using Hausdorff distance (left) and bottleneck distance between persistence diagrams (middle) as a function of the number of samples (x-axis). Farthest point sampling (FPS) consistently yields lower distances compared to KMeans and random sampling. As expected, the Hausdorff distance d_H provides an upper bound on the bottleneck distance d_B across all strategies.

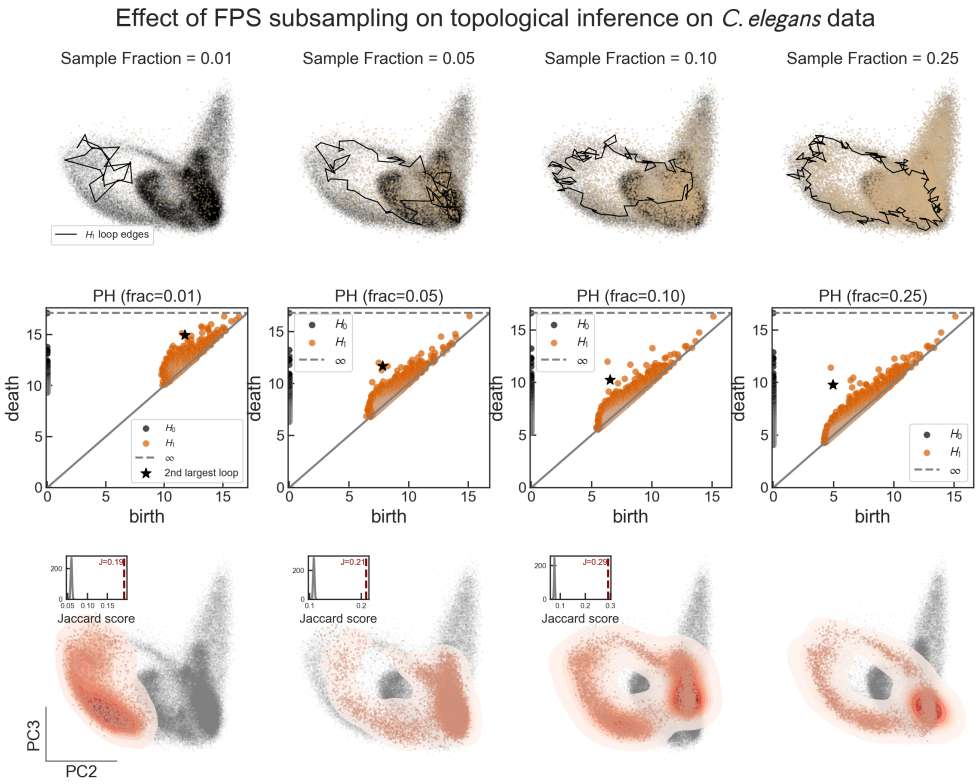

Figure S5: **Simulation on real scRNA-seq atlas reveals that of topological inference is feasible in the approximate regime of topological census.** Top row: Cells selected by FPS (highlighted in beige), with the second most persistent homology loop identified using the critical edge method. Middle row: Persistence diagrams for different sampling fractions, highlighting the second most persistent loop. Bottom row: Kernel density estimation (KDE) visualizations of `topoCells` distributions. Insets depict null distributions of Jaccard scores obtained by random sampling (grey histograms), contrasted with the observed Jaccard scores (maroon dashed lines), demonstrating that inferred `topoCells` significantly outperform random expectations.

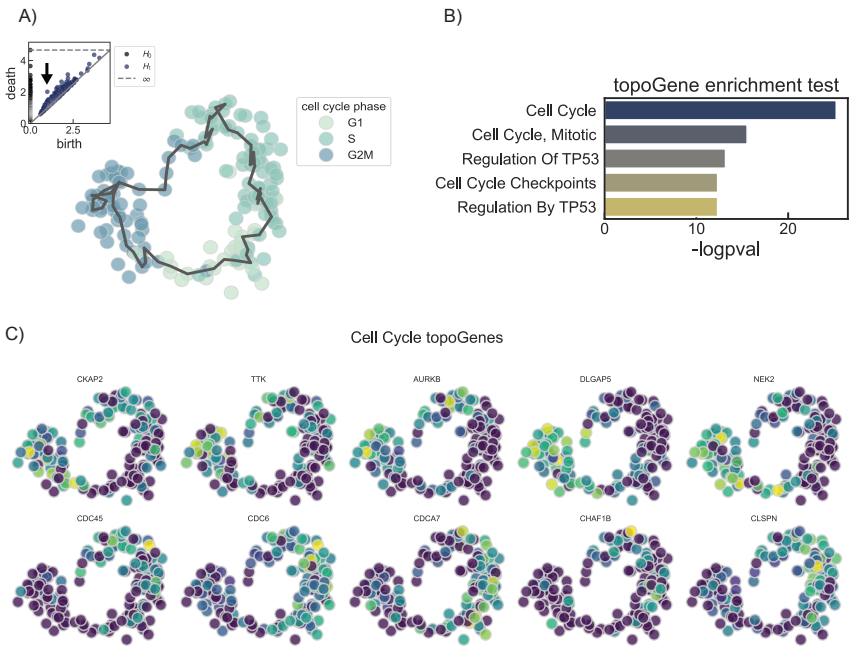

Figure S6: **Cyclic topology identification using `totopos` reveals cell cycle progression in proliferating HeLa cells.**

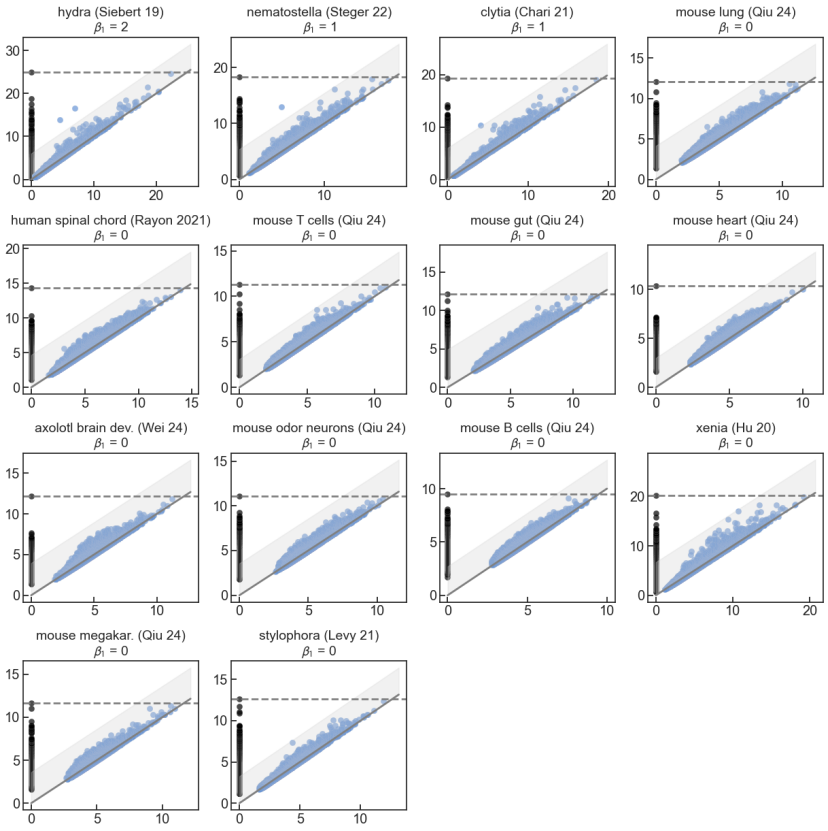

Figure S7: **Persistence diagrams for exact region of the topological census** Persistence diagrams for data sets containing less than 50,000 single-cell transcriptomes. Persistence computation was performed using pyRipser.

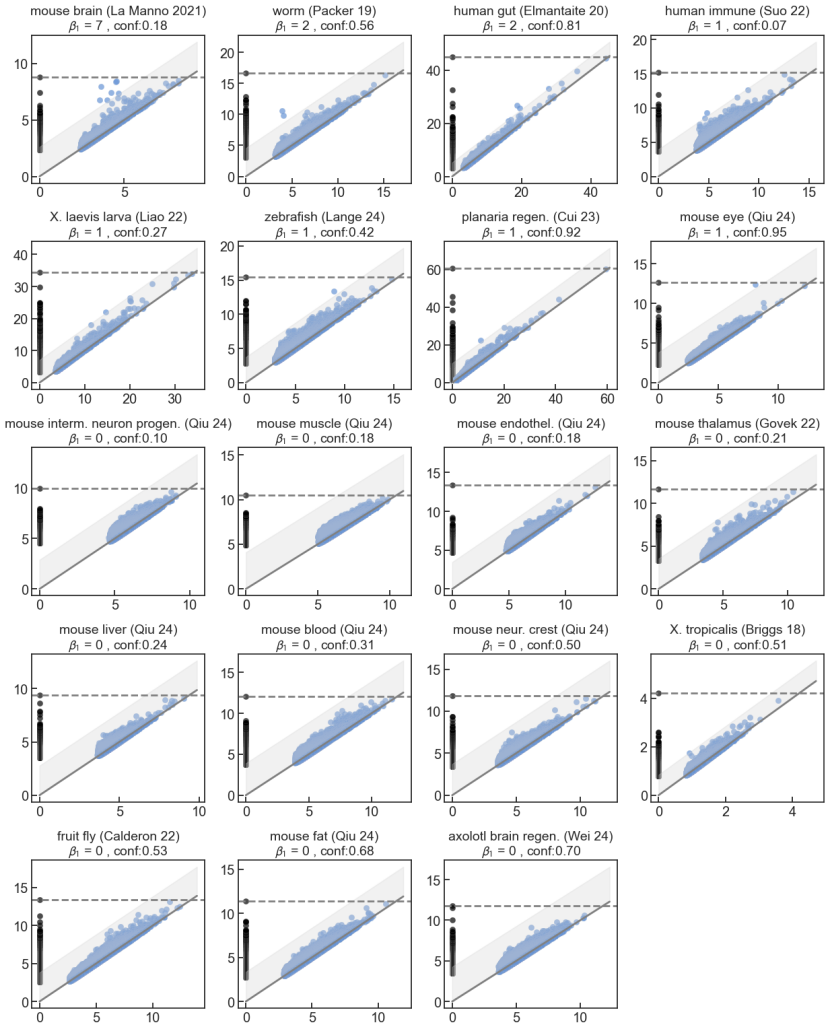

Figure S8: **Persistence diagrams for approximate region** Persistence diagrams for data sets containing more than 50,000 single-cell transcriptomes. Black points indicate H_0 homology classes, and blue points indicate H_1 classes respectively. Persistence computation was performed using a sparse filtration in pyRipser.

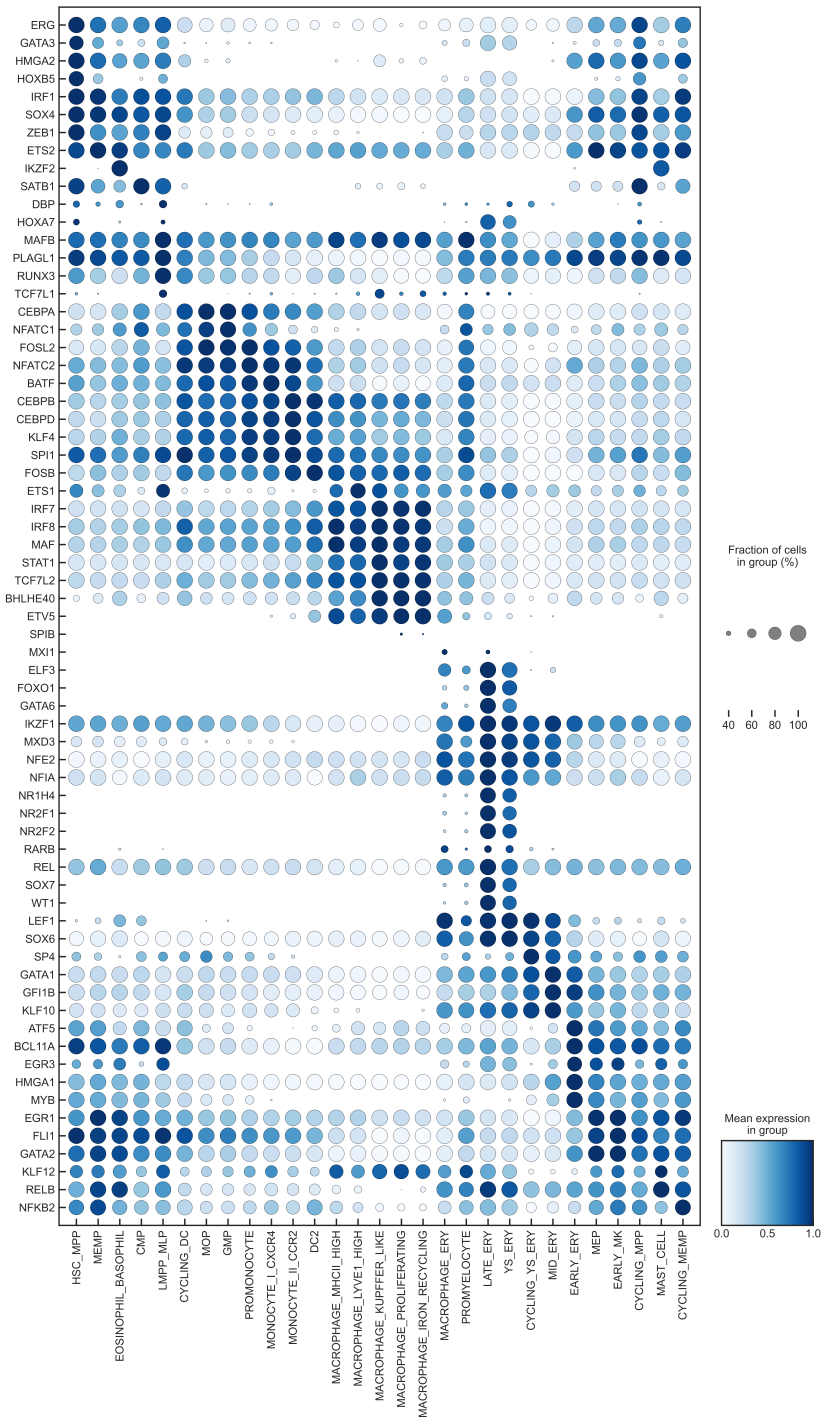

Figure S9: **Transcription factor regulon activities across immune cell populations based on SCENIC output.** Dotplot of transcription factor (TF) regulon activity across the circular coordinates, with cells grouped by original labels as in Suo et al., 2022b. Dot size represents the fraction of cells in each group expressing the regulon, and color intensity indicates the mean regulon activity within each cell group.

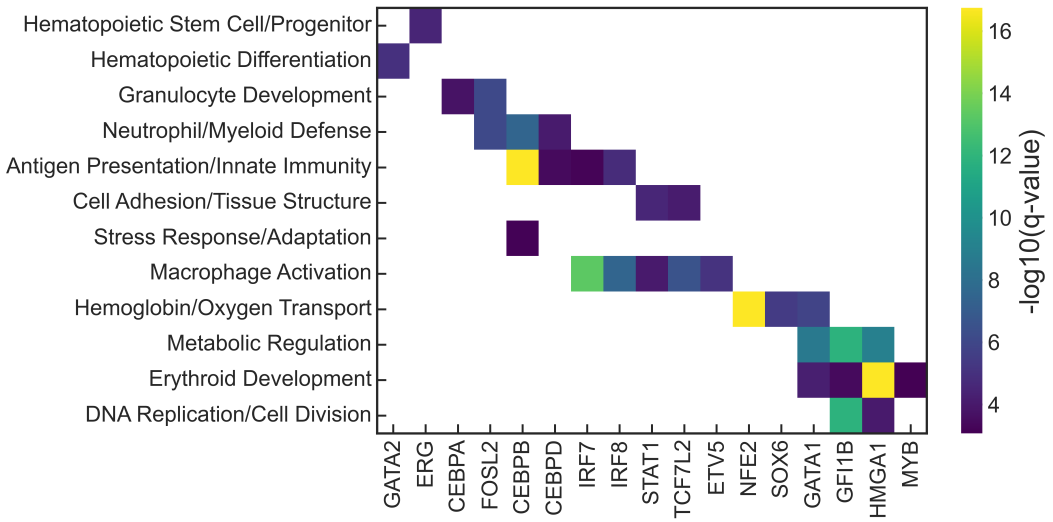

Figure S10: **Regulon enrichment within immune-related gene programs.** Heatmap illustrating statistically significant regulatory relationships between transcription factor (TF) regulons (columns) and immune-related gene programs (rows). Color intensity represents the negative log-transformed false discovery rate (FDR)-adjusted q-values from Fisher's exact tests, with brighter colors indicating stronger enrichment and greater statistical significance. Only associations with a q-value ≤ 0.001 are shown.

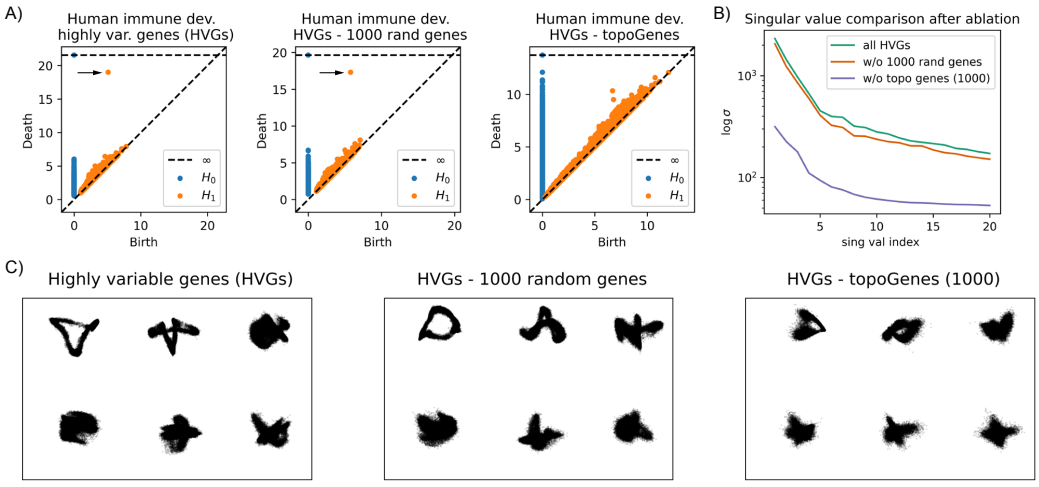

Figure S11: **Ablation experiment shows that topoGenes drive the topological structure in early human immune development.** A) Persistent diagrams using all highly variable genes (HVGs) or removing 1000 random genes are almost indistinguishable (left, middle). Removing `topoGenes` ablates the homology class. (B) The tolerance upon ablation is confirmed by a major decrease in singular value magnitude upon `topoGene` removal, indicating that the `topoGenes` constitute dominant gene programs. (C) PCA visualization indicates that the homology class is part of a dominant program of seam cells. We plotted `topoCells` in PCA space using a three-dimensional sliding window with all genes, after removing a random gene set, and after removing `topoGenes`.

