IMPERIAL

Imperial College London
Department of Mathematics

Topological Invariants for Data: Duality, Stability, and Applications to Inference and Machine Learning

Inés García-Redondo

Supervised by

Dr Anthea Monod

Thesis submitted for the degree of
Doctor of Philosophy in Mathematics
of the Imperial College London,
7th of August, 2025

Statement of Originality

I certify that this thesis and its associated research are my own work, and that any ideas or contents from the work of other people, whether published or unpublished, are properly acknowledged in accordance with standard referencing practices. Portions of this thesis are based on collaborative work. Specifically, part of the content of **Chapter** 2 is joint work with Anthea Monod and Anna Song. The first half of the material in **Chapter** 3 arises from joint work with Pierre Faugère, Gregory Henselman-Petrusek, Anthea Monod and Qiquan Wang; while the second half is joint work with the latter two authors from that list. The content of **Chapter** 4 is based on two joint first-authored publications with Aideen Fay, Charlie B. Tan and Qiquan Wang; co-authored with Michael M. Bronstein, Haim Dubossarsky and Anthea Monod.

Copyright Statement

Abstract

Topological data analysis (TDA) is a modern field at the intersection of algebraic topology, statistics, and data science, focused on extracting meaningful insights from complex data where *shape* and *size* matter. A central tool in TDA is *persistent homology*, which extends classical homology to study topological features across multiple scales via *filtrations*. The resulting algebraic object, the *persistence module*, is summarized through topological invariants such as *persistence barcodes*, which record the lifespan of features in the data. In *multiparameter persistence*, where filtrations are indexed by multiple parameters, the theory becomes more intricate, motivating a rich line of recent research.

This thesis investigates persistent homology in both single- and multiparameter settings, combining theoretical contributions with applications in data analysis and machine learning. The work is organized into three parts.

The first part studies *dualities* in persistence. While persistent *cohomology* is well-understood and leveraged for computational efficiency in the single-parameter case, it remains underdeveloped in the multiparameter setting. We introduce a theoretical framework for dualities in multiparameter persistence and implement a cohomology-based pipeline to match barcodes from different data sources in the single-parameter case.

The second part focuses on *functional invariants*, which integrate more naturally into statistical and machine learning workflows than barcodes. We prove stability results for the *rank invariant* and establish a functional central limit theorem for the *multiparameter persistence landscape*, enabling the construction of confidence bands.

The final part presents two machine learning applications. First, we use persistent homology to estimate the intrinsic dimension of neural network training trajectories and relate it to generalization. Second, we apply it to interpret model behavior under adversarial attacks on large language models.

Acknowledgements

First and foremost, I would like to thank Anthea Monod, for her unwavering support and invaluable mentorship and guidance during these years. I thank her for giving me the freedom to explore the topics that captured my curiosity, always trusting in my instincts; for all the thoughtful advice when I was unsure of next steps, and for all the opportunities that she has provided during my doctoral journey.

I am also grateful to my amazing research group, and especially, to Aideen, Anna, Arne, Daniele, Roan, Qi and Yueqi, who have shared much of this PhD journey with me. I thank them for the mathematical discussions, their support in preparing talks and navigating challenges, and most importantly, for making work and traveling to conferences more fun when we went in group.

I thank the LSGNT and all its cohorts for being an incredible community, and creating the nicest environment possible to do a PhD in London. In particular, I want to thank Ilaria, Marta and Sara for their support, advice and friendship over the years, and for being role models and sources of inspiration.

I thank my friends in all parts of the world, new and old, from Murcia to London, passing through Madrid. Special thanks to my Spanish friends in London: Marta, Ángela, Andrea and Natalia; and my choir friends, particularly Eurydice and Arne (again!), for making this city feel like home and filling these years with joy.

I thank my family for always supporting me, even when I keep going further away from home. Thank you for always being on the other side of a phone call whenever I felt happy, frustrated, angry, disappointed or homesick; and for being the best example of hard work, kindness and love that I could have asked for.

And lastly, I thank Javi, for always having my back and never doubting of this adventure, even being half the world away.

I acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC), through the EPSRC Centre for Doctoral Training in Geometry and Number Theory (The London School of Geometry and Number Theory), University College London, Imperial College London, and King's College London; grant No. EP/S021590/1.

Contents

Introduction

The explosive growth in both the volume and complexity of data across scientific domains has brought renewed attention to a fundamental challenge in modern data analysis: how to effectively extract and interpret the underlying structure of real-world datasets. One particularly subtle yet powerful form of structure is *shape*. Capturing the *observable* shape in data, or detecting its *hidden* shape, can often be key to understanding the phenomena under exploration.

In some cases, it might happen that the geometric or topological properties of the object of study play a central role in the analysis. For example, in medical imaging, the external morphology and internal cavities of glioblastoma multiforme, a highly aggressive and poorly understood brain tumor, have been shown to carry predictive value for clinical outcomes [Cra+20]. Similarly, the process of protein folding, which transforms linear amino acid sequences into intricate three-dimensional structures, is known to produce shapes that are *knotted* while others are *unknotted*—a purely topological condition. This distinction correlates with functional differences, having profound implications for biology [Ben+23].

However, shape can also play a subtler role. In contrast to the cases above, where we directly study the morphology of an object; in other instances, the crucial information may instead reside in the geometric or topological structure of the dataset itself—frequently hidden within its high dimensionality. An interesting example is the conformation space of cyclo-octane, where the relative positions of the molecule's atoms theoretically occupy a 72-dimensional space. Yet, due to the internal symmetries in the molecule and its physical constraints, the relevant configurations lie on a lower-dimensional manifold that encodes meaningful structural information [MW11].

Topological Data Analysis and Persistent Homology

Classical statistical tools, particularly linear methods, often fall short when it comes to capturing these intricate notions of shape. In this thesis, we adopt an alternative perspective:

topological data analysis, which offers a set of techniques rooted in algebraic topology and homological algebra designed to detect and quantify the shape of data in a way that is robust to noise and adaptable to diverse data modalities. Among these tools, in this thesis we focus in one of the most established methods: *persistent homology*.

Persistent homology provides a framework to uncover *multi-scale topological features* in data, such as components, loops, voids, and their higher-dimensional analogues; encoding them in *topological invariants*. Arguably, the most well-known invariant is the *persistence barcode*, a collection of bars reflecting the "lifetimes" of topological features in the data, in terms of scale. Persistent homology has been successfully implemented across various disciplines: assessing coverage in sensor networks [dG07], exploring the structural properties of proteins [Gam+14; Kov+16; XLM18], investigating the 3D folding of DNA [ESR16], goal-directed path planning in robotics [BGK15], trajectory classification [PHR16] and estimating the energetic cost of bipedal walking [VAB13], to name a few. These diverse applications underscore the flexibility and power of persistent homology in addressing complex, shape-driven questions in data analysis.

Persistent Homology: Current Challenges in Theory and Practice

The field of *persistence*, which encompasses the theory and applications of persistent homology, is inherently multidisciplinary, a quality that is reflected throughout this thesis.

At one end of the spectrum lies the theoretical foundation of persistent homology. Its construction is rooted in homological algebra and representation theory. Although persistent homology was introduced in a very concrete scenario—studying the homology groups of a growing family of topological spaces—the underlying structure has since been abstracted into the more general notion of *persistence modules*. These can be viewed categorically, as functors from a poset category to the category of vector spaces, or representation-theoretically, as representations of quivers. This perspective will be exploited in the second chapter of the thesis. The general definition of persistence modules includes the setting on *multiparameter persistence*, an extension of the classical theory that allows to construct the families of topological spaces above depending on more than one parameter. Multiparameter persistence will be a recurring theme in this work.

From a more applied viewpoint, integrating persistent homology into data analysis and statistical pipelines presents significant challenges. First, the space of persistence barcodes, despite being amenable to statistical analysis, has a complicated geometry with undesirable properties such as non-uniqueness of means or geodesics, hindering the development of rigorous statistical theory. In addition, barcodes do not reside in Euclidean spaces, which most machine learning models take as input, and thus cannot be directly implemented within these pipelines.

Often, we need to take a pre-processing step known as *vectorization* where we embed them into a Hilbert space, potentially losing information. Finally, persistent homology suffers from severe *computational constraints*, especially, when applied to high-dimensional data. All these limitations serve as key motivations for several contributions presented in this thesis.

At the other end of the spectrum are applications that *leverage* persistent homology to yield meaningful insights. In the last chapter of this thesis, we explore how persistent homology can be applied to *machine learning systems* to deepen our understanding of artificial intelligence. There has been an increasing trend within the artificial intelligence community to incorporate tools from geometry and topology into modern learning models. One of the reasons to advocate for the use of persistent homology in this context is that it is inherently well-suited to investigate higher-order interactions, extending from pairwise interactions typical from graph-like structures. In fact, a recent position paper [Pap+24] underscores the need for topological frameworks and highlights topology as the "new frontier for relational learning".

Contributions and Structure of the Thesis

This thesis brings together five research contributions—three peer-reviewed publications, one work accepted for publication and one preprint currently under review—with some additional ongoing research, all organized into three thematic chapters. Each chapter explores a distinct research question about persistent homology and its role in machine learning and data analysis. The thesis begins with **Chapter** 1, an introductory exposition of the theory of persistent homology intended for a general scientific audience.

In **Chapter** 2, we explore the role of dualities in persistent homology from a theoretical and a practical perspective, with a focus on the work presented in [GMS24] and ongoing unpublished research. Persistent homology naturally admits a dual mathematical formulation known as persistent *cohomology* which has proven to be instrumental in practice. Perhaps the most notable example is its use in `Ripser` [Bau21], the state-of-the-art software for persistent homology computations, which achieves remarkable performance by leveraging cohomological methods. Building on this foundation, our work in [GMS24] introduces an *efficient methodology for comparing and matching topological signals* across different datasets using persistent cohomology. Specifically, we implemented a pipeline to match bars in persistence barcodes, representing topological features of distinct point clouds, in a way that respects geometric similarity. We released open-source software implementing this methodology that dramatically accelerates the computations, reducing processing time from days to just minutes using standard high-performance computing resources.

In **Chapter** 3, we address the challenge of integrating persistent homology into statistical inference and machine learning by adopting functional summaries instead of the persistence

barcode. This approach is central to my collaborative works [Wan+24] and [GMW25], where we also extend these tools to the more expressive framework of *multiparameter persistence*. In [Wan+24] we studied the *rank function*, focusing on its *stability and performance* in real-data settings. Here by *stability*, we mean a crucial property in topological data analysis that describes the resilience of topological descriptors to noisy perturbations in the data. We presented new stability results for rank functions and developed machine learning pipelines for classification and prediction tasks using open-source biological data, showing an improvement in performance when compared to non-topological approaches. In [GMW25] we instead focused on the *persistence landscape*, another powerful functional invariant. We derived a functional central limit theorem for multiparameter persistence landscapes, enabling the construction of *confidence bands*. This work represents one of the first rigorous methodologies for uncertainty quantification in multiparameter persistence.