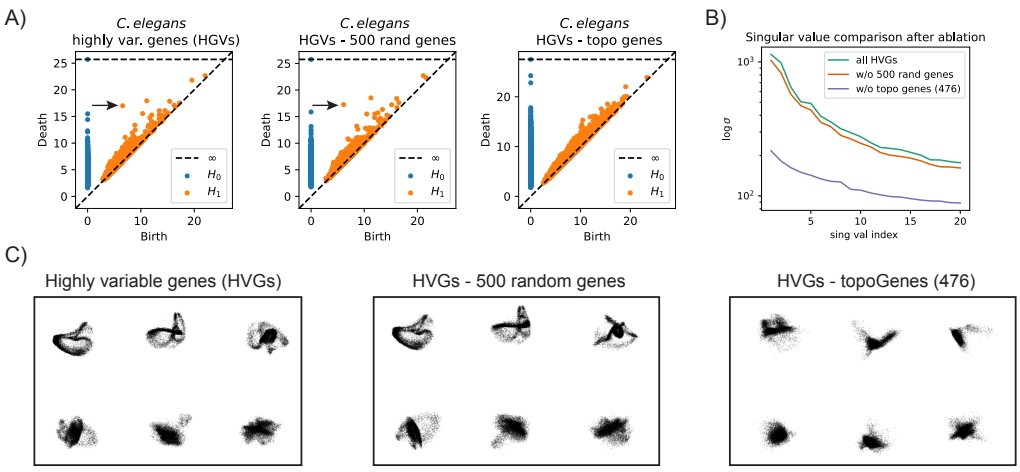

Figure S12: **Ablation experiment shows that `topoGenes` drive the topological structure in *C. elegans*' seam cells.**

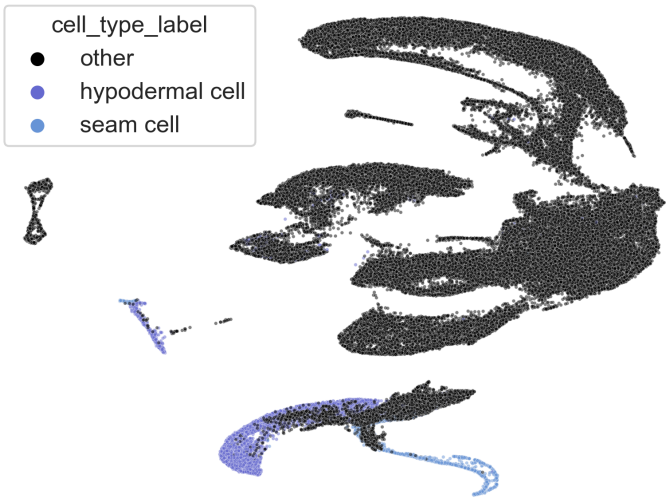

Figure S13: **Seam cell loop of *C. elegans* can be destroyed with UMAP.** We performed UMAP with default parameters on the *C. elegans* developmental atlas using as input the PCA projection using 20 principal components.

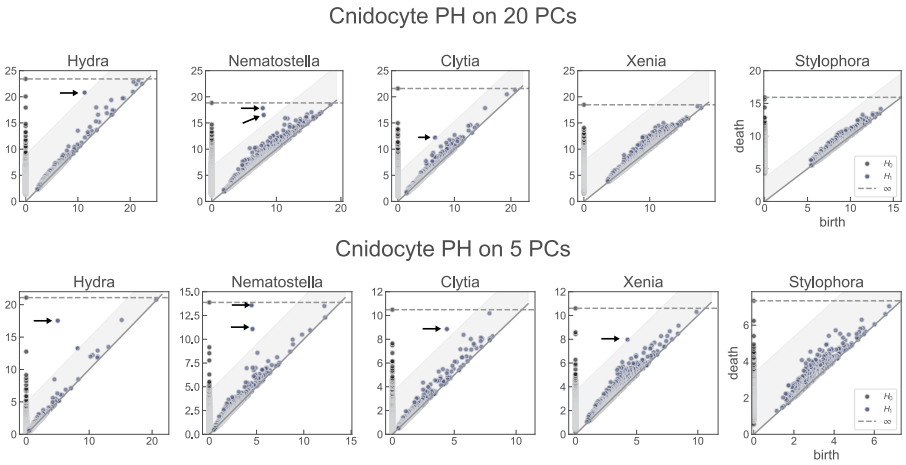

Figure S14: **Persistent diagrams of cnidocytes across cnidarian species.**

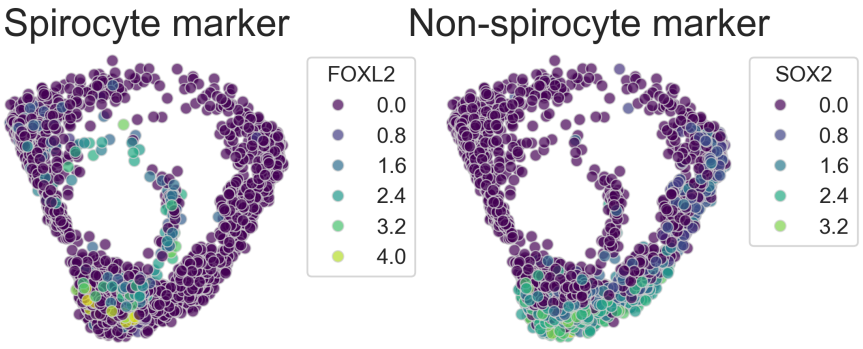

Figure S15: **A trajectory involving an ensnaring cell type, spirocytes, causes an extra loop in *Nematostella vectensis*.**

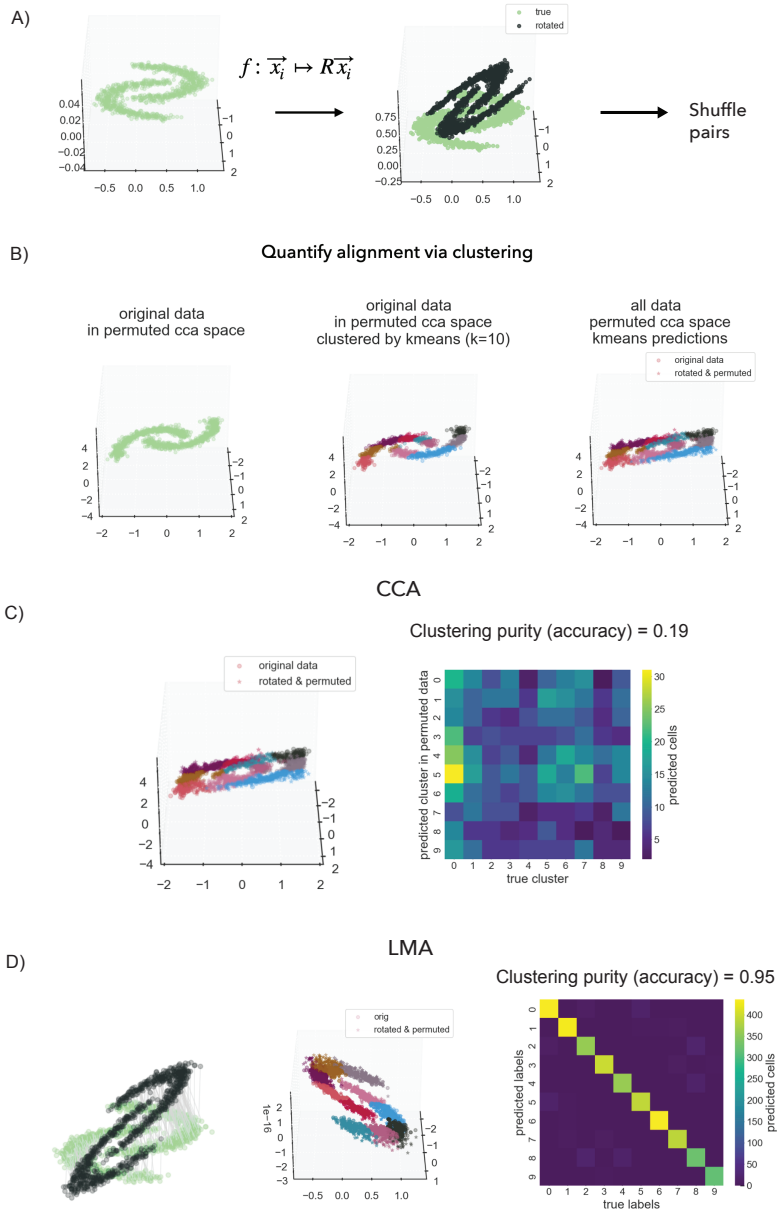

Figure S16: **Benchmarking LMA using simulation.** (A) We simulated a linear transformation followed by shuffling to test LMA. (B) Applying CCA without first considering the pairing caused a shared low-dimensional space, though an imperfect matching due to shuffling of the pairing labels. We used kMeans to quantify the accuracy, by assigning each rotated point in the shared space to its nearest original cluster. (C) Confusion matrix of clustering. (D) Visualization of mutual nearest neighbors (left). After correct pairing, CCA shows a good performance, as quantified by the clustering purity.

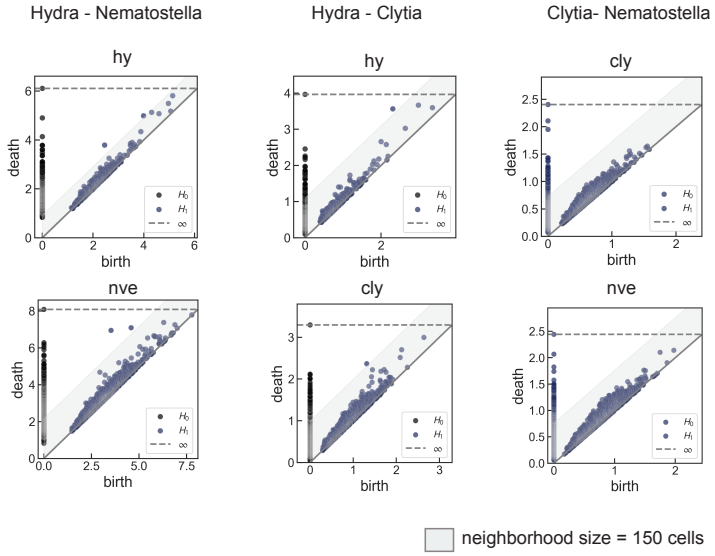

Figure S17: **Cnidarian persistent homology plots in LMA space.**

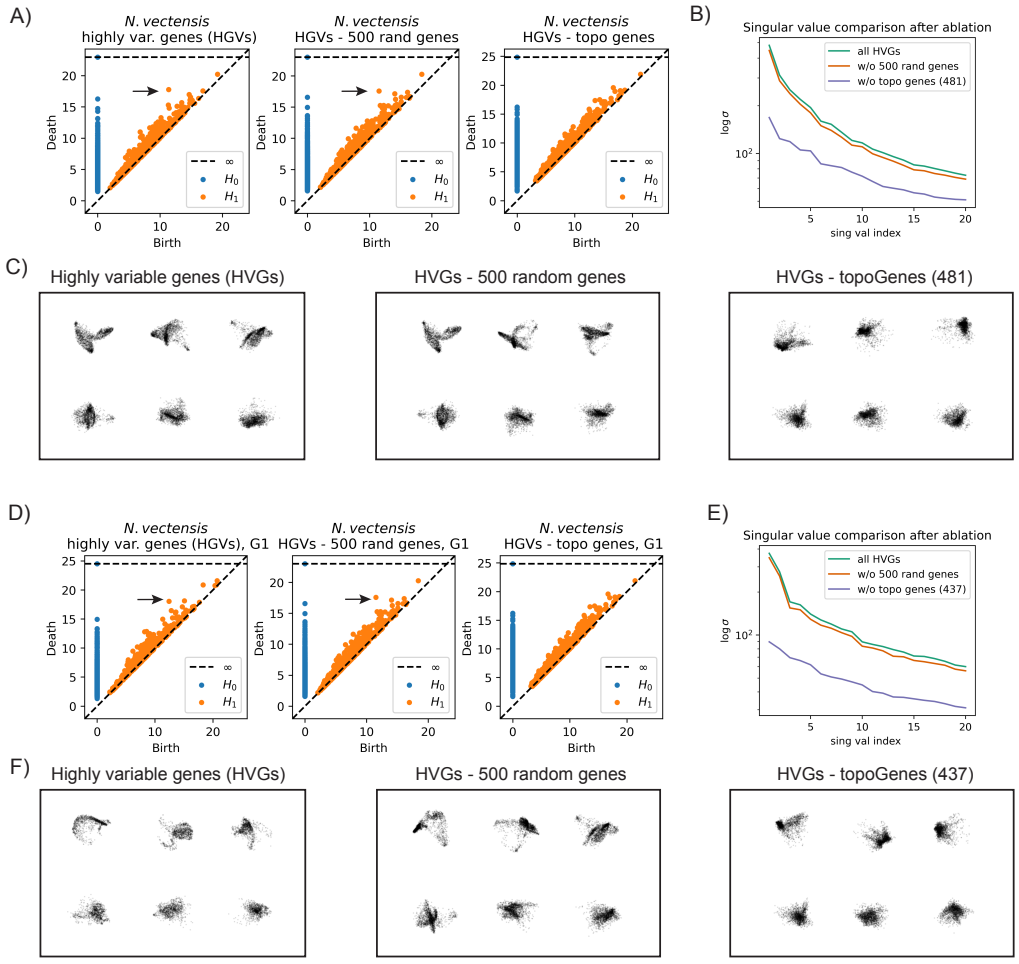

Figure S18: **Ablation experiment shows that `topoGenes` drive the topological structure in *N. vectensis* cnidocytes.** Results are for the two generators in nematostella (G0 top, G1 bottom).

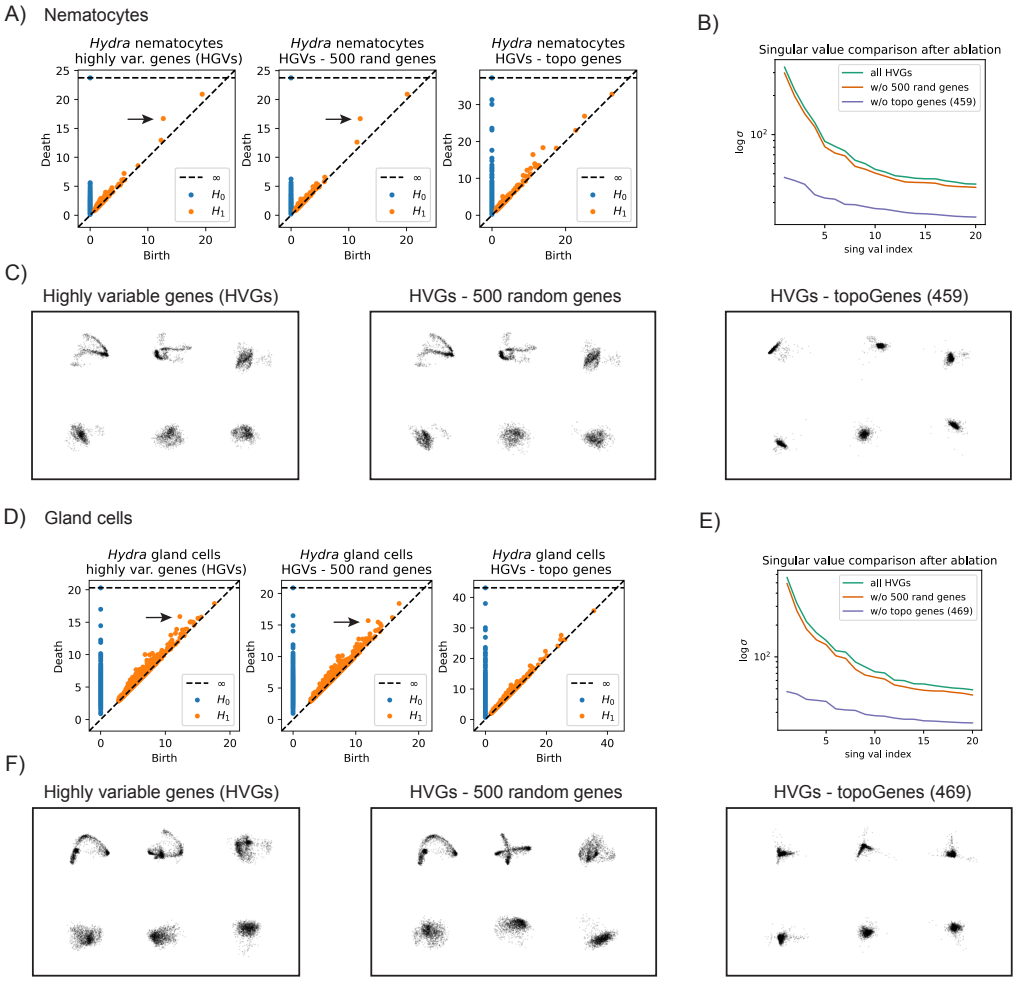

Figure S19: **Ablation experiment shows that `topoGenes` drive the topological structure in *H. vulgaris* cnidocytes and gland cells.**

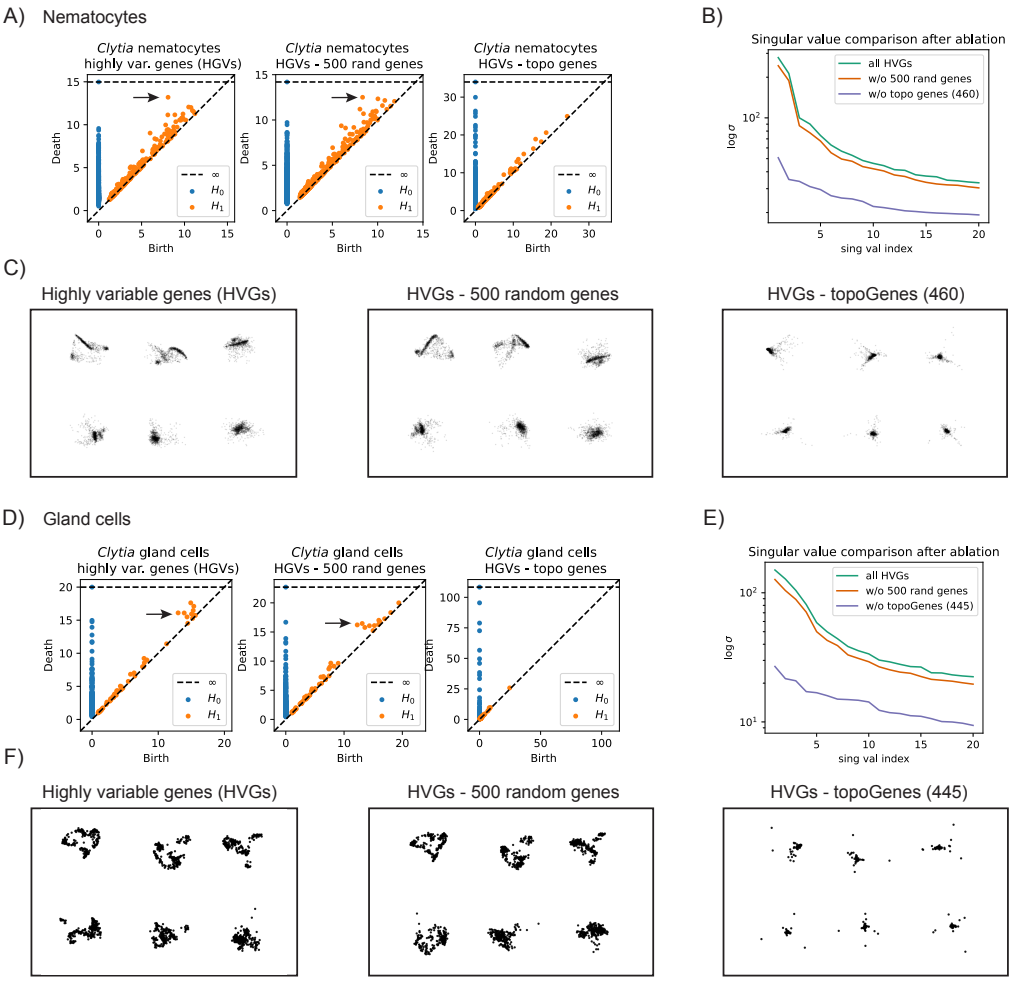

Figure S20: **Ablation experiment shows that `topoGenes` drive the topological structure in *C. hemisphaerica* cnidocytes and gland cells.**

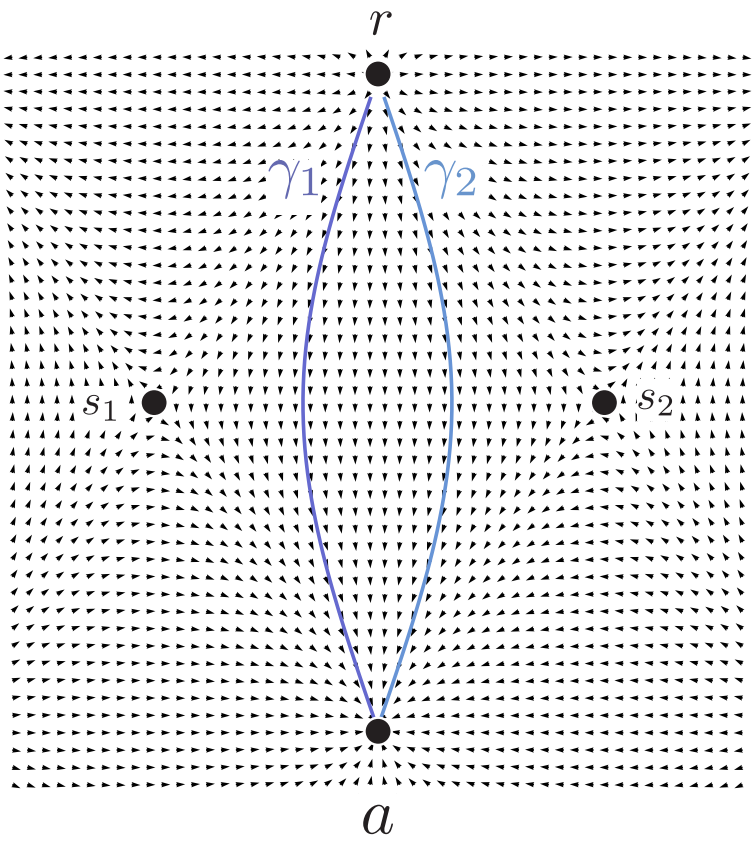

Figure S21: **Gradient systems can contain convergent trajectories.**

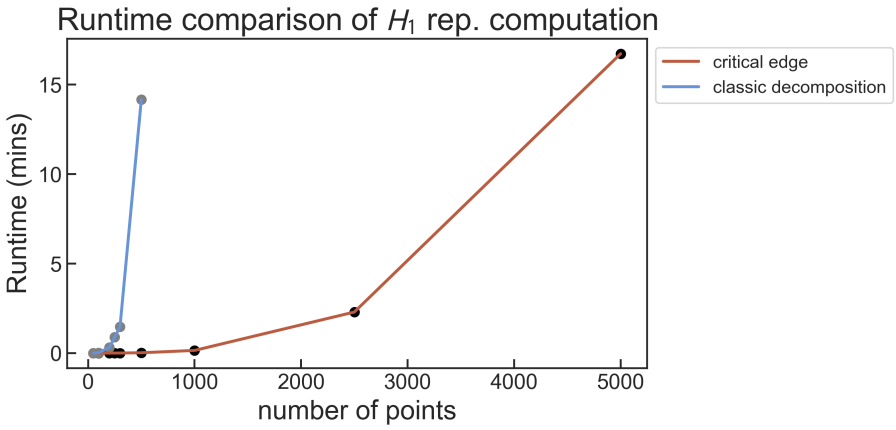

Figure S22: **Homology group generator runtime comparison.** We compared our method to the classical method for computing persistent homology, which yields persistent homology representatives, as implemented in Dionysus (`https://github.com/mrzv/dionysus`).

Chapter 3

MATHEMATICAL FRAMEWORK

3.1 Introduction

The goal of this chapter is to introduce the mathematical foundations underlying the tools developed in this thesis. We begin with a review of algebraic structures that lead to homology theory, proceed to define simplicial complexes and the computation of Betti numbers, and conclude by introducing persistent homology and cohomology as tools to extract topological motifs from single-cell data. By writing this theoretical primer I hope that interested computational biologists can not only use the tools I developed here, but also build new tools to find new insights in single-cell genomics and beyond. Most of the material presented here can be found in (Munkres, 2000; Hatcher, 2001) and the author does not claim original authorship of almost all results. The only case of original work in this section is the Fundamental Theorem of Chain Groups, although it is an elementary extrapolation of Hodge Decomposition for Simplicial Complexes. This chapter is only intended for self-containment of the thesis.

3.2 Algebraic preliminary

Definition. Relation

Let A be a set. A relation R on A is a subset $R \subset A \times A$ with $a \sim b$ for $(a, b) \in R$.

Definition. Equivalence relation

A relation R is an equivalence relation ($\sim$) if the following properties hold:

1. *Reflexivity* $a \sim a \, \forall \, a \in A$
2. *Symmetry* $a \sim b \Rightarrow b \sim a$
3. *Transitivity* if $a \sim b$ and $b \sim c \; \Rightarrow a \sim c$

Definition. Equivalence class An equivalence class is the set defined as:

$$[a] = \{x \in A : x \sim a\}. \tag{3.1}$$

Proposition. Let A be a non-empty set, $\sim$ an equivalence relation on A. The distinct equivalence classes partition A into disjoint sets.

Proof:

As $a \sim a$, then $\forall a \in A \, \exists [a] : a \in [a] \Rightarrow A = \bigcup_{a \in A} [a]$. To finish the proof we now show that equivalence classes are either disjoint or equal. Let $a, b \in A$ and assume their equivalence classes are not disjoint, i.e. $[a] \cap [b] \neq \emptyset$, we need to show that they are equal. Let $x \in [a] \cap [b] \Rightarrow a \sim x, x \sim b \Rightarrow a \sim b$ by trasitivity. Now assume $y \in [a]$, i.e. $y \sim a$. Then by our previous observation and applying transitivity $y \sim b$ and $y \in [b] \Rightarrow [a] \subset [b]$. By a similar argument we can show that $[b] \subset [a] \Rightarrow [a] = [b]$.■

Exercise: Can you show that the converse is true, i.e. that a partition also induces an equivalence relation. With this result, notice, that a clustering algorithm effectively induces an equivalence relation of the input data.

It turns out that, when working with a set with some additional algebraic structure (e.g. a vector space, or a group), the equivalence classes inherit the algebraic structure. The space of equivalence classes is called a quotient from the analogy that we reduce or *factor* the initial set, by focusing at equivalence classes. Our main interest is to define the concept of a homology group, which is a quotient *group*. Thus, we begin by describing what is a group in the context of modern algebra.

Definition. Group A group is a tuple $(G, \cdot)$ where G is a set and "$\cdot$" is a closed binary operation on G, i.e. $\cdot : G \times G \rightarrow G$ and for any $x, y \in G$, $x \cdot y \in G$. It has the following axioms:

1. *Associativity*: For all $a, b, c \in G$ one has that $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

2. *Identity element*: There is an unique element $e \in G$ s.t. for every $a \in G$ one has $e \cdot a = a = a \cdot e$. This element is called the *identity element*.

3. *Inverse element*: For each $a \in G$ there is an unique element a^{-1} s.t. $aa^{-1} = e = a^{-1}a$.

Example: The integers form a group under addition.

Note: It's interesting to think what happens when we remove structure from a group. For example, a set that is closed under binary operation is called a magma. In particular, in a magma, the binary operation need not be associative. An example of a magma is $(\mathbb{R}^3, \times)$, i.e. three-dimensional Euclidean space equipped with the cross product as binary operation. Clearly, this structure is closed under the operation. *Is the operation associative?* It turns out the answer is *no*. To prove it we can just use a counter example: Let $u = \hat{i} = (1, 0, 0)^T, v, w = \hat{j} = (0, 1, 0)^T$. Then $u \cdot v = \hat{k} = (0, 0, 1)$ and $v \cdot w = 0$. Thus $(u \cdot v) \cdot w = -\hat{i} \neq u \cdot (v \cdot w) = 0$. *Exercise:* Can you show that $(\mathbb{R}^3, \times)$ doesn't have an identity element? You can answer in purely geometric terms, i.e. no mathematical symbols.

Remark: If the operation on the group is commutative, i.e. $a \cdot b = b \cdot a$, we say that the group is commutative or *abelian*. Our groups of interest, the simplicial chains and the homology groups are abelian.

Example: The set of nonsingular matrices of dimension $n \times n$ forms a group under matrix multiplication called the general linear group of dimension n, denoted $GL(n)$. *Question*: is the group commutative?

Definition. Subgroup Let G be a group. The subset H of G is a subgroup (written $H \leq G$) if H is nonempty and is closed under products and inverses.

Remark: The smallest subgroup H is the set containing only the identity element of G.

Example: The set of orthogonal matrices of dimension $n \times n$ is a subgroup of $GL(n)$.

Definition Let G be a group, $H \leq G$. Then this defines an equivalence relation on G for $(a, b) \in G$: $a \sim b$ if $\exists\, h \in H \ : a = bh$. The equivalence classes are also called *cosets*.

Proof:

It suffices to show that the definition above has all the propeties of an equivalence relation:

1. Reflexivity: $a \sim a$, since $e \in H$ and $a = ae$.
2. Symmetry: Let $a \sim b$, then $\exists h \in H : a = bh$. But also $b = ah^{-1}$ and $h^{-1} \in H \Rightarrow b \sim a$.
3. Transitivity: Let $a \sim b\,, b \sim c$. Then $\exists h, h' : a = bh, b = ch' \Rightarrow a = ch'h = ch''$ by closure $h'' \in H \implies a \sim c$. ■

Definition. Coset For all $g \in G$ the set $aH = \{ah : h \in H\} = [a]_H$. I.e. a coset is just an equivalence class in the context of group theory.

Definition. Quotient group Let G be a *abelian* group, and $H \leq G$. Then the *quotient group* G/H is the group of cosets of H:

$$G/H := \{aH : a \in G\} \tag{3.2}$$

with operation:

$$(aH)(bH) := abH \tag{3.3}$$

and the operation is well defined.

Remark: To be more explicit about equality of cosets, recall that non-identical cosets are disjoint. Thus $gH = xH$ means that there is an $h \in H$ such that $g = xh$. Thus $gH = x(hH) = xH$, since hH is the same coset as $eH = H$ since H is closed.

Proof:

We need to show that the operation is well-defined, i.e. that it is independent on the choice of

representatives. We will show this in the context of abelian groups. For a general proof one needs to show the additional structure of H being a normal subgroup.

Let $a, b, a', b' \in G$ with $aH = a'H, bH = b'H$. We need to show that $(aH)(bH) = (a'H)(b'H)$.

We have that by definition since $a \sim a'$, i.e. $\exists h, h' \in H : a = a'h, b = b'h'$.

Then

$$aHbH := abH \tag{3.4}$$

$$= a'hb'h'H \tag{3.5}$$

$$= (a'h)(b'h'H) \tag{3.6}$$

$$= (a'h)b'H \text{ since H is closed under the operation} \tag{3.7}$$

$$= a'b'hH \text{ by commutativity} \tag{3.8}$$

$$= a'b'H \tag{3.9}$$

$$= (a'H)(b'H)\blacksquare \tag{3.10}$$

Remark: Since all subgroups of abelian groups are normal, and we're exclusively interested in these, we will not make the general proof for nonabelian groups.

Example: The integers mod 2 $\mathbb{Z}_2$ are a quotient group under the sum. The equivalence classes or cosets are even and odd integers.

Example: Quotient vector space Let V be a vector space and U be a subspace. The quotient space V/U is a quotient group over addition. For concreteness consider $V = \mathbb{R}^2, U = \{(x, x), x \in \mathbb{R}\}$, that is U is a line through the origin with slope 1. In this set up, two vectors x, y are equivalent iff $x - y \in U \iff \langle \frac{x-y}{||x-y||}, (1, 0)^T \rangle = \cos(\frac{\pi}{4})$. Cosets $v + U$ are parallel translates of U. Addition in the quotient group is defined as $(v + U) + (w + U) = (v + w) + U$, and (U) is the identity coset.

Exercise: Visualize the quotient group above, and prove that the operation is well defined.

3.3 Set theory preliminaries

Definition 1. Condition for containment in a union of an indexed family of sets

Let $\{X_i\}_{i\in I}$ be an indexed family of sets. We say that $x \in \bigcup_{i\in I} X_i$ if there is at least one $i \in I$ such that $x \in X_i$. Conversely $x \notin \bigcup_{i\in I} X_i$ if $x \notin X_i \forall i \in I$.

Definition 2. Condition for containment in the intersection of an indexed family of sets

Let $\{X_i\}_{i\in I}$ be an indexed family of sets. We say that $x \in \bigcap_{i\in I} X_i$ if $x \in X_i \forall i \in I$. Conversely $x \notin \bigcap_{i\in I} X_i$ if $\exists i \in I$ for which $x \notin X_i$.

One can infer that there's a simple way of connecting these two definitions by their duality; we show this in the following Theorem.

Proposition. De Morgan's Laws

1. $X - \bigcap_{i\in I} A_i = \bigcup_{i\in I}(X - A_i)$
2. $X - \bigcup_{i\in I} A_i = \bigcap_{i\in I}(X - A_i)$

Proof: We'll only prove one direction of the containment. The second one is easily achieved by a symmetric argument. (1) Let x be in the l.h.s. so that $x \in X$ and $x \notin A_i \forall i \in I$ by the first part of Definition 2. This means that there is an i s.t. $x \in (X - A_i) \implies x \in \bigcup_{i\in I} X - A_i$. Thus l.h.s. $\subset$ r.h.s.

(2) Let x be in the l.h.s., so $x \in X$ and $\exists i : x \notin A_i$ by Definition 1. This means that $x \notin A_i \forall i$ and $x \in X$, i.e. $x \in (X - A_i) \forall i \implies x \in \bigcap_{i\in I}(X - A_i)$. Thus the l.h.s $\subset$ r.h.s.

3.4 (Point-set) Topology

Here, we provide a concise introduction to the concepts of point-set topology. Historically, point-set topology was developed by the need of formalizing the emerging field of algebraic topology (first called *Analysis situs*), and calculus (which lead to the multiple branches of analysis). This section can be skipped during a first read of this material, but I consider it a minimal (hopefully helpful) primer on the basic structure of a topological space.

Def. Topology on a set

A topology on a set X is a family τ of subsets of X with the following properties:

1. $\emptyset$ and X are in τ.
2. Closed under arbitrary unions: The union of any subcollection of subsets in τ is in τ.
3. Closed under finite intersections: The intersection of the elements of a finite subcollection of τ is in τ.

Example. Open balls generate a basis for a topology in $\mathbb{R}^n$. This is called the standard topology in euclidean space.

A concise decription of a topology is given by the following definition.

Def. Basis for a topology Let (X, τ) be a topological space. A basis for a topology on X is a collection $\mathcal{B}$ of subsets of X such that:

1. For each x in X, $\exists B \in \mathcal{B} : x \in B$.
2. If x belongs to the intersection of two basis elements B_1, B_2, then $\exists B_3 : B_3 \subset B_1 \cap B_2$.

If $\mathcal{B}$ satisfies these two conditions, we define the topology τ generated by $\mathcal{B}$ as follows: A subset U of X is said to be open in X (i.e. an element of τ) if for each $x \in U$, there is a basis element $B \in \mathcal{B}$ such that $X \in B \subset U$.

Remark: $B \in \mathcal{B} \implies B \in \tau$, i.e. each basis element $B \in \mathcal{B}$ is itself an element of the topology τ.

Proposition. A basis $\mathcal{B}$ generates a topology τ on X.

Proof: The empty set and X satisfy the conditions trivially. Now let's check that the topology generated τ contains (1) finite intersections of open sets and (2) arbitrary unions.

(1) Let $U = U_1 \cap U_2$, take $B_1 \subset U_1, B_2 \subset U_2 : x \in B_1 and B_2 \implies x \in B_1 \cap B_2 \subset U_1 \cap U_2$. By the second condition for a basis $\exists B_3 \subset B_1 \cap B_2 : x \in B_3 \implies U_1 \cap U_2 \in \tau$ by definition of

topology generated by $\mathcal{B}$. We show that a finite intersection $U_1 \cap ... \cap U_n$ of elements of τ is in τ by induction. The fact is trivial if $n = 1$. Suppose it is true for $n - 1$, then it remains to show that the result is true for n. We know that:

$$(U_1 \cap ... \cap U_n) = (U_1 \cap ... \cap U_{n-1}) \cap U_n. \tag{3.11}$$

By hypothesis $U_1 \cap ... \cap U_{n-1} \in \tau$. Let $U_a = U_1 \cap ... \cap U_{n-1}$, $U_b = U_n$, then by the result proven above $U_a \cap U_b \in \tau$ as desired.

(2) Let U be an arbitrary union of open sets indexed by a set I. Let $x \in U \implies \exists i \in I : x \in U_i$. By definition of basis $\exists B \in \mathcal{B} : x \in B \subset U_i \subset U$. Thus U is open by definition of topology generated by basis. ■

Lemma. $\tau = \bigcup B : B \in \mathcal{B}$

Proof: Elements of $\mathcal{B}$ are also elements of τ, and since τ is a topology, the union of elements in $\mathcal{B}$ is in τ, so $\tau \supset \bigcup_{b \in \mathcal{B}} B$. Conversely, let $U \in \tau$ and for each $x \in U$ assign an element $B_x \in \mathcal{B} : x \in B_x \subset U$. $\implies U = \bigcup_{x \in U} B_x \implies \tau \subset \mathcal{B}$.

Def. Topological space

A topological space is a tuple (X, τ) consisting of a set X and a topology τ on X.

Example. Topological space on a set {a,b,c }

Def. Open set Let (X, τ) be a topological space. We say that a subset U of X is an open set of X if $U \in \tau$.

Def. Closed set A subset A of X is said to be **closed** if the set $X - A$ is open.

It turns out that we can specify a topology on a space using closed sets too. This is achieved by the duality of sets using the De Morgan Laws.

Theorem. Topology using closed sets

Let (X, τ) be a topological space. We can define the topology on X using closed sets if the following properties hold:

1. $\emptyset$ and X are closed.
2. Arbitrary intersections of the closed sets are closed.
3. Finite unions of closed sets are closed.
4. Arbitrary intersections of the closed sets are closed.

Proof: (1) holds since they are the complements of X and $\emptyset$ respectively. (2) This follows from De Morgan: given a finite collection of closed sets $\{A_i\}_{i\in 1,\dots,n}$ we have that $X - \bigcup_{i=1}^{n} A_i = \bigcap_{i=1}^{n}(X - A_i)$. The sets $X - A_i$ are open by definition. Therefore the r.h.s. is open since finite intersection of open sets is open. Therefore $\bigcup_{i=1}^{n} A_i$ is closed.

(3) This follows from De Morgan: given an arbitrary collection of closed sets $\{A_i\}_{i\in 1,\dots,n}$ we have that $X - \bigcap_{i\in I} A_i = \bigcup_{i\in I}(X - A_i)$ The sets $X - A_i$ are open by definition, thus the r.h.s. is an arbitrary union of open sets and is thus open. It follows that $\bigcap iA_i$ is closed.

We can specify a topology τ by giving a collection of closed sets satisfying the above properties, and define open sets as the complements of closed sets. Therefore we showed a way to specify a topology τ on X using closed sets as desired. ■

With the algebraic and topological preliminaries we are now on the grounds to blend them together to arrive at algebraic topology. Our end goal is to arrive at homology groups, a computable topological invariant that reveals the structure of manifolds. Since homology is a homotopical invariant (meaning that, if two spaces are homotopically equivalente, then they have isomorphic homology groups), we will first have to introduce the concept of homotopy. This is a necessary concept to capture the flexibility in the continuous nature of topological spaces, that may not be readily apparent when looking at simplicial homology.

3.5 Homotopy is a formal way of describing continuous deformations

Our main goal for this section is to mathematically describe properties of continuous deformations that are more flexible than homeomorphisms. The problem with homeomorphism is that there is currently no general method of determining if two spaces are homeomorphic. Nevertheless, a good relaxation to homeomorphism is homotopy equivalence. The intuition is that two spaces having the same *shape* are homotopy equivalent. Therefore, despite them not being continuosly bijective, they can, in a broader sense, be deformed into each other. Unfortunately, computing the homotopy groups of a space is also hard, let alone inferring them from data. To our good fortune, there is an even weaker notion of homology equivalence that is weaker than homotopy equivalence but gives us a ton of information: if two spaces are homotopy equivalent, then they have isomorphic homology groups. It is because of this reason that homology groups have been of great use in mathematics, but also, more recently in topological data analysis.

I believe that, under this reasoning, it is very fruitful to first make a brief detour into the basics of homotopy theory before diving into homology. In a sense homotopy gives all the necessary intuition to think about equivalence in terms of *shape*. That is, homotopy equivalence is a strong version of **topological equivalence**.

Definition. Homotopy. Let $(f, g) : X \to Y$ be two continuous functions between topological spaces. A homotopy is a function $F : X \times [0, 1] \to Y$ that has the property:

$$F|_{X\times\{0\}} = f(x), F|_{X\times\{1\}} = g(x). \tag{3.12}$$

If such F exist, we say that f and g are homotopic.