Lastly, **Chapter** 4 presents two applications of persistent homology to data generated by AI systems, focusing on two key areas: understanding *generalization* in neural networks and advancing *interpretability* in large language models in adversarial settings. These contributions are drawn from my joint first-authored works: [Tan+24] and [Fay+25]. In [Tan+24], we investigate the proposed link between the generalization gap of a network (the performance drop from training to test data) and the fractal dimension (quantified using persistent homology) of its optimization trajectory. Through rigorous statistical analysis and two illustrative counterexamples, we found evidence that challenges the established theory. In [Fay+25], we propose persistent homology as a tool to interpret latent space dynamics in large language models under adversarial conditions, specifically backdoor fine-tuning and prompt injection. Across six state-of-the-art large language models, we show that such attacks consistently compress latent topologies suppressing fine-scale structure while amplifying coarse features. These signatures are robust across model sizes, layers, and architectures, and align with the progressive manifestation of adversarial behavior. Our results highlight persistent homology as a scalable, model-agnostic interpretability tool that captures both global structure and local nonlinear effects in large language model activations.

Impact Statement

The contributions of this thesis range from the development of practical tools to theoretical insights, with relevance both to current applications and to ongoing interdisciplinary research.

On the practical side, two concrete outcomes are the open-source software tools we developed: one for cycle matching (`https://github.com/inesgare/interval-matching`) and another for computing confidence bands for persistence landscapes (`https://github.com/inesgare/bands-mph-landscapes`). These tools are freely available to the research community and have

already been adopted in other studies. For instance, cycle matching was used in [Eas+23] as a dimensionality reduction technique to investigate structural changes in brain representations from resting-state fMRIs; and in [MB24] to develop "cross matched prevalence images" via subsampling techniques. These applications underscore the immediate relevance and versatility of the software.

The potential of using topological representations of data for real-world applications is sometimes limited by the scarce availability of statistical and inference methods. Our contributions [Wan+24] and [GMW25] directly address this gap. In them, we enable the implementation of new methodologies and foster the integration of these summaries within machine learning pipelines. In [Wan+24] we already presented two study-cases where using rank functions yields an improvement in performance for clinical tasks such as enhancing detection of stroke patients from electrocardiograms and refining the classification of malign lung tumors. These are early examples, and the broader clinical impact of these methods remains a promising avenue for future exploration.

Finally, this thesis aims to contribute to the growing dialogue between two traditionally distinct communities: pure mathematicians (particularly topologists and algebraists) and machine learning practitioners. By drawing on ideas from topology and geometry, this work aligns with an emerging trend in machine learning that seeks to incorporate more structured and theoretically grounded approaches into model design and analysis. This cross-pollination of ideas has the potential to offer new perspectives and tools, helping to improve the interpretability, robustness, and theoretical foundations of machine learning systems, which I am keen on exploring in the next stages of my academic research career.

1 Persistent Homology

We begin by introducing Persistent Homology (PH), the central technical tool of this thesis. This introduction is aimed at non-specialists and assumes no prior knowledge of algebraic topology. Only the material necessary to follow this thesis is covered, although pointers to further reading and other references are included throughout the text.

This chapter follows the structure of the PH pipeline, depicted in **Figure** 1.1. We start studying how to turn various types of input data into combinatorial/geometric representations called *filtrations* (**Section** 1.1) and how to turn these representations into topological ones using the theory of *homology* to get *persistence modules* (**Section** 1.2). We then cover different types of topological *invariants* one can extract from these representations of data (**Section** 1.3 and **Section** 1.4); and close the chapter with an overview of a key property of PH called *stability* (**Section** 1.5), which justifies its use in data analysis. For thorough introductions to the theory of persistence, we refer to [EH10] or [Oud15].

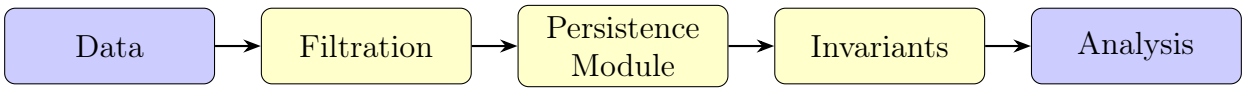


Figure 1.1: Flowchart diagram of the PH pipeline.

1.1 Simplicial Complexes and Filtrations

In this section we introduce simplicial complexes and explain how to define them from various types of input data. We use this definition to construct filtrations, the first step in the PH pipeline in **Figure** 1.1. See [Ghr14, Chapter 2] for further examples and details on these constructions.

Simplicial Complexes. An *abstract simplicial complex* X over a discrete set (or point cloud) S is a collection of subsets of S closed under inclusion: if $\sigma \in X$ and $\tau \subset \sigma$, then $\tau \in X$.

We can assign an orientation to X by fixing an order on the vertices, e.g., if $S = \{s_1, \ldots, s_n\}$, we set $s_1 < s_2 < \cdots < s_n$. A k-simplex is a subset of S with $k+1$ elements, denoted $[s_{i_0}, s_{i_1}, \ldots, s_{i_k}]$, with induced orientation given by $s_{i_0} < s_{i_1} < \cdots < s_{i_k}$. The set of all k-simplices is denoted X_k, and the union of all simplices up to dimension k forms the *k-skeleton*, denoted $X_{\leq k} \coloneqq \bigcup_{j \leq k} X_j$. The *dimension* of a complex is the largest dimension among its simplices. A common example is given by *graphs*, 1-dimensional simplicial complexes presented as $G = (V, E)$, where V is the set of vertices (0-simplices) and E the set of edges (1-simplices).

Vietoris–Rips and Čech Complexes. There are several standard methods for constructing simplicial complexes from input data. Given a point cloud S in a metric space with distance d (also called a finite metric space) and a scale parameter $\epsilon > 0$, one common approach is to define the *Vietoris–Rips* (VR) *complex* at scale ϵ, denoted $\mathcal{VR}_\epsilon(S, d)$ or simply $\mathcal{VR}_\epsilon(S)$ when the metric is clear from the context. This complex consists of all simplices whose vertices are pairwise at most ϵ apart, or more specifically,

$$\mathcal{VR}_\epsilon(S, d) \coloneqq \{\emptyset \neq \sigma \subset S : \operatorname{diam}(\sigma) \leq \epsilon\},$$

where $\operatorname{diam}(\sigma)$ denotes the diameter of the simplex. The VR complex is straightforward to define and can be recovered from the 1-skeleton, which is an advantage for storage requirements, but its size can grow exponentially with the number of points: if $|S| = n$, then $|\mathcal{VR}_\epsilon(S)| \in O(2^n)$ (see Table 1 in [Ott+17]).

An alternative is the Čech complex at scale ϵ, denoted $\check{\mathcal{C}}_\epsilon(S, d)$ or $\check{\mathcal{C}}_\epsilon(S)$. A simplex belongs to the Čech complex if the closed balls of radius ϵ centered at its vertices have a non-empty common intersection. Despite this condition being stronger, the VR and Čech complexes have similar complexity: if S is a subset of an Euclidean space, we have that $\check{\mathcal{C}}_\epsilon(S) \subseteq \mathcal{VR}_\epsilon(S) \subseteq \check{\mathcal{C}}_{\sqrt{2}\epsilon}(S)$. However, the Čech complex enjoys stronger theoretical guarantees. Notably, the *Nerve Theorem* implies that, under suitable conditions (sufficiently dense sampling of a manifold in an Euclidean space and an appropriate choice of ϵ) the Čech complex recovers the correct topological structure (i.e., it has the same homotopy type) of the underlying space we are sampling from [NSW08, Theorem 3.1].

Filtrations. The discussion above naturally leads to the question of *how* to choose appropriate values of ϵ that recover meaningful topological features. Rather than selecting a single scale, the theory of persistence considers *all* scales simultaneously by constructing *filtrations*. A filtration is a family $X_\bullet \coloneqq \{X_t : t \in T\}$ of simplicial complexes indexed by a partially ordered set $(T, \leq)$, such that $X_t \subseteq X_s$ whenever $t \leq s$. In this framework, the VR and Čech complexes give rise to the *VR filtration* $\mathcal{VR}_\bullet(S) \coloneqq \{\mathcal{VR}_\epsilon(S) : \epsilon \in [0, \infty)\}$ and the *Čech filtration* $\check{\mathcal{C}}_\bullet(S) \coloneqq \{\check{\mathcal{C}}_\epsilon(S) : \epsilon \in [0, \infty)\}$. As the scale parameter ϵ increases, the conditions for

including simplices become more permissive, ensuring that the complexes grow monotonically. This monotonicity guarantees that the inclusion condition in the definition of a filtration is satisfied. An illustration of a VR filtration over a point cloud in $\mathbb{R}^2$ at four values of the parameter ϵ can be found in **Figure** 1.2.

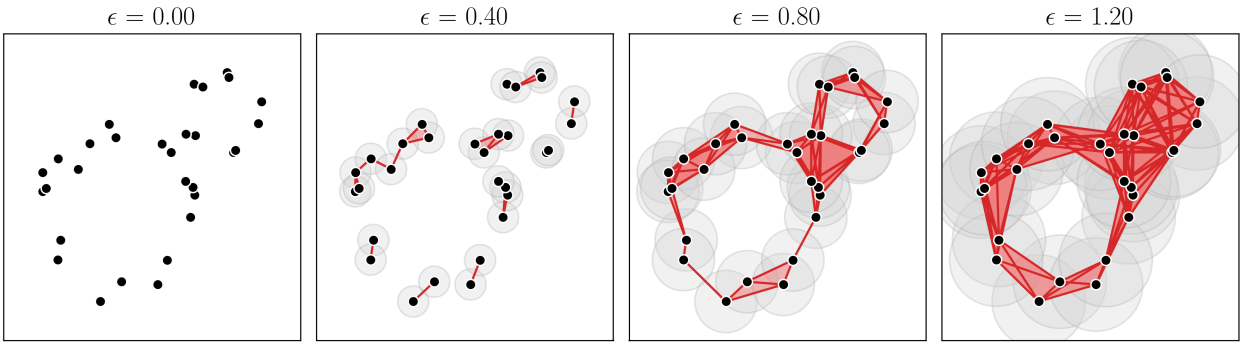


Figure 1.2: VR filtration at four values of the filtration parameter $\epsilon \in [0, \infty)$. VR complexes for $\epsilon = 0$, 0.4, 0.8, and 1.2, with the corresponding balls of radius $r = \epsilon/2$, on a discrete set of 30 points in $\mathbb{R}^2$ sampled over two circles with Gaussian noise added. Note that for $\epsilon = 0.8$ the VR complex captures the two circles (or loops) intuitively present in the sample.