Remark: **Informally, a homotopy is a continuous deformation** of f into g. Intuitively, we can think of the parameter $t \in [0, 1]$ as *time*, and visualize the deformation of the image of f into the image of g by varying the time parameter. In this sense a homotopy is a family of functions $f_t : X \to Y$ parametrized by t that is continuous in both arguments.

The following example will help visualize this deformation.

Definition. Path A path is a continuous map $f : [0, 1] \to X$.

Example: Any two maps $(f, g) : [0, 1] \to \mathbb{R}^n$ are homotopic. An example of a homotopy is the **linear homotopy**:

$$F(x, t) = t f(x) + (1 - t)\, g(x). \tag{3.13}$$

In fact this holds for any convex set $X \subset \mathbb{R}^n$. In this view, if we fix x and vary t

Exercise: Make a sketch of the linear homotopy above.

Proposition. Homotopy is an equivalence relation in the space of continuous functions from X to Y, $C(X, Y)$.

Proof:

- Reflexivity. $f \sim f$ by the *constant* homotopy $F(x, t) = f(x)$.
- Symmetry. $f \sim g \implies g \sim f$. Let $F(x, t)$ be a homotopy from f to g. A homotopy from g to f is given by *reversing* time $H(x, t) = F(x, 1 - t)$.
- Transitivity. $f \sim g, g \sim h \implies f \sim h$. Let F be a homotopy from f to g, G from g to h, and consider

$$H(x, t) = \begin{cases} F(x, 2t), t \in [0, 1/2] \\ G(x, 2t - 1), t \in (1/2, 1]. \end{cases}$$

Remark: We can thus consider the quotient $C(X, Y)/\!\sim = [X, Y]$ the space of homotopy classes.

Proposition. Compositions of homotopic maps are homotopic

Consider $f \simeq f' : X \to Y$ under homotopy F, and $g \simeq g' : Y \to Z$ under homotopy G. Then $g \circ f \simeq g' \circ f' : X \to Z$.

Proof: Construct homotopy $H(x,t) : X \times I \to Z = G(F(x,t),t)\blacksquare$.

Defintion. Homotopy equivalence A function $f : X \to Y$ is a homotopy equivalence if there exists $g : Y \to X$ such that:

$$g \circ f \simeq \mathrm{id}_X \tag{3.14}$$

$$f \circ g \simeq \mathrm{id}_Y \tag{3.15}$$

and g is the homotopy inverse of f. We then say that X, Y are *homotopy equivalent* or that they have the same *homotopy type*.

Example. $\mathbb{R}^n \simeq \{\mathrm{pt}\}$, in words, $n-$dimensional euclidean space is homotopy equivalent to a point, or contractible.

Let $f : \mathrm{pt} \mapsto 0$ be an embedding, $g : \mathbb{R}^n \to \{\mathrm{pt}\}$, where $g(x) = \mathrm{pt}$. First, note that $g \circ f = \mathrm{id}_{\{\mathrm{pt}\}}$ since $f(\mathrm{pt}) = 0, g(0) = \mathrm{pt}$. Now, we need to show $f \circ g \simeq \mathrm{id}_{\mathbb{R}^n}$. We can do this explicitly:

$F(x,t) = tx, F|_0(x) = 0 = f \circ g, F|_1 = \mathrm{id}_{\mathbb{R}^n}$. So indeed $f \circ g \simeq \mathrm{id}_{\mathbb{R}^n}\blacksquare$.

Proposition. Homotopy equivalence is an equivalence relation on the category of topological spaces.

Proof:

- Reflexivity. $X \sim X$ we can use $f, g = \mathrm{id}$.
- Symmetry. $X \sim Y \implies Y \sim X$. If $X \sim Y$ there is (f, g) s.t. g is homotopy inverse to f. Now relabel and assume g is the initial func, f the homotopy inverse.
- Transitivity. Let $X \sim Y, Y \sim Z$. By definition we have:

$$g \circ f \simeq \mathrm{id}_X, f \circ g \simeq \mathrm{id}_Y \tag{3.16}$$

$$k \circ h \simeq \mathrm{id}_Y, h \circ k \simeq \mathrm{id}_Z \tag{3.17}$$

Let us show that $g \circ k \circ h \circ f \simeq \mathrm{id}_X$. First, by transitivity we have that $k \circ h \simeq \mathrm{id}_Y \simeq f \circ g$, then

$$g \circ (k \circ h) \circ f \simeq g \circ (f \circ g) \circ f \text{ since composition of homotopic maps is homotopic} \tag{3.18}$$

$$\simeq g \circ (\mathrm{id}_Y) \circ f \tag{3.19}$$

$$= g \circ f \tag{3.20}$$

$$\simeq \mathrm{id}_X \tag{3.21}$$

where in the first line we meant that $h \sim h, k \circ h \sim f \circ g, f \sim f$. An equivalent argument can be made to show that $h \circ f \circ g \circ k \simeq \mathrm{id}_Z$. Thus $X \sim Z$ completing the proof ■.

Note: A question that may arise here is: what does it mean for two topological spaces to be homotopy equivalent? As we've mentioned in the introduction to this section, homotopy equivalence roughly means that two spaces have the same *shape*. The importance of the above proposition is that if we can show that $X \simeq Y$ and $Z \simeq Y$ for some space Y that is easy to compare to, then we can posit that $X \simeq Z$, in the scenario where X and Z are hard to compare. For example Y could be a model space such as a sphere or a torus. In what follows we'll show a very important way to obtain pairs of homotopy equivalent spaces, in particular between a subspace A of a larger space X.

Definition. Retraction A retraction of a topological space X onto a subspace A is a continuous map:

$$r : X \to A, r|_A = \mathrm{id}_A$$

such A is called a **retract** of X.

Proposition. Deformation retract $X \simeq A$ if there is a homotopy between id_X and $r' = \iota \circ r$. *Proof:* Clearly we already have one side to get a homotopy equivalence, namely $r \circ \iota = \mathrm{id}_A$. Now

we need to show that we can form a homotopy so that $\iota \circ r \simeq \mathrm{id}_X$. But note that one can construct a linear homotopy easily:

$$F(x,t) = tx + (1-t)(\iota \circ r)(x)$$

so that $F|_{t=0} = \iota \circ r$ $F|_{t=1} = \mathrm{id}_X$

A is called a **deformation retract of** X if such a homotopy equivalence holds.

Counterexample. If X is not path-connected $\nexists$ a deformation retraction $r : X \to \{pt\}$. One can show this by contradiction. Assume such a def. retraction exists, then there is a homotopy F such that :

$$F(x,0) = x, F(x,1) = x_0.$$

Now consider any point $u \in X$ and a map (which is a path) $\gamma = F|_{x=u} : [0,1] \to X$, $\gamma(0) = u, \gamma(1) = x_0$ since the point u was arbitrary this implies that X is path connected: a contradiction.

Remark: The converse doesn't hold: if X doesn't retract to a point X is not necessarily path-connected. A counterexample is any space with non-trivial homology for dimensions larger than zero. E.g. a circle doesn't retract to a point since there would be a discontinuity in the map. This leads to the following definition.

Def. Contractible space If X deformation retracts to a single point, such X is called contractible.

Example. $\mathbb{R}^n$ is contractible. Let $X = \mathbb{R}^n, A = x_0$. Then there is a retraction r:

$$r|_{\{x_0\}} = \mathrm{id}_{\{x_0\}}, r(x) = x_0 \forall x \in \mathbb{R}^n.$$

Example. Annulus deformation retracts to a circle. Let X be an annulus $X = \{x : 1/2 \leq ||x|| \leq 3/2\}$, $A = \mathbb{S}^1$. Then $r(x) = \frac{x}{||x||}$ is a deformation retraction.

The following example is really helpful as it provides a simple combinatorial algorithm to induce homotopy equivalence on abstract simplicial complexes. For a definition of simplicial complexes, see the next section.

Lemma. Elementary collapse Let K be a simplicial complex containing simplex $\sigma = \{v_0, ..., v_n\}$ with facet $\tau = \{v_1, ..., v_n\}$. If σ is the only facet of τ then the inclusion $\iota : K - \{\tau, \sigma\} \rightarrow K$ is a homotopy equivalence. *Proof:* First, shoose the barycenter a of τ and connect it to all vertices of σ. This induces a subdivision of σ, τ, and K, as no other simplex contains σ or τ. To obtain the homotopy, slide a towards $v_0 = \sigma - \tau$. This defines a deformation retraction of σ onto its faces not including τ.

Exercise: Make a sketch of the elementary collapse on a triangle simplex.

Definition. Elementary collapse Let K be a simplicial complex, $\tau \subset \sigma \in K$. Assume that σ is the only coface of τ. A removal $K \rightarrow K - \{\tau, \sigma\}$ is called an elementary collapse.

Proposition. An elementary collapse does not change Betti numbers (Virk)

As above, let K be a simplicial complex, with $\tau \subset \sigma \subset K$, τ be a free face of σ, i.e. σ is the only cofacet of τ, with $\dim(\sigma) = n = \dim K$. We then have that $\beta_i(K) = \beta_i(K - \{\tau, \sigma\})$ for all dimensions.

Proof:

First, note that, since σ is the only cofacet of τ, $\implies \partial_n \sigma \in \operatorname{im} \partial_n$. removing σ from the columns of ∂_n decreases its rank by 1 (i.e. decreases dimim ∂_n). Also, $\sigma \in \operatorname{coim} \partial_n$, a nd we have that $C_n = \operatorname{coim}\partial_n \oplus \operatorname{im} \partial_{n+1} \oplus H_n$. Since τ is a free face, this implies that σ is a maximal simplex, hence$C_{n+1} = 0$, and thus $C_n = \operatorname{coim}\partial_n \oplus H_n$. We thus have that collapsing σ affects coim∂_n, and therefore leaves H_n unaffected. Therefore collapse does not affect β_n. On the other hand, we also have that $\operatorname{im} \partial_n \subset \ker \partial n - 1$. Hence, removing τ from the columns of ∂_{n-1} also decreases dimker ∂_{n-1} by 1. Therefore the Betti number β_{n-1} is unaffected. Finally, by dimensionality of the simplices to remove, all other Betti numbers are not affected. We have thus shown that

$\beta_i(K) = \beta_i(K - \{\tau, \sigma\})$ for all dimensions as desired.

Example. A topological tree is contractible. Define a tree as an acyclic, path-connected simplicial complex of dimension 1. One can easily construct deformation retractions from the leaves of the tree onto its root.

As we mentioned in the introduction homotopy is hard to compute, and hence an attractive alternative is homology. In particular, homology is a homotopy invariant, that is, if two spaces are homotopy equivalent, they have isomorphic homology groups.

3.6 Simplicial homology is a computationally tractable theory for topological investigation

There are different of homology theories that are concerned with classical homology groups. Three of the most important ones are simplicial homology, cellular homology and singular homology.

Each one of them serves a good purpose. Simplicial homology is good for calculations on the computer, but it is cumbersome to prove results with it. On the contrary, singular homology is very good for proving theorems, but it is hard to encode its properties in a computer program. Thus, we will use singular homology to prove the main results of homology as a topological invariant, and will leverage simplicial theory for computation.

Definition. Abstract simplicial complex

An abstract simplicial complex K is a collection of non-empty subsets that is closed under the action of subsetting ,i.e. if $\sigma \in K, \tau \subset \sigma, \tau \neq \emptyset \Rightarrow \tau \in K$.

Perhaps with such an abstract definition it would is helpful to provide a (non-)example.

Example. The set $K = \{\{\emptyset\}, \{v_0\}, \{v_1\}, \{v_2\}, \{v_0, v_1\}, \{v_0, v_1, v_2\}\}$ is not a simplicial complex, because $\tau = \{v_1, v_2\} \subset \{v_0, v_1, v_2\} \notin K$.

Remark. Any subset L of K that also has the properties of an ASC is called a **subcomplex** of K.

Definition. Face / coface Given two simplices $\sigma, \tau \in K$ we say that σ is a **face** of τ, denoted $\sigma \leq \tau \iff \forall v \in \sigma, v \in \tau$. If such relationship holds we also say that τ is a **coface** of σ.

Definition. k-skeleton Let K be a m dimensional abstract simplicial complex, then for any $k < m$ the **k-skeleton** is defined as the set of all simplices of dimension less than or equal to k. That is:

$$K^k = (\emptyset) \cup \{\sigma^0\} \cup ... \cup \{\sigma^k\}, \tag{3.22}$$

where $\{sigma^i\}$ denotes the set of all simplices with dim = i. We thus have that K^k is a subcomplex of K.

The theory of simplicial homology is combinatorial in nature, and we will be using it for computation. There is a corresponding geometrical description of simplicial complexes that we describe below. We will arrive at some results to embed abstract simplicial complexes as subsets of $\mathbb{R}^n$. In this sense, the so-called geometric realizations of abstract simplicial complexes inherit the subspace topology of its ambient space.

We begin by making some useful definitions.

Definition. Convex set A set is convex if one can draw a line between any two points in the set, s.t. all points in the line remain in the set.

Definition. Convex hull For any subset $S \subset \mathbf{R}^n$, its convex hull conv(S) is the smallest set containg S, or equivalently, the intersection of all convex sets containing S.

Definition. Geometric simplex A geometric $p-$simplex is the convex hull of $p + 1$ affinely independent points. .e. a collection of points $\{x_0, x_1, ..., x_p\} : x_1 - x_0, x_2 - x_0, ..., x_p - x_0$ form a linearly independent set.

The **dimension of a $p-$simplex** is p.

A $p-$simplex σ^p can be conceptualized as a generalization of a polyhedron created from $p + 1$ vertices. For instance, a $1-$simplex is a line, a $2-$simplex is a triangle, a $3-$simplex is a tetrahedron, and so forth up to higher dimensional polytopes. Thus a simplex can be specified by a tuple with elements corresponding to vertices. The order in which vertices are specified has a geometric meaning as we will see in the following sections.

Every point in the convex hull of $\{x_0, x_1, ..., x_p\}$ has a unique representation in the form $\sum_i \lambda_i x_i$ where $\lambda_i \geq 0 \, \forall i$ and $\sum_{i=0}^{p} \lambda_i = 1$. This representation is known as the barycentric coordinates of a simplex.

The standard $p-$simplex is $\Delta^p = \{x \in \mathbb{R}^{p+1} : \sum_i x_i = 0 x_i \geq 0 \forall i\}$.

Definition. Geometric simplicial complex

A simplicial complex K is a collection of geometric simplices with the following properties:

1. if $\sigma, \tau \in K \rightarrow \sigma \cap \tau \in K$.
2. if $\sigma \in K, \sigma' \subset \sigma \Rightarrow \sigma' \in K$, i.e. K is closed under the action of subsets.

Definition. Polyhedron Let K be a geometric simplicial complex, then $|K|$ called a polyhedron, is its point-set union (i.e. the set of all points on at least one of the geometric simplices). We also say K is a triangulation of $|K|$.

There are properties that are important to highlight at this point.

Proposition: Topological structure of a geometric simplical complex

Let K be a geometric simplicial complex, then $|K|$ is a topological space under the subspace topology. A basis for a topology is given by the simplices of K. We define each subset L of K as closed if L contains each face of each of its simplices, i.e. L is also a simplicial complex, referred to as a subcomplex of K.

Proof: First equip the simplicial complex K with the empty set, i.e. $K = \{\sigma\} \cup \emptyset$. Since $|K|$ is a compact subset of $\mathbb{R}^n$ it follows that it is a topological space under the subspace topology.

Now showing that the simplices form a basis for the standard subspace topology follows from the definition of basis:

(1) $\forall x \in X \exists B \in \mathcal{B} : x \in B$ is clear. (2) If σ_1, σ_2 share a face, clearly $x \in \tau \subset \sigma_1 \cap \sigma_2 \in K$, else the intersection is empty and the condition also holds.

Now the union of all the simplices (the basis elements) forms a topology, since closed sets also define a topology. ■.

There is also a topological structure defined for abstract simplicial complexes. To show this we need the following Lemma.

Lemma. Intersection and union of abstract subcomplexes of an ASC is an abstract subcomplex

Let L, M be subcomplexes of a simplicial complex K, i.e. $L, M \subset K$ and each is a simplicial complex. Then $L \cap M, L \cup M$ will be subcomplexes of K respectively.

Proof: We first show the case for intersections. Clearly, when viewed as sets, $L \cap M \subset K$. To see this, note that for all $x \in L \cap M \implies x \in L \text{and} x \in M$, and since L, M are subsets, then $x \in K$. Since x was arbitrary, this shows that $L \cap M \subset K$.
Now we need to show that $L \cap M$ is a simplicial complex. Assume for a contradiction that $L \cap M \subset K$, where L, M, K are simplicial complexes and that $\exists \sigma, \tau : \sigma \in L \cap M, \tau \subset \sigma$ and that $\tau \notin L \cap M$. But if $\sigma \in L \cap M \implies \sigma \in L$ and $\sigma \in M$ by definition of intersection. Also, $\forall \tau \subset \sigma \implies \tau \in L (M$ resp. by definition of simplicial complex, a contradiction. Thus, τ must be in both L and $M \forall \tau \subset \sigma \in L \cap M$.

The case for unions is analogous. First let's check containment: if $x \in L \cup M$ then either $x \in L$ or $x \in M$ or both. Since $L, M \leq K$, for any $x \in L$ or $x \in M$ we have that $x \in K$. Therefore $L \cup M \subset K$.

Now, let's verify that $L \cup M$ is a simplicial complex. Let $\sigma \in L \cup M$, and $\tau \subset \sigma$. We need to show that $\tau \in L \cup M$. Note that if $\sigma \in L \cup M$, then either $\sigma \in L$ or $\sigma \in K$ or both. Since L and M are both simplicial complexes, naturally $\tau \subset \sigma \implies \tau \in L$ or $\tau \in M$ or both, and hence $\tau \in L \cup M$. Thus $L \cup M$ is a subcomplex. ■.

Proposition. Topological structure on an abstract simplicial complex

Let K be a finite simplicial complex. Declare a subset $L \subset K$ to be closed if it is a subcomplex of

K. Then all subcomplexes $L \leq K$ generate a topology τ on K. This is the Alexandroff topology on the poset of faces of K.

Proof:

To show that (K, τ) is a topological space it is sufficient to verify the axioms.

First, $\emptyset \in \tau$ and $K \in \tau$ by definition of the topology τ. Since τ is finite by construction, by the Lemma above it will be closed under arbitrary unions and intersections and is thus a topology on K. ■

Remark: As any simplicial complex on a vertex set is a topological space, what is the actual abstract simplicial complex that best explains point cloud? In other words, how can we *infer* the topology for an e.g. subspace of a metric space (X, d)? It turns out that under mild considerations, the Nerve of an open cover (the Cech complex) has the same homology as the underlying space. This result is known as the Nerve Theorem, and this allows us infer the topology (or more precisely the homotopy type) of a space using an ASC built from its point set, which in practice is the Vietoris-Rips complex.

It is easy to obtain an abstract simplicial complex from a geometric one, just replace each coordinate vector with an arbitrary label. Going on the other direction is in general harder. If this operation is achievable, the geometric simplicial complex is said to be a **geometric realization** of the abstract simplicial complex. Let's define it in more formal terms.

Definition: Geometric realization of an abstract simplicial complex Let ϕ be a function that sends vertices of K to points in $\mathbb{R}^n$. The geometric realiation of K w.r.t. to ϕ is the union:

$$|K| = \bigcup_{\sigma \in K} \phi(\sigma), \tag{3.23}$$

where for each $p-$simplex $\sigma = \{v_0, ..., v_p\}$ the set $\phi(\sigma) \subset \mathbb{R}^n$ is the geometric simplex spanned by the points $\{\phi(v_0), ..., \phi(v_p)\}$.

We conclude this section by noting that geometric realizations inherit the subspace topology as subsets of $\mathbb{R}^n$. This result is of remarkable nature since it tells us that we can induce a very nice topology after constructing a simplicial complex. In the following sections we will see how we can construct such abstract simplicial complexes using the Vietoris Rips and Cech algorithms. In this sense we can "topologize" any set coming from measurements from biological systems and compute important topological invariants such as the homology groups.