Multifiltrations. Recent applications have motivated the integration of multiple parameters into the construction of filtrations, leading to the concept of *multifiltrations*. This extension arises either from the desire to better handle outliers by incorporating measures such as point density, or from specific application domains where additional parameters naturally play a role in the analysis. We present two extensions of the VR filtration that address these needs by incorporating an extra parameter alongside the scale parameter ϵ.

The first is the *function-Rips bifiltration*, which requires a function $\gamma : S \to \mathbb{R}$ defined on the point cloud. This bifiltration is constructed by taking the VR complexes over the *sublevel sets* of γ, that is $\mathcal{FR}_{t,\epsilon}(S, \gamma) \coloneqq \{\mathcal{VR}_\epsilon(\gamma^{-1}(-\infty, t]) : t \in \mathbb{R}, \epsilon \in [0, \infty)\}$. An analogous construction can be defined using superlevel sets $\gamma^{-1}[t, \infty)$. A common choice for γ in this context is a density estimator on the point cloud, which assigns higher values to points in dense regions and lower values to potential outliers. When such a density function is used, the resulting filtration is often referred to as the *density-Rips bifiltration*, see **Figure** 1.3 for an illustrative example.

A potential drawback of this approach is its reliance on user-defined parameters, such as the number of neighbors in k-Nearest Neighbors (k-NN) estimation, or a bandwidth parameter for Kernel Density Estimation (KDE), which must be carefully tuned. To address this, a parameter-free alternative is provided by the *degree-Rips bifiltration* defined for $t \in \mathbb{N}$ and $\epsilon \in [0, \infty)$ as $\mathcal{DR}_{t,\epsilon}(S) \coloneqq$ the maximal subcomplex of $\mathcal{VR}_\epsilon(S)$ with vertices of degree at least $t-1$, where the degree of a vertex refers to the number of incident edges. This construction filters

simplices based on local connectivity without requiring additional parameter tuning. For further examples and definitions of multifiltrations, we refer the reader to [BL23, Chapter 5].

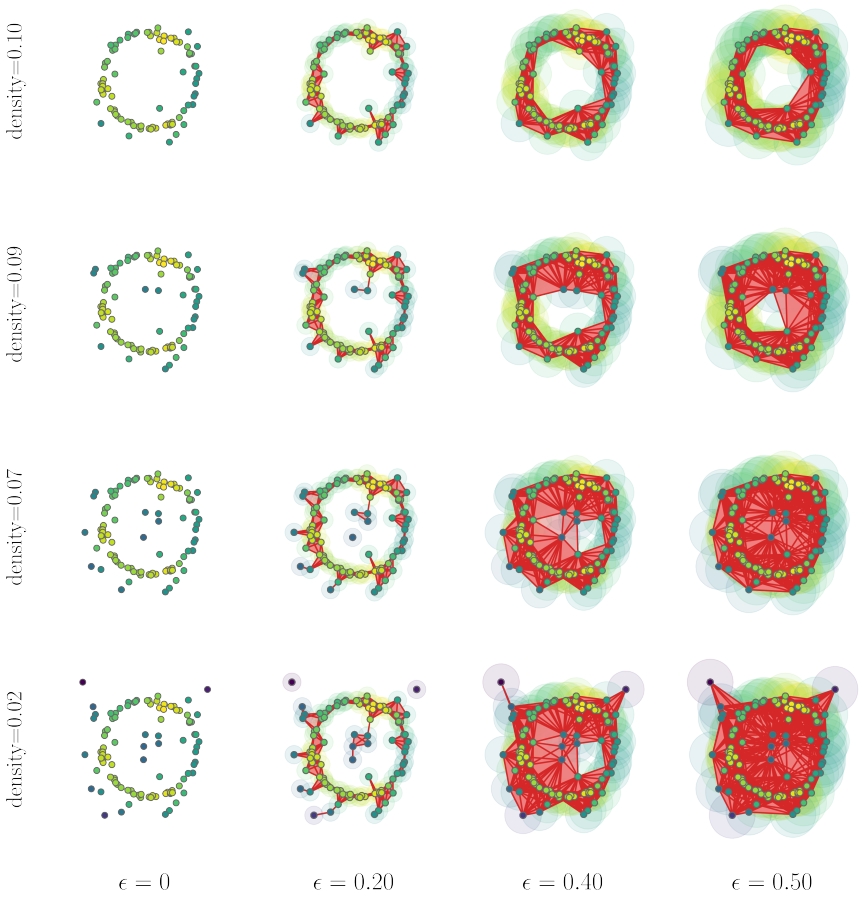


Figure 1.3: Density-Rips bifiltration at 4×4 values of the filtration parameters $\epsilon \in [0, \infty)$ and density $\in \mathbb{R}$. Sample of $N = 100$ points, 60 sampled from circle of radius 1 with Gaussian noise of scale 0.1 added, and 40 outliers sampled from the uniform distribution over the square $[-1.5, 1.5]^2$. The density function is computed using a Gaussian KDE: $\gamma_h(s) = \frac{1}{nh} \sum_{i=1}^{N} K(\frac{s-s_i}{h})$, with bandwidth $h = 0.4$, where K is a Gaussian distribution with zero mean and unit variance. Balls and points are colored by the value of this function. Note that, thanks to filtering out by density, the bifiltration captures the loop from the circle in the upper right region of the parameter space.

A Category-Theoretical Perspective. We close the section with a unifying, category-theoretical definition of filtration that encompasses all the examples above. Given a partially ordered set (also called a poset) $(T, \leq)$, we can view it as a small category T, where the objects are the elements of T, and there is a unique morphism $s \to t$ whenever $s \leq t$.

Definition 1.1.1 (Filtration). A *filtration* is a functor $X_\bullet : \mathsf{T} \to \mathsf{Simp}$, where Simp denotes the category of simplicial complexes. Under the action of this functor, each morphism $s \leq t$ in T corresponds to the inclusion $X_\bullet(s \leq t) : X_s \hookrightarrow X_t$.

For a poset $(T, \leq)$, its opposite poset $(T, \leq_{\text{op}})$ has the same set of objects T but *opposite order*, i.e. $s \leq_{\text{op}} t \iff t \leq s$. We denote the corresponding category as T^{op}.

1.2 Homology, Cohomology, and Persistence Modules

We now introduce the theory of homology and cohomology, both in the absolute and relative settings, and define the central object of study in the theory of persistence: the persistence module.

Simplicial Homology. Given a d-dimensional simplicial complex X, the kth *chain group* is the vector space over a fixed field $\Bbbk$ spanned by the k-simplices of X: $C_k(X, \Bbbk) = \text{span}\{\sigma \in X_k\}$, $0 \leq k \leq d$. When the field is clear from the context, we simply write $C_k(X)$. These chain groups are connected by *boundary operators* $\partial_k : C_k(X) \to C_{k-1}(X)$, defined on basis elements as

$$\partial_k([s_{i_0}, \ldots, s_{i_k}]) \coloneqq \sum_{j=0}^{k} (-1)^j [s_{i_0}, \ldots, \widehat{s_{i_j}}, \ldots, s_{i_k}],$$

where $\widehat{s_{i_j}}$ indicates that the vertex s_{i_j} is omitted. These definitions yield a *chain complex*

$$0 \to C_d(X) \xrightarrow{\partial_d} C_{d-1}(X) \to \cdots \to C_1(X) \xrightarrow{\partial_1} C_0(X) \to 0,$$

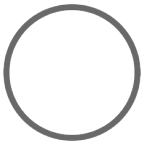
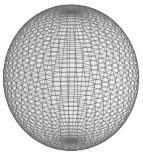
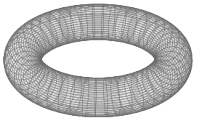
with the fundamental property $\partial_k \circ \partial_{k+1} = 0$. That is, $\text{im}(\partial_{k+1}) \subseteq \ker(\partial_k)$. Elements of $\ker(\partial_k)$ are called *cycles*, and those in $\text{im}(\partial_{k+1})$, *boundaries*.

Definition 1.2.1 (Homology groups). Let $k \geq 0$ and X be a simplicial complex. The kth *homology group* is defined as the quotient

$$\text{H}_k(X) \coloneqq \frac{\ker(\partial_k)}{\text{im}(\partial_{k+1})}. \tag{1.2.1}$$

This group represents *cycles modulo boundaries*, which encode the topological features of X: connected components ($k = 0$), loops or tunnels ($k = 1$), cavities or bubbles ($k = 2$), and higher-dimensional analogues for $k > 2$. The dimensions of these vector spaces, called the *Betti numbers*, $\beta_k(X) \coloneqq \dim \text{H}_k(X)$ precisely count the number of these topological features. Although here we have introduced the theory of *simplicial* homology, these notions can be defined for general topological spaces using the framework of *singular* homology, and it turns out that both theories agree in triangulable spaces. Some illustrative examples of homology groups of common manifolds can be found in **Table** 1.1.

Table 1.1: Homology for $k = 0, 1$ and 2 for the circle, the 2-sphere and the torus. Each manifold has a single connected component, reflected in their identical 0-homology. The circle, being 1-dimensional, has one non-trivial loop and no 2-dimensional cavities. The 2-sphere has no non-trivial loops, as all loops can be contracted to a point, resulting in trivial 1-homology. However, it does enclose a 2-dimensional cavity. The torus has two independent, non-contractible loops (latitudinal and longitudinal) and also encloses a cavity, as shown in its homology.

	Circle ($\mathbb{S}^1$)	Sphere ($\mathbb{S}^2$)	Torus ($\mathbb{T}$)
			
0-th homology	$\mathrm{H}_0(\mathbb{S}^1) = \mathbb{R}$ $\beta_0(\mathbb{S}^1) = 1$	$\mathrm{H}_0(\mathbb{S}^2) = \mathbb{R}$ $\beta_0(\mathbb{S}^2) = 1$	$\mathrm{H}_0(\mathbb{T}) = \mathbb{R}$ $\beta_0(\mathbb{T}) = 1$
1-st homology	$\mathrm{H}_1(\mathbb{S}^1) = \mathbb{R}$ $\beta_1(\mathbb{S}^1) = 1$	$\mathrm{H}_1(\mathbb{S}^2) = 0$ $\beta_1(\mathbb{S}^2) = 0$	$\mathrm{H}_1(\mathbb{T}) = \mathbb{R}^2$ $\beta_1(\mathbb{T}) = 2$
2-nd homology	$\mathrm{H}_2(\mathbb{S}^1) = 0$ $\beta_2(\mathbb{S}^1) = 0$	$\mathrm{H}_2(\mathbb{S}^2) = \mathbb{R}$ $\beta_2(\mathbb{S}^2) = 1$	$\mathrm{H}_2(\mathbb{T}) = \mathbb{R}$ $\beta_2(\mathbb{T}) = 1$

Simplicial Cohomology. The previous construction can be dualized to define *cochain groups* $C^k(X, \Bbbk) = \mathrm{Hom}_{\Bbbk}(C_k(X), \Bbbk)$, where again, when clear from the context, we just write $C^k(X)$. We analogously have *coboundary maps* $\delta^k : C^{k-1}(X) \to C^k(X)$, defined as $\delta^k f(\sigma) \coloneqq f(\partial_k \sigma)$, forming a *cochain complex*

$$0 \leftarrow C^d(X) \xleftarrow{\delta^d} C^{d-1}(X) \leftarrow \cdots \leftarrow C^1(X) \xleftarrow{\delta^1} C^0(X) \leftarrow 0$$

where $\delta^{k+1} \circ \delta^k = 0$.