In order to realize the theory of homology, we have to form an algebraic substrate to work with. In what follows we describe an algebraic model for simplices.

Definition. $p-$**chain**

A $p-$chain is a linear combination of $p-$simplices, where the coefficients are integers. With this construction, chains constitute an additive abelian group over the ring of integers. We will use also addition in $\mathbb{Z}/2\mathbb{Z}$, in which case coefficients just indicate presence or absence. The geometric meaning is that the chain p is a loop over vertices v_0, v_1, v_2 in a counterclockwise direction while the chain $q = \{v_2, v_1, v_0\}$ turns in the clockwise direction. Hence, both simplices and chains have an orientation. This orientation is unique up to even permutations.

Remark: A note on coefficients It turns out that for computational purposes, its better to work with coefficients in a finite field namely $\mathbb{Z}/p\mathbb{Z}$ where p is a prime number. This makes our discussion simpler since model spaces will be vector spaces instead of groups, and hence we can resort to linear algebra to manipulate objects. In particular, using finite fields, all elements have multiplicative inverses, which is a *problem* using integers (since for example we need to invoke the euclidean algorithm when trying to solve the Smith Normal Form). The easiest case (and the one that will be the main use for our discussion) is integers mod 2, since in this case, the interpretation is very neat: everything is orientable and addition boils down to bit flips. To see why Z_2 is a field, it's so trivial that the principle becomes subtle, so let's use the next prime numbers as examples.

Example: Integers modulo 3 is a finite field. Closure under addition is obvious. To see closure under multiplication note that $2 \times 2 = 4\%3 = 1$, and hence 2 is its own multiplicative inverse.

For integers modulo 5 Z_5 we have that for instance $2 \times 2 = 4\%5 = 4$, so 2 must have another multiplicative inverse other than itself. Let's try the next one $2 \times 3 = 6\%5 = 1$. A-ha! We thus have that 2 and 3 are multiplicative inverses of each other. Furthermore 4 is its own multiplicative inverse.

Exercise Show that for any prime p, in Z_p, $p - 1$ will always be its own multiplicative inverse.

Let us continue with our discussion regarding the algebraic model for manifolds: simplicial complexes. Recall that a $p-$chain is a linear combination of $p-$simplices with the appropiate coefficients.

Definition. C_p The abelian group generated by all $p-$chains is denoted by C_p. In other words $C_p =< c_\alpha^p >$.

The next step in building homology theory is to define boundary maps which specify how to connect $p-$chains to $(p - 1)$ and $(p + 1)$ chains, i.e. they provide the instructions on how to construct a simplicial complex by gluing its building blocks.

Definition. Boundary map

The boundary map $\partial_p : C_p \to C_{p-1}$ is a group homomorphism defined by:

$$\partial_p(\sigma) = \sum_{i=0}^{p} (v_0, v_1, .., v_{i-1}, \hat{v_i}, v_{i+1}, ..., v_n), \tag{3.24}$$

where $\hat{v_i}$ means that the $i-$th vertex is deleted from the tuple.

Additionally, define $\partial_0 : C_0 \to 0$ to be the zero map.

A very important property of homology is explained in the following proposition.

Theorem. $\partial_{n-1} \circ \partial_n = 0$

Proof: It suffices to show it on an arbitrary $n-$simplex $\sigma = [v_0, v_1, ..., v_n]$.

$$\partial_{n-1} \circ \partial_n(\sigma) = \partial_{n-1}\left(\sum_{i=0}^{n}(-1)^i[v_0, .., \hat{v_i}, .., v_n]\right) = \tag{3.25}$$

$$= \sum_{i=0}^{n}(-1)^i\left(\sum_{j=0}^{i-1}(-1)^j[v_0, .., \hat{v_j}, \hat{v_i}, .., v_n] + \sum_{j=i+1}^{n}(-1)^{j-1}[v_0, .., \hat{v_i}, .., \hat{v_j}, .., v_n]\right) = \tag{3.26}$$

$$= \sum_{i>j}(-1)^{i+j}[v_0, .., \hat{v_j}, ..\hat{v_i}, .., v_n] + \sum_{j>i}(-1)^{i+j-1}[v_0, .., v_i, .., v_j, .., v_n] = \tag{3.27}$$

$$= \sum_{i>j}(-1)^{i+j}[v_0, .., \hat{v_j}, ..\hat{v_i}, .., v_n] - \sum_{j>i}(-1)^{i+j}[v_0, .., \hat{v_i}, .., \hat{v_j}, .., v_n] = 0, \tag{3.28}$$

where in the third equality we used the definition of group homomorphism and separated the sum for ∂_{n-1} for indices smaller than i and larger than i respectively. The exponent for the cases where $j > i$ is $(-1)^{j-1}$ since we skipped the ith and still label the $n-1$ simplex with the initial $n+1$ vertices. Doing a simple case by hand illuminates this step. The fifth equality is given by noting that $(-1)^{i+j-1} = -(-1)^{i+j}$.

Corollary. $\text{im}\partial_{n+1} \subset \text{ker}\partial_n$

This is the most important result in the theory of homology. In essence, the definition of homology groups rests theoretically on the above corollary. In fact, not only simplicial homology, but all homology theories will have an equality of the same form. Furthermore, a chain complex can be abstractly defined using this property. In the following definition, what may be the most concrete form of a chain complex.

Definition. Chain complex The collection of chain groups and corresponding boundary maps is called a chain complex, denoted :

$$... \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} C_{n-2}... \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0. \tag{3.29}$$

Definition. (Simplicial) Homology Group The $n-$th quotient group is defined by:

$$H_n = \ker\partial_n / \text{im}\partial_{n+1} \tag{3.30}$$

it is informally the set of equivalence classes of homologous holes of dimension n. For example, the H_0 is generated by connected components, H_1 is generated by loops, H_2 is generated by the cavities, and so forth.

Since we are taking the quotient w.r.t. to boundaries (im∂) we have that:

$$x \sim 0 \iff x \in \text{im}\partial \iff \exists\gamma : \partial(\gamma) = x. \tag{3.31}$$

More generally, two chains are homologous if:

$$z \sim y \iff z - y \in \text{im}\,\partial_{n+1} \iff \partial(z) = \partial(y). \tag{3.32}$$

To be more explicit, we have that $\exists\gamma \in C_{n+1} : \partial_{n+1}(\gamma) = z - y \implies \partial\partial(\gamma) = \partial(z - y) = 0 \implies \partial(z - y) = 0 \implies \partial(y) = \partial(z)$.

Note that in the above expression z, y need not be in the kernel, however, it does imply that $z - y$ is in the kernel, i.e. it is a cycle.

Non-trivial equivalence classes are of the form:

$$[v] := [x \in \ker\partial_n : x \sim v] \tag{3.33}$$

$$= \{x \in \ker\partial_n : x - v \in \text{im}\,\partial_n\} \tag{3.34}$$

$$= \{x \in \ker : x = v + \partial_{n+1}\gamma, \gamma \in C_{n+1}\}. \tag{3.35}$$

Definition. Betti number The rank of the $n-$th quotient group is the $n-$th Betti number.

$$B_n = \dim(H_n) = \dim \ker\partial_n - \dim \mathrm{im}\partial_{n+1} \tag{3.36}$$

Now, we continue to show that homology is an homotopic invariant.

Definition. Induced homomorphism $f_\#$

Let $f : X \to Y$ be a continous function between top. spaces. Its induced homomorphism on chains $f_\# : C_n(X) \to C_n(Y)$ is defined by composing it with each singular chain to get:

$$f \circ \sigma : \Delta^n \to Y \tag{3.37}$$

$$f_\# : \left(\sum n_i\sigma_i\right) \mapsto \sum n_i f(\sigma_i). \tag{3.38}$$

It turns out that by checking commutativity with the boundary map, $f_\#$ can be viewed as a chain map between chain complexes.

Proposition : $f \circ \partial = \partial \circ f$

Proof: We can proceed by direct computation on an arbitrary singular simplex:

$$f_\# \circ \partial(\sigma) = f\left(\sum_{i=0}^{n}(-1)^i\sigma|_{[v_0,\ldots,\hat{v_i},\ldots,v_n]}\right) \tag{3.39}$$

$$= \sum_{i=0}^{n}(-1)^i f \circ \sigma|_{[v_0,\ldots,\hat{v_i},\ldots,v_n]} \tag{3.40}$$

$$= \partial \circ f_\#(\sigma). \tag{3.41}$$

Lemma. $f_\#$ takes cycles to cycles

Proof: Let $x \in \ker\partial_n^x \implies \partial_n^x(x) = 0 \implies f_\#\partial_n^x(x) = 0 = \partial_n^y f_\#(x) \implies f_\#(x) \in \ker\partial_n^y$ ■

Lemma. $f_\#$ **takes boundaries to boundaries**

Proof: Let $y \in \text{im}\partial_n^x \implies \exists x \in C_{n+1}(X) : \partial(x) = y \implies f_\#\partial_n^x(x) = f_\#(y) = \partial_n^y f_\#(x) \implies f_\#(x) \in \text{im}\partial_n^y$ ■

Corollary. Pushforward on homology f_*

Let $f : X \to Y$, and $f_\#$ be its corresponding induced homomorphism. Then $f_\#$ induces a homomorphism between homology groups called the *pushforward* on homology, defined as:

$$f_* : H_n(X) \to H_n(Y) \tag{3.42}$$

$$f_* : [\gamma] \mapsto [f_\#(\gamma)]. \tag{3.43}$$

Remark: The well-definedness of the map, that is, independence of the choice of representative, follows from the two Lemmas above, but it can easily be shown for completeness :
Consider $[\alpha + \partial\gamma] \in H_n(X)$, then

$$f_*([\alpha + \partial\gamma]) = [f_\#(\alpha + \partial\gamma)] \text{ by definition} \tag{3.44}$$

$$= [f_\#(\alpha) + f_\#(\partial\gamma)] \text{ since } f_\# \text{ is a hom.} \tag{3.45}$$

$$= [f_\#(\alpha) + \partial f_\#(\gamma)] \text{since } f_\# \text{ commutes with } \partial \tag{3.46}$$

$$\sim [f_\#(\alpha)] \tag{3.47}$$

$$= f_*[\alpha]. \tag{3.48}$$

Homology can therefore be conceptualized as a **functor** which maps the category of simplicial complexes to the category of abelian groups, where chain maps are converted into pushforward maps between homology groups. More precisely, homology is a *covariant* functor. As we will see in persistent homology, it is worthwile to think about homology in these terms, i.e. **homology is a functor**.

Prop. Important properties of the pushforward f_*.

- $(f \circ g)_* = f_* \circ g_*$. This follows from the associativity of compositions $\Delta_n \to^\sigma X \to^g Y \to^f Z$.

- $(\mathrm{id})_* = \mathrm{id}_G$ where id_G denotes the identity in the group.

Theorem If two maps are homotopic, then they induce the same homomorphism across dimensions, i.e.:

$$f \simeq g \implies f_* = g_* : H_n(X) \to H_n(Y)$$

Theorem. If two spaces X, Y have the same homotopy type, then they have isomorphic homology groups.

Proof. $X \simeq Y$ implies that $\exists f, g : g \circ f \simeq \mathrm{id}_X, f \circ g \simeq \mathrm{id}_Y$. By the theorem above $(g \circ f)_* = \mathrm{id}_X = g_* \circ f_*, (f \circ g)_* = \mathrm{id}_Y = f_* \circ g_*$. Therefore $f_*^{-1} = g_*$, i.e. f_* is a group isomorphism with two-sided inverse g_*. Thus, $H_n(X) \simeq H_n(Y)$ for all n.

Corollary. If two spaces X, Y are homotopy equivalent, then they have the same Betti numbers.

Proof: This follows from the theorem above and the fact that isomorphic groups have the same dimension.

Algorithm to compute homology

Theorem. Fundamental decomposition of chain groups

Over Z_2 the chain group C_n vector space can be decomposed as the following direct sum:

$$C_n = \text{im}\,\partial_{n+1} \oplus H_n \oplus \text{coim}\,\partial_n. \tag{3.49}$$

Proof:

First, note that

$$C_n = \ker \partial_n \oplus \text{coim}\,\partial_n \tag{3.50}$$

$$= \text{coker}\,\partial_{n+1} \oplus \text{im}\,\partial_{n+1} \tag{3.51}$$

furthermore, since $\ker \partial_n \subset C_n$ we have that:

$$\ker \partial_n = \ker \partial_n \cap C_n \tag{3.52}$$

$$= \ker \partial_n \cap (\text{coker}\,\partial_{n+1} \oplus \text{im}\,\partial_{n+1}) \tag{3.53}$$

$$= \ker \partial_n \cap \text{coker}\,\partial_{n+1} \oplus \ker \partial_n \cap \text{im}\,\partial_{n+1} \tag{3.54}$$

$$= \ker \partial_n \cap \text{coker}\,\partial_{n+1} \oplus \text{im}\,\partial_{n+1} \tag{3.55}$$

$$= H_n \oplus \text{im}\,\partial_{n+1} \tag{3.56}$$

$$\implies H_n = \ker \partial_n \cap \text{coker}\,\partial_{n+1}, \tag{3.57}$$

where the fourth line follows from $\text{im}\partial_{n+1} \subset \ker\partial_n$, and the second to last line follows by dimension counting $\beta_n = \dim H_n = \dim \ker \partial_n - \dim \text{im}\,\partial_{n+1}$.

Putting these results together we get that:

$$C_n = \ker \partial_n \cap \operatorname{coker} \partial_{n+1} \oplus \operatorname{im} \partial_{n+1} \oplus \operatorname{coim} \partial_n \tag{3.58}$$

$$= \operatorname{im} \partial_{n+1} \oplus H_n \oplus \operatorname{coim} \partial_n. \tag{3.59}$$

Remark: The above result is also known as the Hodge decomposition when working on differential forms.

$$C_n = \left[\begin{array}{c|c|c} \ker \partial_n & H_n & \operatorname{coim} \partial_n \end{array}\right]$$

$$= \left[\begin{array}{c|c|c} \operatorname{im} \partial_{n+1} & & \operatorname{coker} \partial_{n+1} \end{array}\right]$$

Figure 3.1: Fundamental decomposition of chain groups.

At this point we have the following picture: for an abuse of notation let C_n denote the matrix for a basis of C_n. The above theorem tells us that if there is non-trivial n-homology, there exists a basis of C_n so that the first $\operatorname{rk} \partial_{n+1}$ columns belong to $\operatorname{im} \partial_{n+1}$, the next β_n columns belong to the n-homology group H_n, and the last $\operatorname{rk} \partial_n$ columns belong to $\operatorname{coim} \partial_n$ (Figure 3.1). To see why dimensions match, its helpful to have the following definition in mind.

Definition. The dualization of the simplicial chain complex generates simplicial cohomology. Namely, we get cochain groups $C^n = \operatorname{Hom}(C_n, \mathbb{Z}_2)$ connected by coboundary maps (or exterior derivatives) $d_n : C^n \to C^{n+1}$ The cohomology groups are defined as:

$$H^n = \ker d_n / \text{im} d_{n-1} = \text{coker}\, \partial_{n+1} / \text{coim} \partial_n \tag{3.60}$$

with the property that $dd = 0$, i.e. $\text{coim}\partial_n \subset \text{coker}\, \partial_{n+1}$. To see why let $f \in C^n, \sigma \in C_n$, then $ddf(\sigma) = d(df(\sigma)) = d(f \circ \partial\sigma) = f \circ \partial\partial\sigma = 0$ by definition of the dual map.

Let's recap our set up. Since our discussion focuses on homology using $\mathbb{Z}_2$ as coefficients, we have that $C_n \simeq C^n$, since we can thing of cochains as indicators of the presence or absence of chains. Furthermore, we have that $\partial_n = d^{n-1}$, that's why we were able to say that $H^n = \text{coker}\, \partial_{n+1} / \text{coim} \partial_n$ in the definition of the cohomology groups.

To get a full understanding of the Fundamental decomposition of chain groups result, it's necessary to fully understand each of the fundamental subspaces. Hopefully at this point you have a good intuition of what the groups $\text{im}\, \partial_{n+1}$ and $\ker \partial_n$ look like geometrically. If not, let's reiterate: $\text{im}\, \partial_{n+1}$ generates n chains that are boundaries of $n+1$ chains. To put an example, $\text{im}\, \partial_2$ is the set of all edges in a simplicial complex that can be generated by taking the boundaries of triangles. Clearly, all such edges will form loops, which is an example of how $\text{im}\, \partial_{n+1} \leq \ker \partial_n$.

At this point one may wonder, what is the geometrical meaning of coker∂ and coim∂. It turns out that in the case of surfaces, the geometric interpretation of these subspaces is straightforward.

Recall that we showed that:

$$C_n = \text{im}\, \partial_{n+1} \oplus \text{coim}\, \partial_n \oplus H_n \tag{3.61}$$

$$= \text{im}\, \partial_{n+1} \oplus \text{coker}\, \partial_{n+1} \tag{3.62}$$

$$\implies \text{coker}\, \partial_{n+1} = \text{coim} \partial_n \oplus H_n. \tag{3.63}$$

With this in mind, to understand the $\text{coker}\, \partial_{n+1}$, first, we need to understand the coimage.

Example: Let's consider a hollow tetrahedron. If we order the simplices lexicographically, we have that $C_2 = \langle [012], [013], [023], [123] \rangle$, and $C_1 = \langle [01], [02], [03], [12], [13], [23], [24] \rangle$. By hypothesis, coim$\partial_1$ will be the acyclic portion of C_1, since its complement is $H_1 \oplus \text{im}\, \partial_2$.

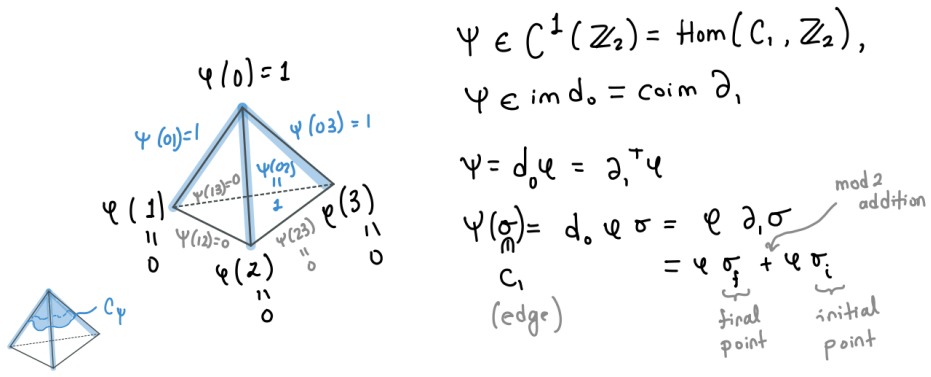

Figure 3.2: Visualization of the coimage of the ∂_1 boundary map on a tetrahedron.

Using the lexicographic order, we can write down ∂_1^T as:

$$\partial_1^T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{3.64}$$

Hence, we can take an element on im∂_1^T = coim∂_1, e.g. $\partial_1([1]) = [01] + [12] + [13]$. In other words, the action of ∂_1^T generates all edges that are cofaces of a given vertex. In this case, $\partial_1([1])$ can be visualized as a "tree" living on the surface of the tet. We can confirm the algebraic relationship coim$\partial_1 \subset$ coker∂_2 by writing down ∂_2 (again, according to the lexicographic order stated above):

$$\partial_2^T = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{3.65}$$

applying $\partial_2^T \circ \partial_1^T([1]) = 0$, as the reader can confirm. Thus $[01] + [12] + [13] \in \operatorname{coker} \partial_2$. Therefore, we can think of $\operatorname{coker} \partial_n$ as the n chains generated by the "acyclic" part of C_n and the homology chains.

Next, we show how does a cohomology class looks geometrically.

There is a really nice interpretation for 1-cochains using Z_2 coefficients on surfaces: each 1-cochain ψ can be associated with a collection of curves C_ψ tghat cross each edge transversally, such that the number of intersections with each edge equals the value of the cochain on that edge. What's nice is that from this geometric view $\psi = d_0\phi$ means that the curve divides the manifold X into two disjoint regions X_0 and X_1, and the subscript indicates the value in the region !