Definition 1.2.2 (Cohomology groups). Let $k \geq 0$ and X be a simplicial complex. The kth *cohomology group* is defined as the quotient

$$\mathrm{H}^k(X) \coloneqq \frac{\ker \delta^{k+1}}{\mathrm{im}\, \delta^k}. \tag{1.2.2}$$

In our context, working over a field $\Bbbk$, the spaces $\mathrm{H}^k(X)$ and $\mathrm{H}_k(X)$ turn out to be dual vector spaces thanks to the Universal Coefficient Theorem [Hat02, Theorem 3.2]. If they are finite-dimensional, we additionally know that they must be isomorphic.

Relative Homology and Cohomology. Another construction that will be important in what follows is that of *relative* homology (and cohomology). The idea is that, if we have a pair of simplicial complexes $A \subset X$, we can define the *relative chain group* $C_k(X, A) \coloneqq \frac{C_k(X)}{C_k(A)}$. Since the boundary operator in $C_k(X)$ restricted to $C_k(A)$ lands in $C_{k-1}(A)$, it induces a *relative* boundary operator $\partial_k^{A,X} : C_k(X, A) \to C_{k-1}(X, A)$, satisfying $\partial_{k-1}^{A,X} \circ \partial_k^{A,X} = 0$.

Definition 1.2.3. Let $k \geq 0$ and $A \subset X$ be simplicial complexes. The kth-*relative homology* group is defined as

$$\mathrm{H}_k(X, A) \coloneqq \frac{\ker \partial_k^{A\,X}}{\operatorname{im} \partial_{k-1}^{A,X}}.$$

It represents the *cycles in* X up to *boundaries in* A. The construction for cohomology is analogous, obtained by dualizing the process above.

More Cathegory Theory in Persistence. The homology and cohomology constructions above are *functorial*, defining functors $\mathrm{H}_k(-)$, $\mathrm{H}^k(-) : \mathsf{Simp} \to \mathsf{Vec}_{\Bbbk}$ from the category of simplicial complexes to the category of vector spaces over $\Bbbk$. Functoriality ensures that maps between complexes induce linear maps between the corresponding homology or cohomology groups, preserving the identity and preserving (for homology) or reversing (for cohomology) the composition of functions. Applying homology to a filtration of simplicial complexes $X_\bullet = \{X_t : t \in T\}$, gives rise to a *persistence module*: a family of vector spaces $\{\mathrm{H}_k(X_t) : t \in T\}$ with linear maps $\mathrm{H}_k(s \leq t) : \mathrm{H}_k(X_s) \to \mathrm{H}_k(X_t)$, for $s \leq t$ in T, induced by the inclusions $X_t \subseteq X_s$. These linear maps are typically called *internal* or *transition maps*.

Definition 1.2.4 (**Persistence Module**)**.** A persistence module is as a functor $M : \mathsf{T} \to \mathsf{Vec}_{\Bbbk}$ from the small category defined by the indexing poset $(T, \leq)$ to the category of vector spaces over $\Bbbk$. If we restrict to the category of finite-dimensional vector spaces, $\mathsf{vec}_{\Bbbk}$, we obtain *point-wise finite-dimensional* (p.f.d.) persistence modules.

We drop the mention of the field, and simply write Vec or vec when it is clear from the context. We will usually denote the vector spaces in the module as $M_t = M(t)$, and the transition maps as $M_t^s = M(s \leq t)$ for $s, t \in T$, $s \leq t$. If the indexing set $(T, \leq)$ is totally ordered, meaning that any two elements in T are comparable, as it happens when T is a subset of the real line $(\mathbb{R}, \leq)$, the result is a *single-parameter persistence* module. In contrast, modules indexed by posets with incomparable elements, such as $(\mathbb{R}^n, \leq)$, where $\leq$ is the product order induced by the total order in $\mathbb{R}$, are called *multiparameter persistence* modules.

Definition 1.2.5 (**Tame persistence module**, Definition 1.12. [Oud15])**.** A persistence module $M : \mathsf{T} \to \mathsf{Vec}_{\Bbbk}$ is *tame* (also called *q-tame*) if for any $s < t \in T$ (i.e., $s \leq t$ and $s \neq t$) we have $\operatorname{rank} M_t^s < \infty$.

Note that every p.f.d. module is also tame, but the opposite might not hold, as in a tame module M we are allowing the rank of $M_t^t = \mathrm{id}_{M_t}$ to be infinite for any $t \in T$.

Persistence modules form an *abelian category* denoted by $\mathsf{Vec}^{\mathsf{T}}$, where morphisms between M, $N \in \mathsf{Vec}^{\mathsf{T}}$ are given by *natural transformations* $\varphi : M \to N$, consisting of a family of linear maps $\varphi_t : M_t \to N_t$ which are required to commute with the internal maps: for $s \leq t$, they yield the commutative diagram

$$\begin{array}{ccc} M_s & \xrightarrow{M(s\leq t)} & M_t \\ {\scriptstyle\varphi_s}\downarrow & & \downarrow{\scriptstyle\varphi_t} \\ N_s & \xrightarrow{N(s\leq t)} & N_t. \end{array}$$

Kernels, cokernels, direct sums and products in $\mathsf{Vec}^{\mathsf{T}}$ are defined point-wise, inherited from the corresponding constructions in Vec.

Given a poset $(T, \leq)$, certain distinguished subposets play a central role in the definition of persistence modules. For $r, t \in T$ with $r \leq t$, the *closed segment* (or *rectangle*) between r and t is defined as

$$[r, t] \coloneqq \{\, s \in T : r \leq s \leq t \,\}.$$

If $r < t$ (i.e., $r \leq t$ and $r \neq t$), the *open rectangle* is defined as

$$(r, t) \coloneqq \{\, s \in T : r < s < t \,\}.$$

Similarly, one defines the *half-open rectangles* $[r, t)$ and $(r, t]$. In discrete posets, it is enough to consider closed rectangles; in contrast, for continuous posets such as $\mathbb{R}^n$, the distinction between open, closed, and half-open rectangles becomes non-trivial and important.

We also define the *upset* from some $t \in T$ as

$$[t, \infty) \coloneqq \{s \in T : s \geq t\}$$

and the *downset* as

$$(-\infty, t] \coloneqq \{s \in T : s \leq t\}.$$

Given some vector space U and $r \leq t \in T$, we denote by $U_{[r,t]}$ the persistence module such that

$$\left(U_{[r,t]}\right)_s = \begin{cases} U, & \text{if } r \leq s \leq t, \\ 0, & \text{otherwise;} \end{cases} \tag{1.2.3}$$

and with internal maps being the identity between any two points in the closed rectangle, and the zero map otherwise. We can analogously define the modules $U_{(r,t)}$, $U_{[t,\infty)}$ and $U_{(-\infty,t]}$

supported in the open segment, the upset and the downset, respectively.

1.3 Barcodes

In this section, we introduce the most established invariant of the PH pipeline: the persistence barcode. First, we recall the mathematical result that sustains it, the *Structure Theorem*. We then say a few words about computations of barcodes, and highlight some of its limitations.

The Structure Theorem and the Persistence Barcode. A persistence module $M : \mathsf{T} \to \mathsf{Vec}$ is *indecomposable* if every direct sum decomposition $M = A \oplus B$ with $A, B \in \mathsf{Vec}^{\mathsf{T}}$ implies $A = 0$ or $B = 0$. The *Structure Theorem* [see Theorem 4.2 in BL23, for a more precise statement] guarantees that any p.f.d. module $M \in \mathsf{vec}^{\mathsf{T}}$ admits a decomposition into indecomposable summands

$$M \cong \bigoplus_{i \in \mathcal{I}} M_i$$

and that this decomposition is essentially unique. For finite posets, this result is standard in the theory of representations of finitely presented algebras. In the generality here presented, the *existence* of the decomposition follows from work of Crawley-Boevey [Cra94], and a short proof was given by Botnan and Crawley-Boevey in [BC20]; the *uniqueness* of the decomposition follows from the Krull–Remak–Schmidt–Azumaya theorem [Azu50]. In more detail, Azumaya proved that any two decompositions of a module in indecomposable modules whose endomorphism ring is *completely primary* must be the same up to reordering. A ring is completely primary if the sum of two non-regular elements is always non-regular, or in other words, if the set of all its zero-divisors forms an additive group. We will use this condition in the proof of **Lemma** 2.6.2.

Let $(T, \leq)$ be a poset. An *interval* $I \subset T$ is a convex and connected subposet, that is, for any three elements $r \leq s \leq t$ of T if $r, t \in I$ then $s \in I$ (*convexity*) and for any $s, t \in I$ there exists a chain of pairwise comparable elements connecting them (*connectivity*). As expected, when $T = \mathbb{R}$ with the total order, these are precisely the intervals of the real line, which coincide with the closed, open and half-open segments defined earlier.

Given a convex subposet $I \subset T$ an *indicator module* supported on I is a persistence module $\Bbbk_I : \mathsf{T} \to \mathsf{Vec}$ such that $\Bbbk_I(t) = \Bbbk$ if $t \in I$ and zero otherwise, the transition maps being the identity within I and zero otherwise. The convexity condition ensures that composition of internal maps is well-defined. When I is an interval, we call $\Bbbk_I$ an *interval module*. These modules are indecomposable, and when $T = \mathbb{R}$ with the total order, the Structure Theorem strengthens to assert that every p.f.d. persistence module decomposes (essentially uniquely)

into a direct sum of interval modules [Cra15, Theorem 1.1]. This result justifies the following definition.

Definition 1.3.1 (**Persistence barcode).** The *persistence barcode* of a module $M \in \mathsf{vec}^{\mathbb{R}}$, denoted $\mathrm{PB}(M)$, is given by the multiset of intervals appearing in the decomposition from the Structure Theorem.

Remark 1.3.2 (**Undecorated barcodes and tame persistence modules)**. The Structure Theorem is established for p.f.d. modules. In contrast, a q-tame module (**Definition** 1.2.5) need not be interval decomposable. Nevertheless, one can still associate barcodes to such modules by adopting a coarser notion: *undecorated barcodes* [Cha+16, §1.6], where we only record the endpoints of the intervals and ignore whether they are closed, open or half-open. We follow here the constructions in [BL15, §7.1] and [Oud15, Chapter 1, §2] to define these undecorated barcodes for q-tame modules, equivalent to an alternative approach in [Cha+16]. Although a q-tame module $M \in \mathsf{Vec}^{\mathbb{R}}$ might not be interval decomposable, we can define its *radical* submodule $\mathrm{rad}(M)$ as $\mathrm{rad}(M)\,(t) \coloneqq \sum_{s \leq t} \mathrm{im}\, M_t^s$, which is not p.f.d. in general but is interval decomposable [CCd16, Corollary 3.6]. In addition, it makes the quotient module $N \coloneqq M/\mathrm{rad}(M)$ ephemeral, meaning that for all $s < t$, we have $\mathrm{rank}\, N_s^t = 0$. This means that every q-tame module is interval decomposable "modulo" some ephemeral modules. Following that intuition, the barcode of a q-tame module is defined as the undecorated barcode of its radical, which coincides with the undecorated barcode of M in case that M is also interval decomposable [BL15, Proposition 7.2]. Chazal, Crawley-Boevey and de Silva [CCd16] formalized this approach considering the *observable category* of persistence modules, given by the q-tame modules modulo ephemeral modules, in the sense of Serre's theory of localization. In this category, when the modules are parameterized over the real line, the interval modules $\Bbbk_{[s,t]}$, $\Bbbk_{(s,t)}$, $\Bbbk_{[s,t)}$ and $\Bbbk_{(s,t]}$ defined over different intervals with same end points turn out to be isomorphic, every q-tame module is indecomposable, and the undecorated barcode is a complete invariant. Following the typical convention in TDA, in this work the term *persistence barcode* of a persistence module $M \in \mathsf{Vec}^{\mathbb{R}}$ refers to its undecorated barcode, which is typically written using half-open intervals:

$$\mathrm{PB}(M) \coloneqq \left\{ [b_i, d_i) : i \in \mathcal{I},\ M \cong \oplus_{i \in \mathcal{I}} \Bbbk_{[b_i, d_i)} \right\}. \tag{1.3.1}$$

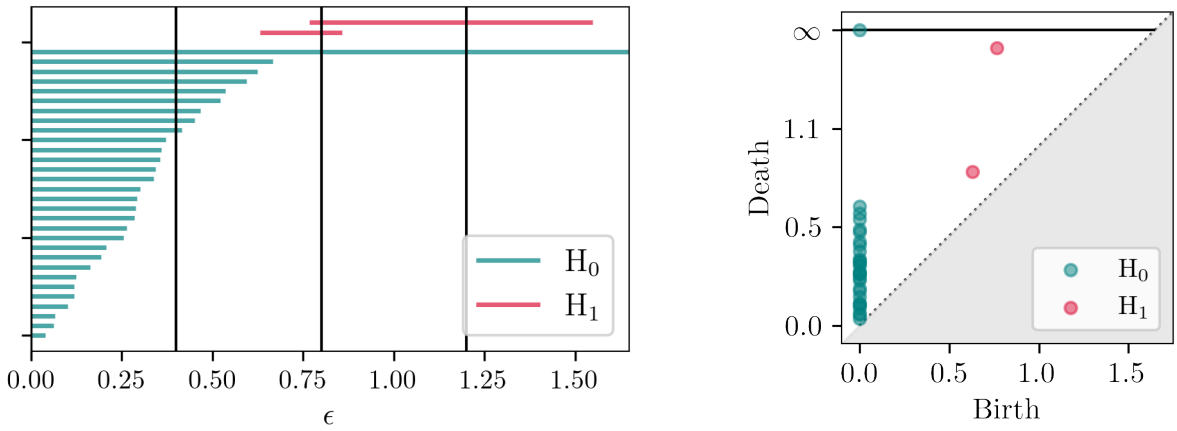
When $M = \mathrm{H}_k(X_\bullet)$ for some filtration $X_\bullet : \mathbb{R} \to \mathsf{Simp}$ and some $k \geq 0$, each interval in the barcode in **Equation** (1.3.1) corresponds to a topological feature in the filtration, which is *born* (appears) at b_i and *dies* (diappears) at d_i, *persisting* over the interval $[b_i, d_i)$. The length $p_i = d_i - b_i$ is called its *persistence*, and long bars are typically interpreted as meaningful topological features of the data, while short bars are often attributed to noise. The term "persistence" is often also use to refer to the field that studies the properties of persistence modules and PH. Our intended meaning when using this word in this thesis (whether referring

to the length of a bar or to the field itself) should be clear from the context.

There is an alternative, equivalent representation of barcodes, which we introduce next.

Definition 1.3.3 (Persistence diagram). The *persistence diagram* of a module $M \in \mathsf{vec}^{\mathbb{R}}$, denoted $\mathrm{PD}(M)$, is defined as the collection of points in the region of $\mathbb{R}^2$ above the diagonal $\Omega = \{(x, y) \in \mathbb{R}^2 : x \leq y\}$, where each point $(b, d) \in \mathrm{PD}(M)$ corresponds to a bar $[b, d) \in \mathrm{PB}(M)$ in the barcode, counted with its multiplicity of appearance, and all the points in the diagonal $\partial\Omega = \{(x, y) \in \mathbb{R}^2 : x = y\}$ are considered with infinite multiplicity.

An illustrative example of a barcode and corresponding diagram, computed from the VR filtration in **Figure** 1.2, can be found in **Figure** 1.4.



(a) Persistence Barcode **(b)** Persistence Diagram

Figure 1.4: Persistence Barcode and Persistence Diagram of the VR filtration over the sample S of points in Figure 1.2. Recall that each bar $[b, d)$ in the barcode corresponds to a point (b, d) in the diagram. $\mathrm{H}_0(\mathcal{VR}(S))$ in blue and $\mathrm{H}_1(\mathcal{VR}(S))$ in red. The vertical lines in the barcode mark the filtration values $\epsilon = 0.4, 0.8, 1.2$, corresponding to the stages shown in **Figure** 1.2. Note that there is one long H_0 bar, corresponding to the connected component that persists throughout the filtration; and two persistent H_1 bars, corresponding to the two circles that we appreciate in the point cloud. At $\epsilon = 0.4$, only H_0-bars appear, indicating disconnected components with no loops. At $\epsilon = 0.8$, the long H_0-bar remains, and two H_1-bar emerge from the loops formed by the circles in the point cloud. By $\epsilon = 1.2$, the smaller of these loops is filled in by triangles, leaving one H_0-bar and one H_1-bar corresponding to the bigger circle in the filtration.

Computing Barcodes. The most basic algorithm for computing the persistence barcode of a finite filtration $\{X_i : 1 \leq i \leq N\}$ was first introduced for coefficients in $\Bbbk = \mathbb{F}_2$ in [ELZ02] and later generalized to arbitrary fields in [ZC05]. The procedure is as follows.

1. We first define the *boundary matrix*, which is the matrix representation of the boundary map $\partial : C(X_N) \to C(X_N)$, where $C(X_N) \coloneqq \bigoplus_{0 \leq k \leq \dim(X_N)} C_k(X_N)$, with respect to

the standard basis given by the simplices of X_N. In this matrix, rows and columns correspond to simplices, ordered first by dimension and then by their appearance in the filtration.

2. For each column, we define its *pivot* as the index of the lowest (in terms of height) non-zero element in the column.

3. Next, we perform column-wise Gaussian elimination from left to right in the matrix. For each column, we check if the row corresponding to its pivot only has zeros to the left of the selected column. If that is not the case, we perform column operations with the non-zero columns to the left, decreasing the value of our pivot, and check again. We repeat this process until we satisfy the initial condition.

4. This algorithm yields the *reduced matrix* from which the barcode can be directly read.

This naive algorithm poses computational problems, mainly, that its complexity equals the number of simplices in the filtration, and that storing the boundary matrix requires significant memory. There has been extensive research to address these issues and optimize persistent homology computations to make them widely accessible and applicable to real-world data. Here we focus on the *clearing algorithm* by Chen and Kerber [CK11], which is key for the implementation of `Ripser` [Bau21], the state-of-the-art code for barcode computations and key ingredient of one of the projects o this thesis. The key observation brought forward by Chen and Kerber [CK11] is that some columns in the reduced matrix must be null after the reduction and do not play a role in the Gaussian elimination process. The clearing algorithm reduces the boundary matrix in blocks by dimension from right to left so that it becomes possible to detect beforehand these null columns and set them directly to 0. Consequently, it is possible to avoid reducing these columns and accelerate the computation. This is already an improvement compared to the standard reduction algorithm. However, this improvement is burdened by the large number of columns in the first block that must be reduced in the boundary matrix. One of the main contributions of `Ripser` [Bau21] is noting that the real increase in speed appears when considering the *relative* coboundary matrix in which this first block is significantly smaller.

The fact that computing cohomology instead of homology induces a speed-up in computations was already foreshadowed by de Silva *et al.* [SMV11], who established that the persistence barcodes of the *four standard persistence modules* of a filtration $X_\bullet = \{X_i : 1 \leq i \leq N\}$, defined by their absolute homology and cohomology, and relative homology and cohomology

as

$$\mathrm{H}_k(X_\bullet): \quad \mathrm{H}_k(X_1) \to \ldots \to \mathrm{H}_k(X_{N-1}) \to \mathrm{H}_k(X_N), \tag{1.3.2}$$

$$\mathrm{H}^k(X_\bullet): \quad \mathrm{H}^k(X_1) \leftarrow \ldots \leftarrow \mathrm{H}^k(X_{N-1}) \leftarrow \mathrm{H}^k(X_N), \tag{1.3.3}$$

$$\mathrm{H}_k(X_N, X_\bullet): \ \mathrm{H}_k(X_N) \to \mathrm{H}_k(X_N, X_1) \to \ldots \to \mathrm{H}_k(X_N, X_{N-1}), \tag{1.3.4}$$

$$\mathrm{H}^k(X_N, X_\bullet): \ \mathrm{H}^k(X_N) \leftarrow \mathrm{H}^k(X_N, X_1) \leftarrow \ldots \leftarrow \mathrm{H}^k(X_N, X_{N-1}); \tag{1.3.5}$$

had equivalent barcodes. In more detail, de Silva *et al.* [SMV11] proved the following.

Proposition 1.3.4 (Proposition 2.3 and Proposition 2.4, [SMV11])**.** *Let $X_\bullet = \{X_i : 1 \leq i \leq N\}$ be a filtration of finite-dimensional simplicial complexes. For all $k \geq 0$, we have*

$$\mathrm{PB}(\mathrm{H}_k(X_\bullet)) = \mathrm{PB}(\mathrm{H}^k(X_\bullet)).$$

In addition, if $\mathrm{PB}(-)^\dagger$ denotes the finite intervals in the barcode and $\mathrm{PB}(-)^\infty$ the infinite intervals, we also have

$$\mathrm{PB}(\mathrm{H}_k(X_\bullet))^\dagger = \mathrm{PB}(\mathrm{H}_{k+1}(X_N, X_\bullet))^\dagger, \quad \textit{and} \quad \mathrm{PB}(\mathrm{H}_k(X_\bullet))^\infty = \mathrm{PB}(\mathrm{H}_k(X_N, X_\bullet))^\infty;$$

where the second equality must be understood as a bijection $[a, \infty) \leftrightarrow [-\infty, a)$.