If there is no solution to $\psi = d\phi$ one cannot construct such curves.

Example. 1-cohomology class on the annulus.

We can construct a cohomology class by following its definition: finding a 1-cochain ψ that will be in the cokernel of $\partial_2 \iff d_1(\psi) = 0$. This will occur if the number of times each ψ takes on the value of 1 on the boundary of each 2 simplex is either 0 or 2, since $d_1(\psi)(\sigma) = \psi\partial_2(\sigma) = 0 \forall \sigma \in C_2$. Here, we're taking into account that we're working with a 2-manifold each edge is the face of two triangles exclusively, and that we're using $\mathbb{Z}_2$ as coefficients.

The curve C_ψ in the figure above crosses each edge exactly 0 or 2 times and thus the associated 1-cochain $\psi \in \ker d_1$. However, as you can verify, there is no 0-chain that solves $d\phi = \psi$, i.e. $\psi \notin \operatorname{im} d_0 \implies C_\psi$ is associated with a $1-$cohomology class. Also note that the curve does not separate the domain into two regions.

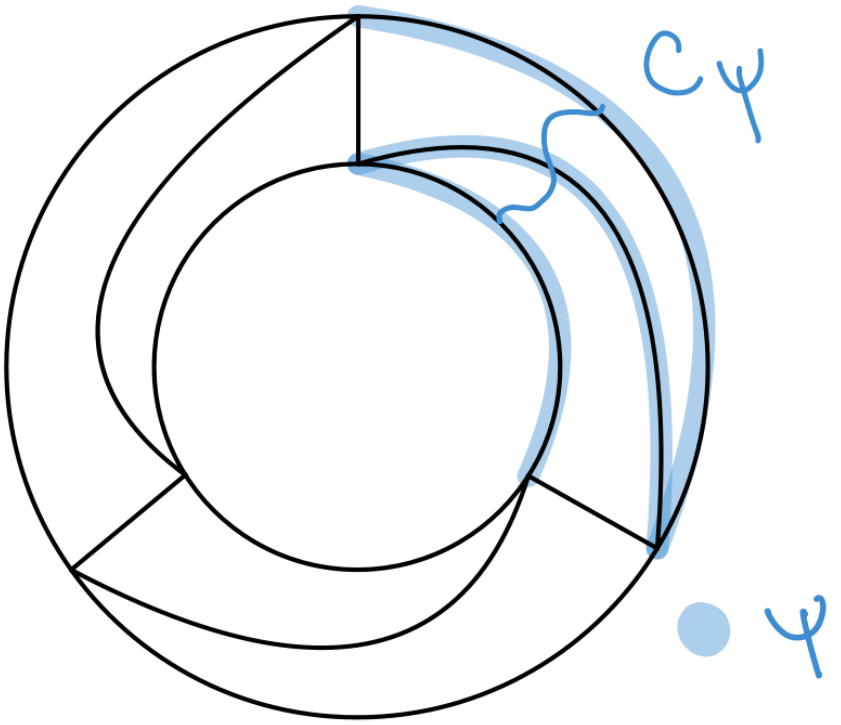

Figure 3.3: Cohomology class on the annulus.

By identifying the inner and outer loops in the annulus one gets a similar construction in the torus.

Example. 1-cohomology class on the torus.

Consider the standard simplicial triangulation of the torus in Fig x.

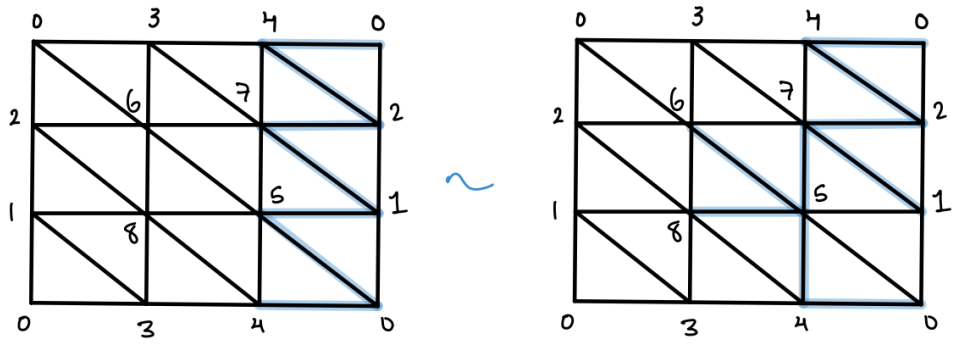

Figure 3.4: Cohomology class on the torus.

Claim: The 1-chain / cochain $\psi = [0,4] + [0,5] + [1,5] + [1,7] + [2,4] + [2,7]$ is both in the kernel of ∂_1 and in the cokernel of ∂_2, hence it will be both a homology and cohomology class. We can confirm that by computing directly on restricted versions of the boundary / coboundary matrices. More explicitly, the restricted matrices are, for d_1 just considering the image of the $1-$cochain, and for ∂_1 as follows:

$$\partial_2^T = d_1 = \begin{array}{c} \\ [0,1,5] \\ [0,2,4] \\ [0,4,5] \\ [1,2,7] \\ [1,5,7] \\ [2,4,7] \end{array} \begin{array}{c} \begin{array}{cccccc} [0,4] & [0,5] & [1,5] & [1,7] & [2,4] & [2,7] \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \tag{3.66}$$

$$\partial_1 = \begin{array}{c} \\ [0]1 \\ [1]0 \\ [2]0 \\ [4]1 \\ [5]0 \\ [7]0 \end{array} \begin{array}{c} \begin{array}{cccccc} [0,4] & [0,5] & [1,5] & [1,7] & [2,4] & [2,7] \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 1 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 1 & \\ 0 & 0 & 0 & 1 & 0 & \\ 1 & 1 & 0 & 0 & 0 & \\ 0 & 0 & 1 & 0 & 1 & \end{bmatrix} \end{array} \tag{3.67}$$

We encourage the interested reader to confirm that both matrices are correct, and that are the only thing we need for our purposes. One way to convince yourself is to see that each column in the matrices are above have two nonzero entries: this is because in each edge is the face of only two triangles in a surface, and it is the coface of two vertices. Furthermore, one can confirm that $d_1\psi = 0$ and $\partial_1\psi = 0$. In Fig x. we visualize the corresponding chain. We can easily find a

cohomologous class ψ' by adding a 0-coboundary $d_0[5] = \partial_1^T[5]$. *Exercise.* Show that ψ' is not a homology class.

Definition. Smith Normal Form (SNF)

Let A be an $m \times n$ integer matrix, i.e. $A \in M_{m\times n}(\mathbf{Z})$. There is a unique factorization $A = UDV^{-1}$ with the following properties:

- $D \in M_{m\times n}(\mathbb{Z})$ is a diagonal matrix, with the property that $D_{ii}|D_{i+1,i+1}$. The diagonal entries are called elementrary dividors or invariant factors of A.
- $U \in GL(m,\mathbb{Z}), V \in GL(n,\mathbb{Z})$, i.e. U and V are unimodular (with det $= \pm 1$), invertible, integer matrices.

We also call D the SNF of A. The main result about the SNF is that is defined for every integer matrix—so it is in a sense a univeral property of integer matrices. It is unique up to signs. Naturally U and V are unique up to isomorphism. The SNF is important in the context of computational homology since one gets all of the fundamental subspaces of a matrix after the decomposition.

Lemma. SNF provides all fundamental subspaces of an integer matrix

Let $A = UDV^{-1}$, with $A \in M_{m\times n}(\mathbb{Z})$. Then U constitutes a basis for $\mathbb{Z}^m$, and V is a basis for $\mathbb{Z}^n$. Furthermore, if we let r equal to the nonzero diagonal elements of D, then the first r columns of U are a basis for im$A$, and the last $n - r$ columns of V are a basis for ker$A$.

Proof:

The first statment is easy to see as U, V are isomorphisms and they have each m, n columns so they are indeed a basis for $\mathbb{Z}^m, \mathbb{Z}^n$ respectively. Now note that if $i \in 1, ..., r$ we have

$$Av_i = d_i u_i \neq 0 \Rightarrow u_i \in \text{im}A. \tag{3.68}$$

Furthermore, if $r < j < n$ then

$$Av_j = 0u_j = 0 \Rightarrow v_j \in \text{ker}A. \tag{3.69}$$

■.

With the above result, we can now easily calculate the Betti numbers. However, should we be interested in the explicit calculation of representatives of homology groups, we are still left astray: despite we are able to compute $\ker\partial_j$ and $\mathrm{im}\partial_{j+1}$, their bases need not be equal. This motivates the following results.

Computing generators using the Smith Normal Form

Definition. D^* Let D be the SNF of a matrix A. Then define D^* to be the matrix which is the result of permuting the columns of the matrix D so that the diagonal block is in the right upper corner.

In this sense, we're thinking of D_i^* to be the result of applying a change of basis to the boundary matrix ∂_i to get $D_i^* = U_i^{-1}\partial_i V_i P_i$.

Accordingly, since operations on columns on ∂_i, as a change of basis operation, correspond to operations on rows on ∂_{i+1} we have the following definition.

Definition. Let $D_i = U_i^{-1}\partial_i V_i P_i$, we say that $\tilde{\partial}_{i+1}$ is the matrix into which ∂_{i+1} is carried after applying the operations to diagonalize ∂_i using SNF. That is $\tilde{\partial}_{i+1} = P_i^{-1}V_i^{-1}\partial_{i+1}$.

In a more succint description, let $\tilde{V}_i = V_i P_i$, then we define:

$$D_i^* = U_i^{-1}\partial_i\tilde{V}_i \tag{3.70}$$

$$\tilde{\partial}_{i+1} = \tilde{V}_i^{-1}\partial_{i+1} \tag{3.71}$$

Lemma. The last r_i rows of $\tilde{\partial}_{i+1}$ consist of zeros.

Proof:

For a contradiction, assume that the last r_i rows of $\tilde{\partial}_{i+1}$ are not all zeros. However, since smith factor matrices and permutation matrices are isomorphisms we have that $D_i^* \circ \tilde{\partial}_{j+1} = 0$. This leads to a contradiction of our assumption. ■

The importance of this result is that by changing bases on a boundary matrix i to those specified by the previous matrix following SNF decomposition, we actually advance the SNF of the $i-$th matrix and work with the same basis. We thus come closer to our goal of having the same basis to directly compute a representative of the homology group. The following result achieves this goal.

Theorem. It is possible to choose bases for the chain groups $C_0, C_1, ..., C_n$ in terms of which the boundary maps have the form D^* and the consecutive spaces ($\ker\partial_i, \text{im}\partial_{i+1}$) have the same basis.

Proof:

Let $\partial_i, \partial_{i+1}$ be consequent boundary maps of a chain complex.

Further, assume that the SNF of ∂_i is available in the form $D_i^* = U_i^{-1}\partial_i\tilde{V}_i$.

Change basis to map $\partial_{i+1} \mapsto \tilde{\partial}_{i+1}$, and decompose the matrix using SNF to get $D_{(i+1)'} = U_{i+1}^{-1'}\tilde{\partial}_{i+1}V_{(i+1)'} = (U_{i+1}^{-1'}\tilde{V}_i^{-1})\partial_{i+1}V_{(i+1)'}$.

We now have that $\mathfrak{B}_\mathfrak{i} = \tilde{V}_i U_{(i+1)'}$ is a common basis for the $\ker\partial_i$ and the $\text{im}\partial_{i+1}$. ■

Corollary: The columns with index $\{\dim\text{im}\partial_{i+1}+1, \dim\text{im}\partial_{i+1}+2, ..., \dim\ker\partial_i\}$ of $\mathfrak{B}_\mathfrak{i}$ constitute representatives of the $i-$th homology group.

Proof:

B_i is a common basis for $\ker\partial_i$ and $\text{im}\,\partial_{i+1}$ by the last result. Since $\tilde{\partial}_{i+1}$ has the last r_i rows with all zeros, $U_{(i+1)'}^{-1}$ leaves the last r_i rows of $\tilde{\partial}_{i+1}$ unchanged. Thus the first r_i rows $\mathfrak{B}_\mathfrak{i}$ constitute a basis for $\ker\partial_i$, as are the first columns of $\tilde{V}_i$. However, because $D_{(i+1)'}$ is the SNF of $\tilde{\partial}_{i+1}$, the first r_{i+1} columns of $\mathfrak{B}_\mathfrak{i}$ are also a basis for the image of ∂_{i+1}, where $r_{i+1} = \dim\text{im}\partial_{i+1}$. Therefore columns corresponding to the indices $\{\dim\text{im}\partial_{i+1} + 1, \dim\text{im}\partial_{i+1} + 2, ..., \beta_i\}$ of $\mathfrak{B}_i$ are in the kernel of ∂_i and not in the image of ∂_{i+1}, and there are β_i of them. ■

Remark: The above result is an algorithm to compute the homology groups of a simplicial complex.

3.7 Systematic topological inference with Persistent Homology

Up to now we have dealt with both the representation problem (addressed with a simplicial complex structure) and the computation problem (addressed with the Smith Normal Form). However, there is one detail we haven't addressed yet: in the definition of a Vietoris-Rips complex, we need a threshold ε which setys the locality of the neighborhood. But how do we choose this threshold? The answer is that we don't choose one, but rather sweep through a range of values and record the topological structures that emerge and cease at different scales. To make an analogy, instead of looking at a snapshot of the topology, we instead watch a movie (this will be the filtration) of topological structures, and record the different events in a persistence diagram. It turns out that the theory behind this approach—called persistent homology—is surprisingly beautiful, and simple to compute. But as before, the key is to approach the problem step-by-step.

Def: Filtration. A nested sequence of subspaces of the form:

$$\mathcal{K} = \{\emptyset \subset K_0 \subset K_1 \subset \ldots \subset K_n\}, \tag{3.72}$$

is called a filtration of the space X.

To understand what these inclusion maps do at the level of homology groups, it is useful to note that homology is a functor mapping the category of topological spaces to the category of abelian groups, where continuous maps get sent to corresponding group homomorphisms.

Applying the homology functor to the filtration $\mathcal{K}$ yields a sequence of homology groups and group homomorphisms, the **persistence module** (or quiver, see next section) associated to the filtration:

$$H_n(K_0) \xrightarrow{\iota_*^{0,1}} H_n(K_1) \xrightarrow{\iota_*^{1,2}} \ldots \xrightarrow{\iota_*^{n-1,n}} K_n \tag{3.73}$$

Where the asterisk denotes that the function is now a linear map between the homology groups (which recall are vector spaces using Z_2 coefficients, or more generally homomorphism between

homology groups). The ι_* are also called the *pushforwards* of inclusion maps. That is if $\gamma \in H_n(K_i)$, $\iota_*^{i,j}[\gamma] = [\iota(\gamma)] \in H_n(K_j)$, by definition of the induced maps on homology.

Def: n-th Persistent homology group and persistent Betti numbers

The images of the pushforward inclusion maps are the **persistent homology groups**, that is:

$$H_n^{i,j} = \text{im}\, \iota_{n,*}^{i,j}. \tag{3.74}$$

The corresponding $n-$th **persistent Betti number** are their corresponding ranks:

$$\beta_n^{i,j} = \dim H_n^{i,j}. \tag{3.75}$$

In words, persistent homology groups consist of the homology classes of K_i still alive at K_j.

To reiterate, for any $i \leq j$ we have a linear map $\iota_*^{i,j} : H_n(K_i) \to H_n(K_j)$. If $j - i > 1$ we have $\iota_*^{i,j} = \iota_*^{j-1,j} \circ ... \circ \iota_*^{i+1,i+2} \circ \iota_*^{i,i+1}$ by functoriality. The image of this linear map are the non-trivial homology classes that persist from K_i to K_j. However, such classes may not actually be born in K_i.

The **n-homology classes born exactly at** K_i are $H_n(K_i)/\,\text{im}\, \iota_{n,*}^{i-1,i}$, and there are exactly $\beta_n(K_i) - \beta_n^{i-1,i}$ of them.

The **n-homology classes that die exactly at** K_j are precisely $\ker \iota_{n,*}^{j-1,j}$. Furthermore there are exactly $\beta_n(K_{j-1}) - \beta_n^{j-1,j}$ of them.

Let's recap what we just said with specific representatives. Let $\gamma \in H_n(K_i)$, we say that this $n-$homology class is *born* at K_i if $\gamma \notin \text{im}\iota_{n,*}^{i-1,i} = H_n^{i-1,i}$.

Futhermore a homology class *dies* at K_j if it merges with a previously born homology class exactly at K_j, i.e. $\iota_{n,*}^{i,j}([\gamma]) \in \text{im}\, \iota_{n,*}^{i-1,j} = H_n^{i-1,j}$ and $\iota_{n,*}^{i,j-1}([\gamma]) \notin \text{im}\iota_{n,*}^{i-1,j-1} = H_n^{i-1,j-1}$. This leads to the following definition.

Definition. Persistence

If a homology class $[\gamma]$ is born at K_i and dies at K_j, then, its persistence is the difference in their function values:

$$\text{pers}(\gamma) = a_j - a_i.$$

If the indices are sufficient in the context one can also define the persistence as $\text{pers}(\gamma) = j - i$.

Remark: An alternative definition of the persistence homology groups, given in the original paper (See Edelsbrunner, Letscher, and Zomorodian), is the following: recall that $H_n^{i,j}$ are the homology classes present at K_i that do not become trivial in K_j.

Then, $H_n^{i,j}$ is approximately $\ker \partial_n(K_i)/\text{im}\, \partial_{n+1}(K_j)$. However, we don't know if $\text{im}\, \partial_{n+1}(K_j)$ is a subspace of $\ker \partial_n(K_i)$, since C_{n+1}^j may be substantially larger than C_{n+1}^i. Thus, to ensure proper containment one actually has that:

$$H_n^{i,j} = \ker \partial_n(K_i)/(\text{im}\, \partial_{n+1}(K_j) \cap \ker \partial_n(K_i))$$

We know that $(\text{im}\, \partial_{n+1}(K_j) \cap \ker \partial_n(K_i))$ is a subspace of C_n^j, since the intersection of two subspace is a subspace.

Let's ground the above definitions with an example.