Limitations of Barcodes. Barcodes have been the most widely used, and thus successful, topological invariant stemming from the PH pipeline. However, they also present certain drawbacks and limitations that have inspired some of the work in this thesis. First, although the space of barcodes equipped with a suitable metric (see **Section** 1.5) is complete and separable, which makes it amenable to probability and statistics [MMH11], its underlying geometry is complicated. Specifically, it constitutes an Alexandrov space with curvature bounded from below, where Fréchet means are not unique [Tur+14]. This complexity presents significant challenges for the full development of statistical theory for barcodes and for their integration in Machine Learning (ML) methodologies.

In the literature, two main strategies have emerged to address this issue. One approach adapts and extends statistical and ML methods to work directly within the space of barcodes despite its geometric complications, e.g. [Fas+14; Rei+15; Hof+17]. The other approach involves embedding barcodes into spaces with more favorable statistical properties, such as Euclidean or Hilbert spaces. This process is known as *vectorization*. In **Section** 1.4 we introduce in detail one such vectorization, called the *persistence landscape* [Bub15], and provide a brief overview and references for other methods.

A second major limitation of barcodes is that they do not naturally extend to the theory of multiparameter persistence [CZ09]. This difficulty stems from the representation theory of

modules over arbitrary posets, where the indecomposables need not be interval modules, or even indicator modules. As a result, research in multiparameter persistence in recent years has focused on developing new invariants that are sufficiently discriminative, computationally efficient, and suitable for data analysis. We discuss some of these developments in **Section** 1.4.

1.4 Beyond Barcodes: Further Invariants

In this section we introduce two functional invariants that are of great importance in some sections of this thesis: the rank function (or rank invariant in multiparameter persistence) and the persistence landscape. As functional summaries, they live in a vector space amenable to statistical analysis. We conclude the section with a brief overview of other relevant invariants not explored in depth in this thesis but important to the theory of persistence.

Rank Functions. Given that a persistence module consists of a family of vector spaces connected by linear maps, a natural quantity to consider to define invariants is the rank of these transition maps.

Definition 1.4.1 (Rank function). For a persistence module $M \in \mathsf{Vec}^{\mathbb{R}}$, the *rank function* is defined as the map $\mathrm{rk}^M : \mathbb{R} \times \mathbb{R} \to \mathbb{Z}$ where

$$\mathrm{rk}^M(x, y) \coloneqq \begin{cases} \operatorname{rank} M(x \leq y) & \text{if } x \leq y, \\ 0 & \text{otherwise.} \end{cases} \tag{1.4.1}$$

This definition naturally extends to multiparameter persistence modules $M \in \mathsf{Vec}^{\mathbb{R}^n}$ by considering the ranks of maps $M(\mathbf{x} \leq \mathbf{y})$ between comparable pairs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. In this context, rank functions are typically called *rank invariants*, and were indeed the topological invariant first proposed to study multiparameter persistence when barcodes were proved to be unfruitful [CZ09]. When the persistence module is a PH module $\mathrm{H}_k(X_\bullet)$ of a filtration $X_\bullet \in \mathsf{Simp}^{\mathbb{R}^n}$, these ranks are often denoted as $\beta_k^{\mathbf{x},\mathbf{y}} = \operatorname{rank} \mathrm{H}_k(\mathbf{x} \leq \mathbf{y})$ and called *persistent Betti numbers* [ELZ02], drawing connections to the definition of Betti numbers in the theory of homology. These should not be confused with *multigraded Betti numbers*, which we will introduce in **Section** 2.6.2, and which are also relevant invariants in multiparameter persistence.

PH is rooted in various constructions, but potentially the most relevant one is *size theory*, which in fact predates the definition of PH modules and barcodes. A primitive version of the rank function for H_0 was first introduced by Frosini in the 1990s as the *size function*, a tool for computer vision and pattern recognition in shape analysis [Fro90; Fro92b; Fro92a; Ver+93]. Size functions were algebraically reinterpreted as formal series by Landi and Frosini

[LF97; LF99; FL99], using a construction based their discontinuities which reappeared in the definition of persistence diagrams introduced in the 2000s [CEH05; Cha+09]. This definition was not based on the Structure Theorem, but on using an inclusion-exclusion formula involving the persistent Betti numbers. Instead, here we have presented persistence diagrams directly through their equivalence to barcodes. Several pseudometrics were also introduced for size functions [LF97], which later found counterparts in metrics for persistence barcodes and diagrams. In **Section** 1.5, we discuss some potential reasons why barcodes and diagrams prevailed over size and rank functions, and the gaps addressed in this thesis in this direction.

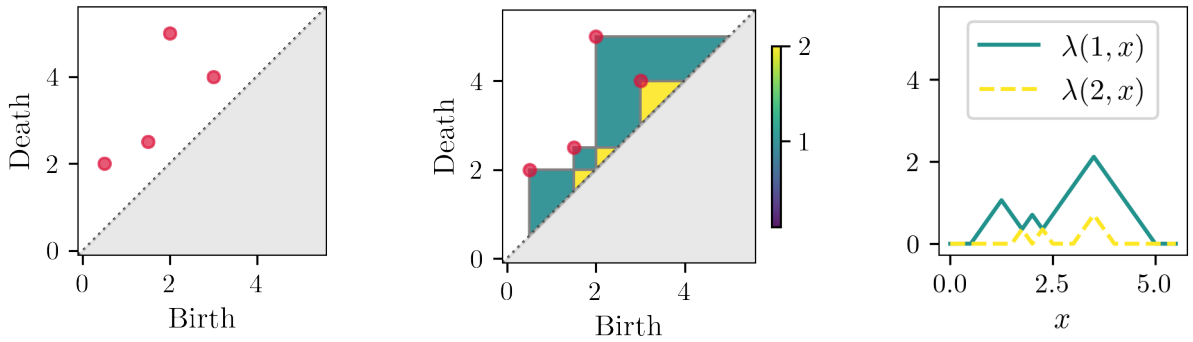
Persistence Landscapes. In single-parameter persistence, rank functions are discontinuous integer-valued functions over $\mathbb{R}^2$. As functions, they live in a vector space which can be endowed with the $L^p(\mathbb{R}^2)$ norm for $1 \leq p \leq \infty$, providing for $p = 2$ a Hilbert space structure (cf. **Equation** (1.5.6)). This is already an improvement to develop statistical theory and ML methods with respect o barcodes, whose geometry was far more complicated and impossibilitated direct integration with ML. *Persistence landscapes* are piecewise-linear functions built from the rank functions that further simplify them, reducing the dimensionality and making them more manageable, while preserving their functional nature to develop statistical theory. Persistence landscapes in single-parameter persistence were first introduced by Bubenik [Bub15] and then extended to multiparameter persistence by Vipond [Vip20]; here we directly introduce the latter definition, which encompasses the former.

Definition 1.4.2 (**Multiparameter persistence landscapes).** For a persistence module $M \in \mathsf{Vec}^{\mathbb{R}^n}$, its *persistence landscape* is the function $\lambda : \mathbb{N} \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\lambda(m, \mathbf{x}) \coloneqq \sup\{\epsilon > 0 : \beta^{\mathbf{x}-\mathbf{h}}_{\mathbf{x}+\mathbf{h}} \geq m \text{ for all } \mathbf{h} \geq \mathbf{0} \text{ with } \|\mathbf{h}\|_\infty \leq \epsilon\}. \tag{1.4.2}$$

Intuitively, the persistence landscape measures the maximal radius (in the ℓ^∞ metric of $\mathbb{R}^n$) over which m features in the module persist in all directions at $\mathbf{x} \in \mathbb{R}^n$. This gives a family of non-negative functions $\lambda(m, \mathbf{x}) \geq 0$ which are 1-Lipschitz in $\mathbf{x} \in \mathbb{R}^n$ [Vip20, Lemma 20]. See **Figure** 1.5 for an illustrative example of a persistence diagram 1.5a, and its corresponding rank function 1.5b and persistence landscape 1.5c .

Other Invariants. There are additional vectorization techniques alternative to the persistence landscape for single parameter persistence such as the *persistent entropy* [Chi+15; Ruc+16; AGS20], *Betti curves*, *silhouettes* [Cha+14b], *persistence kernels* [Rei+15] or *persistence images* [Ada+17], to name a few. We refer to [Ali+23] for a survey and benchmarking of methods. On the other hand, the lack of a complete invariant analogue to the persistence barcode in the setting of multiparameter persistence has motivated extensive research to define new incomplete invariants that are interpretable and computationally feasible for real



(a) Persistence diagram **(b)** Rank function **(c)** Persistence landscape

Figure 1.5: Rank function (b) and persistence landscape (c) for a given persistence diagram (a). Observe that the landscapes $\lambda(1, x)$ and $\lambda(2, x)$ can be obtained from the rank function by projecting the diagonal onto the x-axis and tracing the outline of the outermost boundaries of the collection of triangles that define the rank function, and the second outermost boundaries, respectively. See [Bub15, §2.3] for more details.

data. We now provide a non-comprehensive list of methods. Some proposals capitalize the algebraic structure of persistence modules, which can be seen as modules over a certain algebra, to adapt and use the invariants in that setting. This includes *minimal presentations* and *multigraded Betti numbers* [LW15; LW22], or *multigraded associated primes* and *local cohomology* [Har+19], for instance. An alternative due to Patel [Pat18] leverages the original definition of persistence diagrams using an inclusion-exclusion formula on the rank function [CEH05; Cha+09] to define *generalized persistence diagrams*. Kim and Mémoli [KM21] introduced the *generalized rank invariant* and showed that this is the counterpart to the generalized persistence diagram of Patel, extending and generalizing the previous formulation. Using resolutions and rank-exact structures Botnan *et al.* [BOO22] have introduced *signed barcodes* and *signed rank decompositions*, which have been vectorized seeing them as measures for seamless integration within ML methods by Loiseaux *et al.* [Loi+24]. Finally, similarly to persistence landscapes, other vectorization methods can be extended to multiparameter persistence, such as the *multiparameter persistence kernel* [Cor+19] or the *multiparameter persistence image* [CB20].

1.5 Metrics and Stability

Stability is a fundamental concept in TDA, supporting the idea that topological invariants—such as barcodes, rank functions, and persistence landscapes—provide faithful representations of input data. Intuitively, stability ensures that small changes in the input lead to only small changes in the resulting invariants, making them robust to noise. This notion is formalized mathematically through bounds that relate distances between input data to distances between

their corresponding invariants. In this section, we review key metrics and stability results in the theory of persistence. Later, in **Chapter** 3, we will present new stability bounds for rank functions established as part of this thesis.