Example: Visualizing persistent homology groups

Consider the filtration in Figure 3.5 A. This filtration consists of the sublevel sets $K_a = f^{-1}(-\infty, a)$ of the height function f applied to a curve C. The filtration is $\mathcal{X} = \{\emptyset \subset K_0 = f^{-1}(-\infty, h_0) \subset K_1 = f^{-1}(-\infty, h_1) \subset \ldots \subset K_4 = f^{-1}(-\infty, h_4)\}$ and the corresponding $0-$th homology groups (using mod2 coefficients) are :

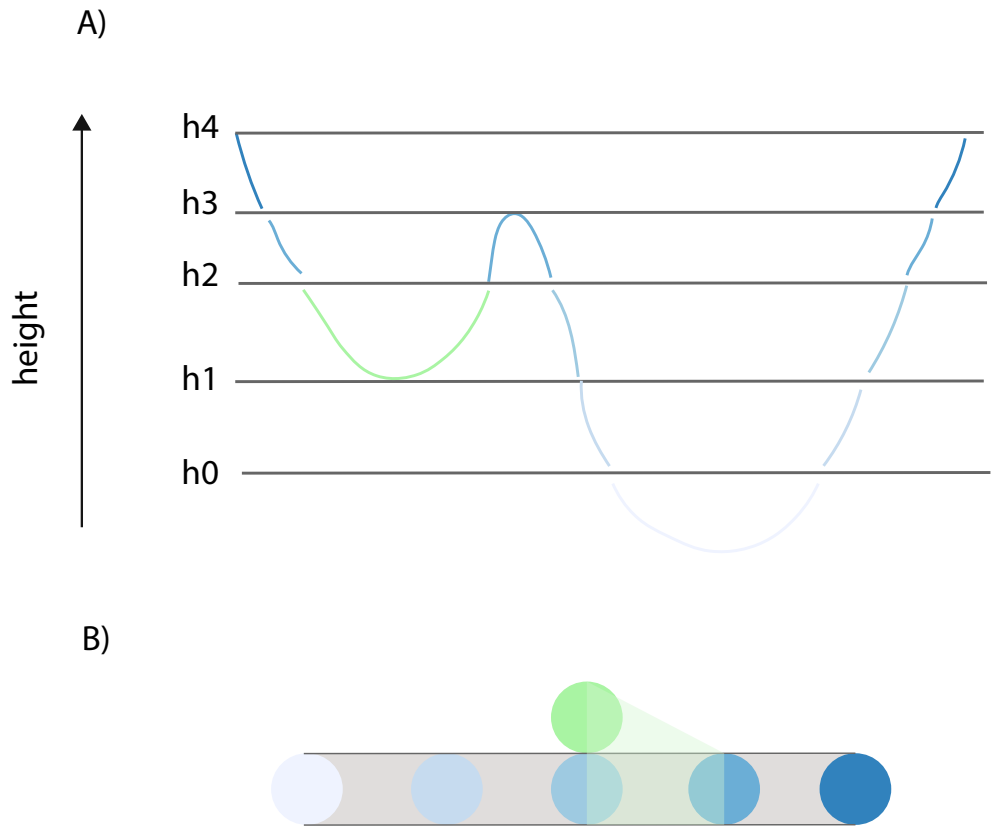

Figure 3.5: Persistent homology of a filtration using the preimages of a Morse function.

$$H_0(K_0) \approx \mathbb{Z}_2 \xrightarrow{\text{id}} H_0(K_1) \approx \mathbb{Z}_2 \xrightarrow{\begin{bmatrix}1\\0\end{bmatrix}} H_0(K_2) \approx \mathbb{Z}_2^2 \xrightarrow{\begin{bmatrix}1 & 1\end{bmatrix}} H_0(K_3) \approx \mathbb{Z}_2 \xrightarrow{\text{id}} \mathbb{Z}_2 \approx H_0(K_4) \tag{3.76}$$

There are several interesting things to note in this example. First, note that the rank of the groups are correct since for all sublevel sets we have a Betti number equal to 1, except on K_2. Also, note that we can explicitly write down the induced linear maps on homology groups. To see why the linear maps are correct, the only detail is to realize that we need a change of basis for $H_0(K_2)$, especifically we need to map $\begin{bmatrix}1\\0\end{bmatrix} \mapsto 1 \in H_0(K_3)$, $\begin{bmatrix}1\\1\end{bmatrix} \mapsto 0 \in H_0(K_3)$ via the map $\begin{bmatrix}1 & 1\end{bmatrix}$. Please note that the operation is in modulo 2 arithmetic.

A visualization of the homology groups is displayed in Figure 3.5 B. In the figure we have visualized in grey the sequence of isomorphisms from the first generator. Note that this generator has infinite lifetime, that is, it is born at the start of the filtration and never dies. On the other hand the generator corresponding to the second component (visualized in green) is born at the third value of the filtration, and dies in the fourth one. Thus, **the green generator has a lifetime of one**.

Remark: Despite that the inclusion maps are always injective, their induced homomorphisms in homology groups are not always injective, and can be surjective : when a persistent homology class dies, the rank of the group goes down by one and thus cannot be injective. In our example, note that the map from K_2 to K_3 is surjective despite that the inclusion map is injective. To reiterate: the induced maps on homology groups need not be of the same class as the maps on simplicial chains.

Algorithm

Assume we have a filtration $\mathcal{K} = K_1 \subset ... \subset K_N$, where in each step we add a simplex at a time. (We may need to expand a bit here recalling where does the filtration comes from, why do we add one simplex at a time, and that this process is called refinement, and point to the corresponding reference). This will induce a total ordering of the simplices in $mathcalK$. Also assume that we're using $\mathbb{Z}_2$ as coefficients. It turns out that under this set up we can extract the persistence module associated with $\mathcal{K}$ using a variation of Gaussian elimination. Towards this aim we will construct a boundary matrix D as follows:

$$D_{ij} = \begin{cases} 1 \text{ if} \sigma_i \text{ is a codimension 1 coface of } \sigma_j \\ 0 \text{ otherwise} \end{cases}$$

The rows and columns are ordered according to the order in $\mathcal{K}$, and the boundary of each simplex is recorded in the columns of D. The following algorithm was first described in (...) and uses columns for reduction, hence now its commonly referred to as the column algorithm. Let $low(j)$ be the row index of the lowest nonzero entry in column j.

Algorithm 1 Column reduction algorithm for persistent homology

input : D boundary matrix representing a filtration
output R reduced boundary matrix
:
$R = D, V = I$
for $j = 1$ to n do
while exists $k < j : low(k) = low(j)$:
$R[:, j] = R[:, j] + R[:, k]$
$V[:, j] = V[:, j] + V[:, k]$
endwhile
endfor

The matrix R is the reduced version of D. We add columns from left to right to take into account the filtration order: at the filtration index j we can change the standard basis $\langle \sigma_1, \sigma_2, ...\sigma_j \rangle \rightarrow \langle \hat{\sigma_1}, ..., \hat{\sigma_j} \rangle$ of the simplicial complex using simplices already present in the filtration. In fact we will have a new basis of the form $D : V_j \rightarrow R_j$, where R_j represent (finite) n homology classes and V_j are the preimage chains under the boundary map.

Now, let us analyze the effect of adding a simplex at a time on the homology groups.

Lemma. Consider a simplex-wise filtration $K_i = K_{i-1} \cup \sigma_i$, i.e. in each step we add a single simplex. Assume that $\dim \sigma_i = n$. Then only two things can occur:

- Inclusion of σ_i **increases** H_n, and we say that σ_i is a **positive** simplex.
- Inclusion of σ_i **decreases** H_{n-1}. In this case σ_i is a **negative** simplex.

Proof:

The proof proceeds by analyzing the dimensions of the different subspaces. We have one of two cases.

Case 1. If the boundary of σ_i is a linear combination of boundaries in K_{i-1}, i.e. if $\partial_n(\sigma_i) \in \text{im}\partial_n(K_{i-1})$, then adding σ_i to the column of ∂_n will increase the dimension of the kernel $\ker \partial_n$. In particular, note that $R_i = 0$. Furthermore since we're adding an $n-$simplex, the group $C_{n+1}(K_i)$

will remain unchanged, and hence the im ∂_{n+1} will also remain unchanged. Thus the dimension of H_n will increase by 1.

Case 2. On the other hand, if the boundary of σ_i is *not* a linear combination of boundaries in K_{i-1}, adding σ_i to ∂_n will increase the dimensionality of im∂_n. Of importance, note that $R_i \neq 0$. Furthermore, since we're adding and $n-$simplex and the group C_{n-1} is unchanged implies that ker∂_{n-1} is unchanged. Hence, the inclusion of σ_i causes the dimension of H_{n-1} to decrease by 1 as desired. ■

Before proving that the reduction algorithm provides the persistence module, let us add another Lemma which suggests how the pivots or lowest ones of matrix R hold fundamental information of the persistent homology.

Lemma The pivots are invariant to the reduction process. In particular, we can further "sparsify" matrix R by left-to-right operations and this won't change the pivots.

Proof

After reduction, nonzero columns will be independent. Therefore adding columns from left to right will only change the basis of the image of D. Since by assumption, all pivots of matrix R will be different, further performing left-to-right operations will not affect the pivots. Thus the pivots are invariants of the reduction process. ■

Using the Lemma above, we now show that the "dynamics" of birth and death of homology classes in the filtration appear when $R = DV$ by noting "that R_j is a cycle appearing in the filtration at index $i = low(j)$ and becoming a boundary when σ_j enters the filtration at index j (and recorded in V_j), and when an essential cycle V_i appears at index i."[Ripser paper]

Now we proceed to show that the algorithm contains all the information of the persistence modules.

Theorem The output matrix R of the column reduction algorithm contains the persistence module of the filtration $\mathcal{K}$. Namely, the persistence intervals will be of the form $[a_i, a_j) \in Dgm_p(\mathcal{K})$ iff $i = low(j)$ in the reduced matrix R, and dim $\sigma_i = p$ and dim $\sigma_j = p + 1$.

Proof:

First, note that column j will reach its final form after the $j-$th iteration of the outer loop in the algorithm. At that moment, we have reduced all columns $k < j$.

By the Lemma above, $R_j = 0$ is a cycle born at index j.

Furthermore the non-zero columns in R correspond to $n-$homology classes that died by addition of a single simplex, so if $R_j \neq 0 \implies \tau_j$ killed the homology class in R_j.

Now, we need to show that the birth simplex is the pivot σ_i with $low(j) = i$. Simplex σ_i is possibly a birth simplex because $\sigma_i \in R_j$, and since $i < j$ it represents a class that must be paired. Assume that the class actually born at i is finite. Now, for a contradiction assume there is another simplex $\sigma_k \neq \sigma_i$ that is the birth simplex. But since simplices that are not pivots are not unique, they may be further reduced and since the birth simplex is unique, this leads to a contradiction. This suggests that σ_i is the birth simplex since it is invariant to the reduction process.

Finally, note that R_j cannot exist as a cycle before the addition of simplex i since if it could we would have been able to further reduce the column and get a lower pivot. Hence σ_i is precisely the birth simplex and the persistence pair is $[i, j)$. ■

Remark: A note on implementation. It turns out that any filtration can be refined into a filtration in which at each step, a single simplex is added, i.e. a simplex-level filtration. Therefore, a preprocessing step of the algorithm for computing persistent homology requires ordering the simplices. We mention a particular useful order in the context of our interest, Vietoris Rips filtrations. Namely, we order the simplices hierarchically by:

1. Diameter
2. Dimension of the simplex
3. Reverse lexicographic order (of vertices).

A note on cohomology

In a groundbreaking paper, De Silva, Morozov and Vejdemo-Johansson **dualities** showed that the persistence modules of homology and cohomology are isomorphic, namely that each persistence pair is flipped: if a persistent homology class is born at index i and dies at j the there is a corresponding persistent cohomology class that is born at index j and dies at index i. The flavor of the proof of this result is that, when we use Z_2 coefficients, there is no torsion, and the Universal Coefficient theorem tells us that the homology and cohomology groups are isomorphic. Using the previous observations, we can formally prove the result formally.

Theorem. Persistent homology and cohomology have identical barcodes when using Z_2 coefficients. (Proposition 2.3 **dualities**)

First, when cosidering Z_2 coefficients, the Universal Coefficient theorem tells us that the homology and cohomology groups are isomorphic, since there is no torsion.

Now consider the $k \to k+1$ step in the homology and cohomology persistence modules of a simplex-wise filtration:

$$\begin{array}{ccc} H_n(X_k) & \xrightarrow{i_*} & H_n(X_{k+1}) \\ \simeq\downarrow & & \downarrow\simeq \\ H^n(X_k) & \xleftarrow{j_*} & H^n(X_{k+1}) \end{array}$$

We have the following two cases, which follow from dimension counting:

- Case 1. A class is born in homology. Since the vertical arrows are isomorphisms, this implies that a class dies in cohomology.
- Case 2. Similarly, if a class dies in homology, then a class is born in cohomology.

This implies that the barcodes of homology and cohomology are the same, but flipped. ■

The above result is of fundamental importance, since computing the cohomology groups is more efficient than computing the homology groups. Hence the above result allows us to use the

cohomology barcode to study the homology barcode. This conceptual result has caused crucial algorithmic advances that enabled the computation of PH on large datasets.

In the following section we show how the persistence diagram completely algebraically characterizes all the information of data topology.

Quivers, persistence diagrams, and indecomposable representation

The underlying algebraic theory of persistence homology groups have taken many different approaches. Among them, one of the most prolific in terms of their capacity to prove important results and connect them to other mathematical theories is that of Quiver theory of persistence modules. In this section we will use the latter. **Quivers** are graph representations. In other words, a quiver is a directed graph where nodes represent vector spaces and arrows are linear maps. To see why this theory applies in our context, note that the sequence of homology groups induced by a filtration can be conceptualized as a quiver without much effort: we just need to work with coefficients in $\mathbb{Z}/n\mathbb{Z}$ and each homology group will have a vector space structure.

There is a vast and profound theory of quivers. To read a concise introduction we refer to [cite]. For our purposes we will only need the following theorem that characterizes quivers in terms of their isomorphism classes. Before proceeding we just need a couple of definitions.

Definition. A **persistence module** is an indexed family of vector spaces, together with a doubly indexed family of linear maps $(v_r^s : V_r \to V_s | r \leq s)$.

Remark Note that with the above definition v_r^r is the identity on V_r. *Remark* It turns out that a persistence module is a special type of quiver representation, which we define in what follows.

Definition. Quiver A *quiver* is a tuple $Q = (K^0, K^1)$, where K^0 is a set of vertices and K^1 is a finite set of (directed) edges.

Definition. Quiver representation Let Q be a quiver and K be a field. The set of finite dimensional vector spaces attached to each vertex and corresponding linear maps associated with arrows is a quiver representation.

Definition. Direct sum of quiver representations If V and W are representations of the same quiver Q, define their direct sum $V \bigoplus W$ by:

$$(V \bigoplus W)_x = V_x \bigoplus W_x \tag{3.77}$$

for all $x \in K^0$ and

$$(v \bigoplus w)_i^j = \begin{pmatrix} v_i^j & 0 \\ 0 & w_i^j \end{pmatrix} : V_i \bigoplus W_i \to V_j \bigoplus W_j \tag{3.78}$$

Definition. A representation V is **trivial** if $V_x = 0 \forall x \in K^0$.

Definition. (In)decomposable representation If a quiver representation U is isomorphic to a direct sum, i.e. $U \cong V \bigoplus W$, where V, W are non-trivial representations then U is called decomposable; otherwise U is called decomposable.

The classification of quivers boils down to classifying indecomposable representations.

Theorem. Classification of A_n type quivers (Gabriel).

Let Q be an A_n type quiver, and let K be a field. Then every indecomposable finite-dimensional representation of Q over K is *isomorphic* to a direct sum of interval representations $\mathbb{I}_Q[b, d]$ of the following form:

$$0 \to 0 \to \ldots \to K \xrightarrow{id} K \xrightarrow{\text{id}} \ldots \xrightarrow{\text{id}} K \to 0 \ldots \to 0, \tag{3.79}$$

where the length of the nonzero subspaces is of size $b - d$.

Remark: In the context of our **persistence module**, the interval modules are precisely indicators of the lifetime of persistent homology classes. Now the evocative notation lends itself to its useful interpretation: b, d correspond to the birth and death of the persistent homology class.

What this means for our purposes is that we can fully characterize the mathematical properties of persistence modules with the persistence diagram Dgm = $\bigoplus_i [b_i, d_i)$.

Persistence diagram. The disjoint union of intervals $[b, d)$ is the persistence diagram of $\mathbb{V}$.

3.8 Conclusion

This is all the necessary theory to understand the computational framework used in this thesis. As stated in the introduction, my hope is that this material will be useful to the reader to build new tools using algebraic topology and the theory of persistence.

BIBLIOGRAPHY

Aibar, Sara et al. (Nov. 2017). "SCENIC: single-cell regulatory network inference and clustering". en. In: *Nat Methods* 14.11. Publisher: Nature Publishing Group, pp. 1083–1086. ISSN: 1548-7105. DOI: 10.1038/nmeth.4463. URL: https://www.nature.com/articles/nmeth.4463 (visited on 03/11/2025).

Angerer, Philipp et al. (Dec. 2015). "destiny: diffusion maps for large-scale single-cell data in R". In: *Bioinformatics* 32.8, pp. 1241–1243. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv715. URL: https://doi.org/10.1093/bioinformatics/btv715.

Arnold, V. I. (1991). *Ordinary differential equations*. MIT Press.

Babonis, Leslie S. et al. (Feb. 2023). "Single-cell atavism reveals an ancient mechanism of cell type diversification in a sea anemone". en. In: *Nature Communications* 14.1, p. 885. ISSN: 2041-1723. DOI: 10.1038/s41467-023-36615-9. URL: https://www.nature.com/articles/s41467-023-36615-9 (visited on 02/25/2025).

Bauer, Ulrich (2021). "Ripser: efficient computation of Vietoris-Rips persistence barcodes". In: *J. Appl. Comput. Topol.* 5.3, pp. 391–423. ISSN: 2367-1726. DOI: 10.1007/s41468-021-00071-5. URL: https://doi.org/10.1007/s41468-021-00071-5.

Bauer, Ulrich et al. (2024). *Keeping it sparse: Computing Persistent Homology revisited*. arXiv: 2211.09075 [cs.CG]. URL: https://arxiv.org/abs/2211.09075.

Benjamin, Katherine et al. (June 2024). "Multiscale topology classifies cells in subcellular spatial transcriptomics". en. In: *Nature* 630.8018, pp. 943–949. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-07563-1. URL: https://www.nature.com/articles/s41586-024-07563-1 (visited on 12/18/2024).

Bergen, Volker et al. (Dec. 2020). "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *Nat Biotechnol* 38.12. Publisher: Nature Publishing Group, pp. 1408–1414. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0591-3. URL: https://www.nature.com/articles/s41587-020-0591-3 (visited on 03/13/2025).

Botnan, Magnus Bakke and Michael Lesnick (Mar. 2023). *An Introduction to Multiparameter Persistence*. arXiv:2203.14289 [math]. DOI: 10.48550/arXiv.2203.14289. URL: http://arxiv.org/abs/2203.14289 (visited on 12/18/2024).

Brabin, Charles, Peter J. Appleford, and Alison Woollard (2011). "The Caenorhabditis elegans GATA Factor ELT-1 Works through the Cell Proliferation Regulator BRO-1 and the Fusogen EFF-1 to Maintain the Seam Stem-Like Fate". In: *PLoS Genetics* 7.8. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002200.

Briggs, James A. et al. (2018). "The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution". In: *Science* 360.6392, eaar5780. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar5780. URL: https://www.science.org/doi/10.1126/science.aar5780.

Brown, Brielin C., Nicolas L. Bray, and Lior Pachter (Dec. 2018). "Expression reflects population structure". en. In: *PLOS Genetics* 14.12. Ed. by Anna Di Rienzo, e1007841. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007841. URL: https://dx.plos.org/10.1371/journal.pgen.1007841 (visited on 02/18/2025).

Budninskiy, Max et al. (2019). "Parallel Transport Unfolding: A Connection-Based Manifold Learning Approach". In: *SIAM Journal on Applied Algebra and Geometry* 3.2, pp. 266–291. DOI: 10.1137/18M1196133. eprint: https://doi.org/10.1137/18M1196133. URL: https://doi.org/10.1137/18M1196133.

Calderon, Diego et al. (2022). "The continuum of *Drosophila* embryonic development at single-cell resolution". In: *Science* 377.6606, eabn5800. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abn5800. URL: https://www.science.org/doi/10.1126/science.abn5800.

Calvo, D. (2001). "A POP-1 repressor complex restricts inappropriate cell type-specific gene transcription during Caenorhabditis elegans embryogenesis". In: *The EMBO Journal* 20.24, pp. 7197–7208. ISSN: 14602075. DOI: 10.1093/emboj/20.24.7197.

Cannoodt, Robrecht et al. (2021). "Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells". In: *Nature Communications* 12.1, p. 3942. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24152-2. URL: https://www.nature.com/articles/s41467-021-24152-2 (visited on 11/20/2024).

Cavanna, Nicholas J., Mahmoodreza Jahanseir, and Donald R. Sheehy (2015). "A Geometric Perspective on Sparse Filtrations". In: *Proceedings of the Canadian Conference on Computational Geometry*, pp. 116–121.

Chari, Tara et al. (Nov. 2021). "Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across *Clytia* medusa cell types". en. In: *Science Advances* 7.48, eabh1683. ISSN: 2375-2548. DOI: 10.1126/sciadv.abh1683. URL: https://www.science.org/doi/10.1126/sciadv.abh1683 (visited on 02/18/2025).

Chasis, Joel Anne and Narla Mohandas (Aug. 2008). "Erythroblastic islands: niches for erythropoiesis". In: *Blood* 112.3, pp. 470–478. ISSN: 0006-4971. DOI: 10.1182/blood-2008-03-077883. URL: https://doi.org/10.1182/blood-2008-03-077883 (visited on 03/11/2025).

Chazal, Frédéric, David Cohen-Steiner, et al. (2009). "Proximity of persistence modules and their diagrams". In: *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*. SCG '09. Aarhus, Denmark: Association for Computing Machinery, pp. 237–246. ISBN: 9781605585017. DOI: 10.1145/1542362.1542407. URL: https://doi.org/10.1145/1542362.1542407.

Chazal, Frédéric, Vin de Silva, and Steve Oudot (Dec. 2014). "Persistence stability for geometric complexes". In: *Geometriae Dedicata* 173.1, pp. 193–214. ISSN: 1572-9168. DOI: 10.1007/s10711-013-9937-z. URL: https://doi.org/10.1007/s10711-013-9937-z.

Chow, Andrew et al. (Apr. 2013). "CD169+ macrophages provide a niche promoting erythropoiesis under homeostasis and stress". en. In: *Nat Med* 19.4. Publisher: Nature Publishing Group, pp. 429–436. ISSN: 1546-170X. DOI: 10.1038/nm.3057. URL: https://www.nature.com/articles/nm.3057 (visited on 03/11/2025).

Cohen-Steiner, David, Herbert Edelsbrunner, and John Harer (Jan. 2007). "Stability of Persistence Diagrams". In: *Discrete & Computational Geometry* 37.1, pp. 103–120. ISSN: 0179-5376, 1432-0444. DOI: 10.1007/s00454-006-1276-5. URL: http://link.springer.com/10.1007/s00454-006-1276-5.

Dłotko, Paweł (Aug. 2012). "A fast algorithm to compute cohomology group generators of orientable 2-manifolds". en. In: *Pattern Recognition Letters* 33.11, pp. 1468–1476. ISSN: 01678655. DOI: 10.1016/j.patrec.2011.10.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167865511003370 (visited on 12/11/2024).

Farrell, Jeffrey A. et al. (2018). "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis". In: *Science* 360.6392, eaar3131. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar3131. URL: https://www.science.org/doi/10.1126/science.aar3131.

Gabriel, Peter (Mar. 1972). "Unzerlegbare Darstellungen I". In: *manuscripta mathematica* 6.1, pp. 71–103. ISSN: 0025-2611, 1432-1785. DOI: 10.1007/BF01298413. URL: http://link.springer.com/10.1007/BF01298413 (visited on 12/18/2024).

Giladi, Amir et al. (July 2018). "Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis". en. In: *Nat Cell Biol* 20.7. Publisher: Nature Publishing Group, pp. 836–846. ISSN: 1476-4679. DOI: 10.1038/s41556-018-0121-4. URL: https://www.nature.com/articles/s41556-018-0121-4 (visited on 03/10/2025).

Gilleard, J. S. and J. D. McGhee (2001). "Activation of hypodermal differentiation in the Caenorhabditis elegans embryo by GATA transcription factors ELT-1 and ELT-3". In: *Molecular and Cellular Biology* 21.7, pp. 2533–2544. ISSN: 0270-7306. DOI: 10.1128/MCB.21.7.2533-2544.2001.

Haniffa, Muzlifah et al. (Sept. 2021). "A roadmap for the Human Developmental Cell Atlas". en. In: *Nature* 597.7875, pp. 196–205. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03620-1. URL: https://www.nature.com/articles/s41586-021-03620-1 (visited on 12/19/2024).

Hatcher, Allen (2001). *Algebraic topology*. Cambridge University Press.

Hiraoka, Yasuaki et al. (2024). *Curse of Dimensionality on Persistence Diagrams*. arXiv: 2404.18194 [math.ST]. URL: https://arxiv.org/abs/2404.18194.

Hopf, Heinz (1927). "Vektorfelder inn-dimensionalen Mannigfaltigkeiten". In: *Mathematische Annalen* 96.1, pp. 225–249. ISSN: 0025-5831, 1432-1807. DOI: 10.1007/BF01209164. URL: http://link.springer.com/10.1007/BF01209164 (visited on 11/22/2023).

Horst, Suzanne E. M. van der et al. (2019). "C. elegans Runx/CBF suppresses POP-1 TCF to convert asymmetric to proliferative division of stem cell-like seam cells". In: *Development* 146.22, dev180034. ISSN: 0950-1991. DOI: 10.1242/dev.180034. URL: https://doi.org/10.1242/dev.180034.

Hu, Minjie et al. (June 2020). "Lineage dynamics of the endosymbiotic cell type in the soft coral Xenia". en. In: *Nature* 582.7813, pp. 534–538. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2385-7. URL: https://www.nature.com/articles/s41586-020-2385-7 (visited on 02/27/2025).

Klei, Thomas R. L. et al. (Feb. 2017). "From the Cradle to the Grave: The Role of Macrophages in Erythropoiesis and Erythrophagocytosis". English. In: *Front. Immunol.* 8. Publisher: Frontiers. ISSN: 1664-3224. DOI: 10.3389/fimmu.2017.00073. URL: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2017.00073/full (visited on 03/11/2025).

Klein, Allon M. et al. (2015). "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5, pp. 1187–1201. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2015.04.044. URL: https://www.sciencedirect.com/science/article/pii/S0092867415005000.

Kouns, Nathaniel A. et al. (2011). "NHR-23 dependent collagen and hedgehog-related genes required for molting". In: *Biochemical and Biophysical Research Communications* 413.4, pp. 515–520. ISSN: 0006291X. DOI: 10.1016/j.bbrc.2011.08.124.

Kowalczyk, Monika S. et al. (2015). "Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells". In: *Genome Research* 25.12, pp. 1860–1872. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.192237.115. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.192237.115.

La Manno, Gioele et al. (Aug. 2018). "RNA velocity of single cells". In: *Nature* 560.7719. Publisher: Nature Publishing Group, pp. 494–498. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0414-6. URL: https://www.nature.com/articles/s41586-018-0414-6 (visited on 03/13/2025).

Lambert, Samuel A. et al. (Feb. 2018). "The Human Transcription Factors". English. In: *Cell* 172.4. Publisher: Elsevier, pp. 650–665. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.01.029. URL: https://www.cell.com/cell/abstract/S0092-8674(18)30106-5 (visited on 03/14/2025).

Lange, Merlin et al. (2023). "Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State-Transition Dynamics of Late-Vertebrate Pluripotent Axial Progenitors". In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2023/06/14/2023.03.06.531
DOI: 10.1101/2023.03.06.531398. URL: https://www.biorxiv.org/content/early/2023/06/14/2023.03.06.531398.

Laurenti, Elisa and Berthold Göttgens (Jan. 2018). "From haematopoietic stem cells to complex differentiation landscapes". In: *Nature* 553.7689. Publisher: Nature Publishing Group. ISSN: 1476-4687. DOI: 10.1038/nature25022. URL: https://www.nature.com/articles/nature25022 (visited on 03/10/2025).

Levy, B. (2006). "Laplace-Beltrami Eigenfunctions Towards an Algorithm That "Understands" Geometry". In: *IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*. IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06). IEEE, pp. 13–13. ISBN: 978-0-7695-2591-4. DOI: 10.1109/SMI.2006.21. URL: http://ieeexplore.ieee.org/document/1631196/.

Levy, Shani et al. (2021). "A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity". In: *Cell* 184.11, 2973–2987.e18. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2021.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0092867421004402.

Liao, Yuan et al. (2022). "Cell landscape of larval and adult Xenopus laevis at single-cell resolution". In: *Nature Communications* 13.1, p. 4306. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31949-2. URL: https://www.nature.com/articles/s41467-022-31949-2.

Liberali, Prisca and Alexander F. Schier (July 2024). "The evolution of developmental biology through conceptual and technological revolutions". en. In: *Cell* 187.14, pp. 3461–3495. ISSN: 00928674. DOI: 10.1016/j.cell.2024.05.053. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867424006329 (visited on 04/08/2025).

Link, Oliver et al. (2023). "A cell-type atlas from a scyphozoan jellyfish Aurelia coerulea (formerly sp.1) provides insights into changes of cell-type diversity in the transition from polyps to medusae". In: *bioRxiv*. DOI: 10.1101/2023.08.24.554571. eprint: https://www.biorxiv.org/content/early/2023/08/26/2023.08.24.554571.full.pdf. URL: https://www.biorxiv.org/content/early/2023/08/26/2023.08.24.554571.

Lyons, Karen M., Brigid L.M. Hogan, and Elizabeth J. Robertson (1995). "Colocalization of BMP 7 and BMP 2 RNAs suggests that these factors cooperatively mediate tissue interactions during murine development". In: *Mechanisms of Development* 50.1, pp. 71–83. ISSN: 0925-4773. DOI: https://doi.org/10.1016/0925-4773(94)00326-I. URL: https://www.sciencedirect.com/science/article/pii/092547739400326I.

Macosko, Evan Z. et al. (May 2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5. Publisher: Elsevier, pp. 1202–1214. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.05.002. URL: https://doi.org/10.1016/j.cell.2015.05.002 (visited on 04/08/2025).

Maggs, Kelly et al. (2025). "CocycleHunter: cohomology-based circular gene set enrichment and genetic phase estimation in single-cell RNA-seq data". In: *bioRxiv*. DOI: 10.1101/2025.01.09.632214. eprint: https://www.biorxiv.org/content/early/2025/01/14/2025.01.09.632214.full.pdf. URL: https://www.biorxiv.org/content/early/2025/01/14/2025.01.09.632214.

Meli, Vijaykumar S. et al. (May 2010). "MLT-10 Defines a Family of DUF644 and Proline-rich Repeat Proteins Involved in the Molting Cycle of *Caenorhabditis elegans*". en. In: *Molecular Biology of the Cell* 21.10. Ed. by Josephine C. Adams, pp. 1648–1661. ISSN: 1059-1524, 1939-4586. DOI: 10.1091/mbc.e08-07-0708. URL: https://www.molbiolcell.org/doi/10.1091/mbc.e08-07-0708 (visited on 03/07/2025).

Moon, Kevin R. et al. (Dec. 2019). "Visualizing structure and transitions in high-dimensional biological data". en. In: *Nature Biotechnology* 37.12, pp. 1482–1492. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0336-3. URL: https://www.nature.com/articles/s41587-019-0336-3 (visited on 04/26/2025).

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1226. URL: https://www.nature.com/articles/nmeth.1226 (visited on 04/08/2025).

Munkres, J.R. (2000). *Topology*. Featured Titles for Topology. Prentice Hall, Incorporated. ISBN: 9780131816299. URL: https://books.google.com/books?id=XjoZAQAAIAAJ.

Nigmetov, Arnur and Dmitriy Morozov (2022). *Topological Optimization with Big Steps*. Version Number: 2. DOI: 10.48550/ARXIV.2203.16748. URL: https://arxiv.org/abs/2203.16748 (visited on 12/11/2024).

Orkin, Stuart H. and Leonard I. Zon (Feb. 2008). "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology". In: *Cell* 132.4, pp. 631–644. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.01.025. URL: https://www.sciencedirect.com/science/article/pii/S0092867408001256 (visited on 03/11/2025).

Oudot, Steve Y. (2015). *Persistence theory: from quiver representations to data analysis*. Mathematical Surveys and Monographs volume 209. American Mathematical Society. 218 pp. ISBN: 978-1-4704-3443-4 978-1-4704-2545-6.

Packer, Jonathan S. et al. (2019). "A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution". In: *Science* 365.6459, eaax1971. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax1971. URL: https://www.science.org/doi/10.1126/science.aax1971.

Park, Eunji et al. (2012). "Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record". In: *Molecular Phylogenetics and Evolution* 62.1, pp. 329–345. ISSN: 1055-7903. DOI: https://doi.org/10.1016/j.ympev.2011.10.008. URL: https://www.sciencedirect.com/science/article/pii/S1055790311004374.

Pflugrad, Amy et al. (May 1997). "The Groucho-like transcription factor UNC-37 functions with the neural specificity gene *unc-4* to govern motor neuron identity in *C. elegans*". en. In: *Development* 124.9, pp. 1699–1709. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.124.9.1699. URL: https://journals.biologists.com/dev/article/124/9/1699/39738/The-Groucho-like-transcription-factor-UNC-37 (visited on 03/07/2025).

Plass, Mireya et al. (2018). "Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics". In: *Science* 360.6391, eaaq1723. DOI: 10.1126/science.aaq1723. URL: https://www.science.org/doi/abs/10.1126/science.aaq1723.

Popescu, Dorin-Mirel et al. (Oct. 2019). "Decoding human fetal liver haematopoiesis". en. In: *Nature* 574.7778. Publisher: Nature Publishing Group, pp. 365–371. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1652-y. URL: https://www.nature.com/articles/s41586-019-1652-y (visited on 03/11/2025).

Qiu, Chengxiang et al. (Feb. 2024). "A single-cell time-lapse of mouse prenatal development from gastrula to birth". en. In: *Nature* 626.8001, pp. 1084–1093. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-07069-w. URL: https://www.nature.com/articles/s41586-024-07069-w (visited on 11/29/2024).

Rand, David A. et al. (2021). "Geometry of gene regulatory dynamics". In: *Proceedings of the National Academy of Sciences* 118.38, e2109729118. DOI: 10.1073/pnas.2109729118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2109729118.

Richter, Daniel J. and Nicole King (Nov. 2013). "The Genomic and Cellular Foundations of Animal Origins". en. In: *Annual Review of Genetics* 47.1, pp. 509–537. ISSN: 0066-4197, 1545-2948. DOI: 10.1146/annurev-genet-111212-133456. URL: https://www.annualreviews.org/doi/10.1146/annurev-genet-111212-133456 (visited on 04/08/2025).

Rizvi, Abbas H et al. (2017). "Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development". In: *Nature Biotechnology* 35.6, pp. 551–560. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3854. URL: https://www.nature.com/articles/nbt.3854.

Saelens, Wouter et al. (2019). "A comparison of single-cell trajectory inference methods". en. In: *Nature Biotechnology* 37.5, pp. 547–554. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0071-9. URL: https://www.nature.com/articles/s41587-019-0071-9 (visited on 04/26/2025).

Sander, Klaus (1997). "Shaking a concept: Hans Driesch and the varied fates of sea urchin blastomeres". In: *Landmarks in Developmental Biology 1883–1924: Historical Essays from Roux's Archives*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 29–31. ISBN: 978-3-642-60492-8. DOI: 10.1007/978-3-642-60492-8_10. URL: https://doi.org/10.1007/978-3-642-60492-8_10.

Schwabe, Daniel et al. (Nov. 2020). "The transcriptome dynamics of single cells during the cell cycle". en. In: *Molecular Systems Biology* 16.11, e9946. ISSN: 1744-4292, 1744-4292. DOI: 10.15252/msb.20209946. URL: https://www.embopress.org/doi/10.15252/msb.20209946 (visited on 02/28/2025).

Scoccola, Luis et al. (2023). "Toroidal Coordinates: Decorrelating Circular Coordinates with Lattice Reduction". en. In: *LIPIcs, Volume 258, SoCG 2023* 258, 57:1–57:20. ISSN: 1868-8969. DOI: 10.4230/LIPICS.SOCG.2023.57. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SoCG.2023.57 (visited on 12/11/2024).

Sebé-Pedrós, Arnau, Bernard M. Degnan, and Iñaki Ruiz-Trillo (Aug. 2017). "The origin of Metazoa: a unicellular perspective". en. In: *Nature Reviews Genetics* 18.8, pp. 498–512. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg.2017.21. URL: https://www.nature.com/articles/nrg.2017.21 (visited on 04/08/2025).

Sebé-Pedrós, Arnau, Baptiste Saudemont, et al. (May 2018). "Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq". en. In: *Cell* 173.6, 1520–1534.e20. ISSN: 00928674. DOI: 10.1016/j.cell.2018.05.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867418305968 (visited on 02/27/2025).

Setty, Manu et al. (2016). "Wishbone identifies bifurcating developmental trajectories from single-cell data". In: *Nature Biotechnology* 34.6, pp. 637–645. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3569. URL: https://www.nature.com/articles/nbt.3569.

Sheehy, Donald R. (2020). "One Hop Greedy Permutations". In: *Proceedings of the 32nd Canadian Conference on Computational Geometry*, pp. 221–225.

Siebert, Stefan et al. (2019). "Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution". In: *Science* 365.6451, eaav9314. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aav9314. URL: https://www.science.org/doi/10.1126/science.aav9314.

Silva, Vin de, Dmitriy Morozov, and Mikael Vejdemo-Johansson (2011). "Dualities in persistent (co)homology". In: *Inverse Problems* 27.12, p. 124003. DOI: 10.1088/0266-5611/27/12/124003. URL: https://dx.doi.org/10.1088/0266-5611/27/12/124003.

Singh, Gurjeet, Facundo Mémoli, Gunnar E Carlsson, et al. (2007). "Topological methods for the analysis of high dimensional data sets and 3d object recognition." In: *PBG@ Eurographics* 2, pp. 091–100.

Sørensen, Mikael, Charilaos I. Kanatsoulis, and Nicholas D. Sidiropoulos (2021). "Generalized Canonical Correlation Analysis: A Subspace Intersection Approach". In: *IEEE Transactions on Signal Processing* 69, pp. 2452–2467. DOI: 10.1109/TSP.2021.3061218.

Steger, Julia et al. (2022). "Single-cell transcriptomics identifies conserved regulators of neuroglandular lineages". In: *Cell Reports* 40.12, p. 111370. ISSN: 22111247. DOI: 10.1016/j.celrep.2022.111370. URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124722012025.

Su, Zhe, Yiying Tong, and Guo-Wei Wei (Apr. 2024). "Hodge Decomposition of Single-Cell RNA Velocity". In: *J. Chem. Inf. Model.* 64.8. Publisher: American Chemical Society, pp. 3558–3568. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.4c00132. URL: https://doi.org/10.1021/acs.jcim.4c00132 (visited on 03/13/2025).

Sulston, J. E. et al. (1983). "The embryonic cell lineage of the nematode Caenorhabditis elegans". In: *Developmental Biology* 100.1, pp. 64–119. ISSN: 0012-1606. DOI: https://doi.org/10.1016/0012-1606(83)90201-4.

Suo, Chenqu et al. (June 2022a). "Mapping the developing human immune system across organs". en. In: *Science* 376.6597, eabo0510. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abo0510. URL: https://www.science.org/doi/10.1126/science.abo0510 (visited on 12/19/2024).

— (May 2022b). "Mapping the developing human immune system across organs". In: *Science* 376.6597. Publisher: American Association for the Advancement of Science, eabo0510. DOI: 10.1126/science.abo0510. URL: https://www.science.org/doi/10.1126/science.abo0510 (visited on 03/10/2025).

Tralie, Christopher, Nathaniel Saul, and Rann Bar-On (2018). "Ripser.py: A Lean Persistent Homology Library for Python". In: *Journal of Open Source Software* 3.29, p. 925. DOI: 10.21105/joss.00925. URL: https://doi.org/10.21105/joss.00925.

Trapnell, Cole et al. (Apr. 2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". en. In: *Nature Biotechnology* 32.4, pp. 381–386. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2859. URL: https://www.nature.com/articles/nbt.2859 (visited on 04/26/2025).

Velten, Lars et al. (Apr. 2017). "Human haematopoietic stem cell lineage commitment is a continuous process". en. In: *Nat Cell Biol* 19.4. Publisher: Nature Publishing Group, pp. 271–281. ISSN: 1476-4679. DOI: 10.1038/ncb3493. URL: https://www.nature.com/articles/ncb3493 (visited on 03/10/2025).

Vipond, Oliver et al. (2021). "Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors". In: *Proceedings of the National Academy of Sciences* 118.41, e2102166118. DOI: 10.1073/pnas.2102166118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2102166118.

Virtanen, Pauli et al. (2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Waddington, C. H. (Apr. 2014). *The Strategy of the Genes*. London: Routledge. ISBN: 978-1-315-76547-1. DOI: 10.4324/9781315765471.

— (Apr. 1957). *The Strategy of the Genes*. en. Routledge. ISBN: 978-1-315-76547-1. DOI: 10.4324/9781315765471.

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson (2016). “How to Use t-SNE Effectively”. In: *Distill*. DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

Zhang, Tengjiao et al. (2020). “A single-cell analysis of the molecular lineage of chordate embryogenesis”. In: *Science Advances* 6.45, eabc4773. DOI: 10.1126/sciadv.abc4773. URL: https://www.science.org/doi/abs/10.1126/sciadv.abc4773.