Barcodes and Persistence Diagrams. We first introduce arguably the most relevant metric in the space of persistence barcodes or diagrams.

Definition 1.5.1 (**Bottleneck distance).** The *bottleneck distance* between two persistence diagrams PD_1 and PD_2 is defined as

$$d_B(\mathrm{PD}_1, \mathrm{PD}_2) \coloneqq \inf_{\phi:\mathrm{PD}_1\to\mathrm{PD}_2} \sup_{x\in\mathrm{PD}_1} \|x - \phi(x)\|_\infty \tag{1.5.1}$$

where $\|\cdot\|_\infty$ is the ℓ^∞ norm on $\mathbb{R}^2$ and ϕ ranges over all bijections between PD_1 and PD_2 (recall that all points in the diagonal are included in the persistence diagram with infinite multiplicity, so we can map points in PD_1 to the diagonal in PD_2).

This metric is relevant in persistence because it underpinned the first stability result in the literature [CEH05; CEH07]. In their work, Cohen-Steiner *et al.* considered continuous *tame* functions $f : X \to \mathbb{R}$ defined on triangulable spaces.[1] They showed that the map sending each of function f to the corresponding persistence diagram $\mathrm{PD}(f)$ arising from the PH of the *sublevel-set filtration* $X_t = f^{-1}(-\infty, t]$, for $t \in \mathbb{R}$, is 1-Lipschitz with respect to the bottleneck distance for diagrams and the L^∞ metric for functions. That is, if f and g are tame functions, the inequality $d_B(\mathrm{PD}(f), \mathrm{PD}(g)) \leq \|f - g\|_\infty$ holds.

The bottleneck distance can be generalized to a broader family of metrics known as Wasserstein distances by replacing the ℓ^∞ norm and the supremum in its definition.

Definition 1.5.2 (**(p, q)-Wasserstein distance).** For two persistence diagrams PD_1 and PD_2 and parameters $1 \leq p, q < \infty$, the (p, q)-*Wasserstein distance* is defined as

$$W_{p,q}(\mathrm{PD}_1, \mathrm{PD}_2) \coloneqq \inf_{\phi:\mathrm{PD}_1\to\mathrm{PD}_2} \left(\sum_{x\in\mathrm{PD}_1} \|x - \phi(x)\|_q^p \right)^{1/p} \tag{1.5.2}$$

where $\|\cdot\|_q$ denotes the ℓ^q norm on $\mathbb{R}^2$ and again ϕ ranges over all bijections between PD_1 and PD_2.

In practice, it is common to assume $p = q$ and refer simply to the p-*Wasserstein distance*, denoted W_p. We will follow this convention throughout this thesis.

[1] A function $f : X \to \mathbb{R}$ is tame if there is a finite number of values $a \in \mathbb{R}$ such that for some $k \geq 0$ and all sufficiently small $\epsilon > 0$, the map $\mathrm{H}_k(f^{-1}(-\infty, a - \epsilon]) \to \mathrm{H}_k(f^{-1}(-\infty, a + \epsilon])$ is not an isomorphism; and for all $t \in \mathbb{R}$ and $k \geq 0$, the homology groups $\mathrm{H}_k(f^{-1}(-\infty, t])$ are finite-dimensional.

Wasserstein metrics give a more accurate notion of proximity, particularly for lower values of p. In a number of practical applications, they are preferred to the bottleneck distance precisely for this reason [BW20; GH10; Gid17; Ham+22]. However, their stability properties have remained less well-known until very recently, when Skraba and Turner [ST21] established the following stability theorem.

Theorem 1.5.3 (Cellular Wasserstein Stability Theorem, [ST21])**.** *Let $f, g : X \to \mathbb{R}$ be monotone functions, i.e. $f(\sigma) \leq f(\tau)$ whenever $\sigma \subset \tau$, defined on a finite CW complex X. Then, the following inequality*

$$W_p(\mathrm{PD}(f), \mathrm{PD}(g)) \leq \|f - g\|_p$$

holds, where $\mathrm{PD}(f)$ *refers to the persistence diagram obtained from the homology of the sublevel set filtration.*

Persistence Modules. The stability theorem by Cohen-Steiner *et al.* [CEH05; CEH07] is only defined for continuous tame functions over triangulable spaces. This limitation is overcome by the *algebraic* stability theorem by Chazal *et al.* [Cha+09], where the authors worked directly at the level of persistence modules and introduced the *interleaving distance*.

Definition 1.5.4 (**ϵ-interleaving**)**.** Let $M, N \in \mathsf{Vec}^{\mathsf{T}}$ be two persistence modules and $\epsilon > 0$. An *ϵ-interleaving* is given by two families of maps $\phi_t : M(t) \to N(t+\epsilon)$ and $\varphi_t : N(t) \to M(t+\epsilon)$, indexed by the elements in the poset $t \in T$, such that the following diagrams for $s \leq t \in T$ commute

$$\begin{array}{ccc} M(s) & \longrightarrow & M(t) \\ & {\scriptstyle \phi_s} \searrow \quad \searrow {\scriptstyle \phi_t} & \\ & N(s+\epsilon) \longrightarrow N(t+\epsilon), & \end{array} \qquad \begin{array}{ccc} & & M(s+\epsilon) \longrightarrow M(t+\epsilon) \\ & {\scriptstyle \varphi_s} \nearrow \quad \nearrow {\scriptstyle \varphi_t} & \\ N(s) & \longrightarrow & N(t) \end{array}$$

where the horizontal maps are the internal maps of the persistence modules. In other words, ϕ and φ define *natural transformations* between the functors M and N translated by ϵ; and N and M translated by ϵ, respectively. In addition, the maps $\{\phi_t : t \in T\}$ and $\{\varphi_t : t \in T\}$ need to make the following diagrams commute for all $t \in T$

$$\begin{array}{ccc} M(t) & \xrightarrow{M^t_{t+2\epsilon}} & M(t+2\epsilon), \\ & {\scriptstyle \phi_t} \searrow \quad \nearrow {\scriptstyle \varphi_{t+\epsilon}} & \\ & N(t+\epsilon) & \end{array} \qquad \begin{array}{ccc} & M(t+\epsilon) & \\ {\scriptstyle \varphi_t} \nearrow & & \searrow {\scriptstyle \phi_{t+\epsilon}} \\ N(t) & \xrightarrow{N^t_{t+2\epsilon}} & N(t+2\epsilon). \end{array}$$

We say that two modules are ϵ-interleaved if there exist an ϵ-interleaving between them for some $\epsilon > 0$.

Definition 1.5.5 (**Interleaving distance).** For M, $N \in \mathsf{Vec}^{\mathsf{T}}$, the *interleaving distance* is defined as

$$d_I(M,N) \coloneqq \inf\{\epsilon > 0 : M \text{ and } N \text{ are } \epsilon\text{-interleaved}\} \tag{1.5.3}$$

and $d_I(M,N) = \infty$ if there is no possible interleaving between M and N.

This is an extended pseudometric in the space of persistence modules up to isomorphism, as there are modules that are not isomorphic for which their interleaving distance is 0. For instance, taking M to be a trivial persistence module over $\mathbb{R}$, that is $M(x) = 0$ for all $x \in \mathbb{R}$, and N the module such that $N(x) = 0$ for all $x \neq 0$ and $N(0) = \Bbbk$, we have that the interleaving distance between the two is 0 but they are clearly not isomorphic.

The algebraic stability theorem establishes that if two persistence modules M, $N \in \mathsf{Vec}^{\mathsf{T}}$ are ϵ-interleaved, then the bottleneck distance between their persistence diagrams is bounded from above by $\epsilon > 0$; or equivalently, $d_B(\mathrm{PD}(M), \mathrm{PD}(N)) \leq d_I(M,N)$. The *isometry theorem* [Les15, Theorem 3.4] establishes the converse inequality for single-parameter persistence, allowing us to conclude that $d_B(\mathrm{PD}(M), \mathrm{PD}(N)) = d_I(M,N)$.

Multiparameter Persistence. The interleaving distance is formulated in terms of general posets and directly applies to multiparameter persistence. It has been one of the most widely used metrics in this setting, as it enjoys good theoretical properties. From a computational perspective, the interleaving distance in known to be NP-hard to compute [BBK20] except when it reduces to the bottleneck distance thanks to the isometry theorem. Note that, on the contrary, checking for 0-interleavings between two q-tame modules, which is equivalent to looking for isomorphisms, can be solved in polynomial time [BL08].

The extension of the bottleneck and Wasserstein metrics, with a combinatorial nature amenable to computations, from single- to multiparameter persistence is an active area of research. For interval decomposable modules we can use *partial matchings* [BL15] between the sets of intervals to achieve this extension.

Definition 1.5.6 (**Multiparameter bottleneck distance**, [Bak21] and **Multiparameter Wasserstein distance**, [BSS23])**.** Let $M, N \in \mathsf{Vec}^{\mathbb{R}^n}$ be interval decomposable modules, that is, $M \cong \bigoplus_{j\in\mathcal{J}} \Bbbk_{J_j}$ and $N \cong \bigoplus_{k\in\mathcal{K}} \Bbbk_{K_k}$, with $\{J_j : j \in \mathcal{J}\}$ and $\{K_k : k \in \mathcal{K}\}$ multisets of intervals. Then, the *bottleneck distance* between them is defined as

$$d_B(M,N) \coloneqq \inf_{\phi:\mathcal{I}\to\mathcal{K}} \max\left(\sup_{\phi(i)=k} d_I(\Bbbk_{J_i}, \Bbbk_{K_k}),\ \sup_{j\in\mathcal{J}\setminus\mathcal{I}} d_I(\Bbbk_{J_j}, 0),\ \sup_{k\in\mathcal{K}\setminus\phi(\mathcal{I})} d_I(0, \Bbbk_{K_k}) \right) \tag{1.5.4}$$

and the *Wasserstein distance* between them is defined as

$$W_p(M,N) \coloneqq \inf_{\phi:\mathcal{I}\to\mathcal{K}} \left[\sum_{\phi(i)=k} d_I(\Bbbk_{J_i}, \Bbbk_{K_k})^p + \sum_{j\in\mathcal{J}\setminus\mathcal{I}} d_I(\Bbbk_{J_j}, 0)^p + \sum_{k\in\mathcal{K}\setminus\phi(\mathcal{I})} d_I(0, \Bbbk_{K_k})^p \right]^{1/p}. \tag{1.5.5}$$

where ϕ ranges over all partial matchings, i.e. injections of subsets $\mathcal{I} \subset \mathcal{J}$ into $\mathcal{K}$.

Rank Functions and Persistence Landscapes. As functional summaries, both rank functions and persistence landscapes are endowed with L^p metrics. For a module $M \in \mathsf{Vec}^{\mathbb{R}^n}$, its corresponding rank invariant can be seen as a function $\mathrm{rk}^M : \mathbb{R}^{2n} \to \mathbb{R}$.

Definition 1.5.7 (L^p **metric for rank invariants).** For $M, N \in \mathsf{Vec}^{\mathbb{R}^n}$ persistente modules, the L^p metric between their corresponding rank invariants rk^M and rk^N is defined as

$$d_{L^p}(\mathrm{rk}^M, \mathrm{rk}^N) \coloneqq \left\| \mathrm{rk}^M - \mathrm{rk}^N \right\|_p = \left(\int_{\mathbb{R}^{2n}} \left| \mathrm{rk}^M - \mathrm{rk}^M \right|^p d\mu \right)^{1/p} \tag{1.5.6}$$

where μ is the Lebesgue measure in $\mathbb{R}^{2n}$.

Under this metric, two rank functions might have infinite distance (for instance in single parameter persistence when we have infinite cycles at distinct birth times in M and N). This is not a problem when we work with filtrations where the last stage is a simplicial complex with trivial homology, as all homology classes must disappear at some point, and there are no infinite cycles.

An alternative metric for rank functions, which we include for completeness but which will not be the focus of this thesis, is the *matching distance* [dFL06; dFL10].

Definition 1.5.8 (**Matching distance**, Definition 6.11, [BL23])**.** For $M, N \in \mathsf{Vec}^{\mathbb{R}^n}$ persistence modules, the *matching distance* between their rank invariants rk^M and rk^N is defined as

$$d_{\mathrm{match}}(\mathrm{rk}^M, \mathrm{rk}^N) \coloneqq \sup_L d_B(\mathrm{PB}(M_L),\, \mathrm{PB}(N_L))$$

where $L : \mathbb{R} \to \mathbb{R}^n$ varies in the set of lines in $L(t) = vt + b$ such that $v \in [1, \infty)^n$ and $b \in \mathbb{R}^n$. The modules M_L, $N_L \in \mathsf{Vec}^{\mathbb{R}}$ denote the single-parameter persistence modules obtained after restricting M and N to L.

2 Dualities in Persistence and Fast Cycle Matching Computations

In this chapter, we explore the role of dualities in PH from a practical perspective—in order to compute barcode matchings in an efficient way—and a theoretical one. Part of content of this chapter can be found in my joint published work [GMS24].

We begin the chapter motivating the importance of dualities in persistence in **Section** 2.1. We then introduce the problem of *cycle matching* in **Section** 2.2 and overview previous literature relevant to our approach in **Section** 2.3 and **Section** 2.4. We cover our cohomological cycle matching and the contributions in [GMS24] in **Section** 2.5. The chapter is closed with some observations stemming from unpublished, ongoing research that explores the role of dualities in multiparameter persistence in **Section** 2.6.

2.1 The Relevance of Dualities in the Theory of Persistence

In this section, we go over over some of the unique characteristics of cohomology that make it an interesting theory in itself, and review how these have been implemented to advance the theory of persistence. Recall that in **Section** 1.2 we saw that homology (**Definition** 1.2.1) and cohomology groups (**Definition** 1.2.2) with coefficients over a field like $\mathbb{R}$ are dual vector spaces thanks to the Universal Coefficient Theorem. Additionally, when they have finite dimension, they are isomorphic.

In general, even taking coefficients over an abelian group, the dualizing step required to define cohomology makes it *contravariant* instead of *covariant*, that is, if $f : X \to Y$ is a simplicial map, the induced linear map at the level of cohomology reverses the source and target $\mathrm{H}^k(f) : \mathrm{H}^k(Y) \to \mathrm{H}^k(X)$ by sending $\varphi \mapsto \varphi \circ f$; and it reverses compositions, $\mathrm{H}^k(g \circ f) = \mathrm{H}^k(f) \circ \mathrm{H}^k(g)$ for any two simplicial maps $f : X \to Y$ and $g : Y \to Z$. This

characteristic turns out to add some extra structure that makes cohomology richer and interesting in itself, for reasons that include the following.

- We can define a product $\frown\colon \mathrm{H}^i(X) \times \mathrm{H}^j(X) \to H^{i+j}(X)$ which turns the cohomology group $\mathrm{H}(X) \coloneqq \cup_{k\geq 0} \mathrm{H}^k(X)$ into a ring that distinguishes topological spaces with the same homology groups.
- Brown's Representability Theorem [Bro62] establishes a one-to-one correspondence between homotopy classes of maps $f : X \to \mathrm{K}(G, n)$ from some topological space X to the *Eilenber-MacLane* space $\mathrm{K}(G, n)$ [EM45b] and cocycles in the *reduced* cohomology group $\widetilde{\mathrm{H}}^n(X, G)$.
- Some contexts, such as de Rham cohomology, are more naturally formulated in the cohomology setting.

This rich structure has been instrumental in the theory of persistence. Potentially one of the most notable observations is that cohomology can be used to accelerate PH computations [SMV11; dMV11] a strategy that has been applied in most of the state-of-the-art packages [Bau+17; Bau21; BS23; BLL23], as we have already discussed. The cup product has been used to define finer invariants in single-parameter persistence [Con+22; MSZ24; DR24] and appears as well in the adaptation of Steenrod modules to persistence [LMT22]. Finally, Brown's Representability Theorem has found applications in persistence as a theoretical justification to develop circular coordinates [dMV11; Blu+24], spherical coordinates [SS24] and toroidal coordinates [Sco+23; PST23] from point cloud data. All these applications justify the inherent interest on understanding well persistent cohomology, instead of approaching it as an afterthought of homology.

In general, not only referring to homology and cohomology, we can define the following duality on persistence modules over T and T^{op}.

Definition 2.1.1 (**Duality functor).** We define the *duality functor* connecting the categories $\mathsf{vec}^{\mathsf{T}}$ and $\mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$ as

$$D_{\mathsf{T}} = \mathrm{Hom}_{\Bbbk}(-, \Bbbk) : \mathsf{vec}^{\mathsf{T}} \to \mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}. \tag{2.1.1}$$

We drop the dependence on the poset when it is unambiguous.

In more detail, for an object $M \in \mathsf{vec}^{\mathsf{T}}$ the dual object $DM \in \mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$ is the functor such that

- for $t \in \mathsf{T}$, returns $DM_t \coloneqq \mathrm{Hom}_{\Bbbk}(M_t, \Bbbk)$ the dual vector space;
- for each morphism $s \leq_{op} t$, assigns the dual of the internal map φ_s^t from M_t to M_s, namely $(\varphi_s^t)^\vee : \mathrm{Hom}_{\Bbbk}(M_s, \Bbbk) \to \mathrm{Hom}_{\Bbbk}(M_t, \Bbbk)$ with $u \mapsto u \circ \varphi_s^t$, which is well defined as $t \leq s$.

In **Section** 2.6 we will study the properties and some consequences of applying this functor to general persistence modules.

2.2 Cycle Matching: Motivation of the Problem

In **Section** 1.3 we introduced persistence barcodes, and observed that, when computed from the PH modules $\mathrm{H}_k(X_\bullet)$, $k \geq 0$, of a filtration $X_\bullet : \mathsf{T} \to \mathsf{Simp}$ such as the VR filtration, they encode information about the multiscale topological features of the input data. Each bar in the barcode corresponds to a topological feature in the data (see **Figure** 1.4). A natural question that arises is whether we can establish a correspondence between the bars in the barcodes of two different input data sets that exhibit similar topological structure, so that the matched bars represent resembling features.

A motivating example can be found **Figure** 2.1, showing four slices of a stack of images of the posterior lateral line primordium (pLLP) of a zebrafish embryo. This is a primitive expression of the lateral line, an organ in fish that regulates the way that they perceive the water flow over their bodies to coordinate their swimming. At embryonic stage, the pLLP appears as a cluster of roughly spherical cells. As the organ is scanned across its height, we can observe the changing contours of these cells in the image slices. The question here is whether we can define a matching between the 1-bars in the persistence barcodes of these images to effectively "track" the evolving cell contours across slices, as the cells are imaged at varying heights.

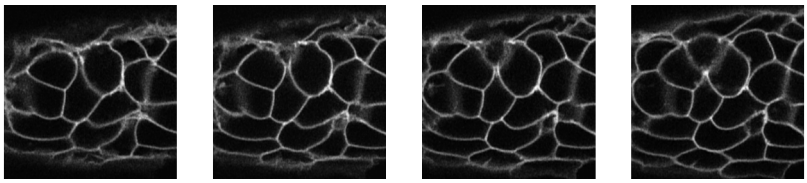


Figure 2.1: Four slices from a stack of images of the posterior lateral line primordium of a zebrafish embryo, from data set with image ID 9836972 provided by [Har+20].

There have been various attempts to answer this question in the literature, of which we now provide a brief overview. Yoon *et al.* [YGG23] proposed the methods of *persistent extension* and *analogous bars* to match signals from input data residing in potentially different spaces, using only a dissimilarity measure between them. This is particularly useful when the data that we want to match cannot be embedded into a common space in a meaningful way. Bauer and Lesnick [BL15] defined a partial matching between bars of barcodes from a known map between the persistence modules. This method was further developed by González-Díaz *et al.* [GST23], who proposed a basis-free method for partial matchings between barcodes using

the *block function*, also defined from the morphism between the persistence modules. In addition, the matching in [BL15] was reinterpreted in statistical terms by Reani and Bobrowski [RB23], who introduced a matching based on *image persistence* [Coh+09]. In our project, we focused on this last approach.

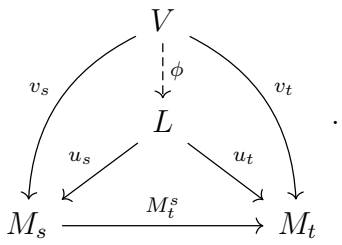
A common limitation of the proposals above is that they are all very expensive to compute. Our initial aim for this project was to leverage the strategy followed by `Ripser` [Bau21], the state-of-the-art package to compute Vietoris–Rips barcodes, to accelerate the computations of the cycle matching proposed by Reani and Bobrowski. This entailed extending the correspondences between barcodes derived in [SMV11] to the setting of *image persistence*, to be able to compute their barcode from the relative cohomology setting, as `Ripser` does. However, just some months after we had started working on this, a paper came out providing a solution to this problem, which we present next.

2.3 Functorial Dualities in Single-Parameter Persistence and Lifespan Functors

In [BS23], Bauer and Schmahl developed a functorial generalization of **Proposition** 1.3.4 by [SMV11]. In **Section** 2.3.1, we provide the necessary background to understand this categorical formulation, and in **Section** 2.3.2, we review the lifespan functors defined in [BS23] and the functorial generalization of **Proposition** 1.3.4.

2.3.1 Category Theory Preliminaries

Limits and Colimits. Let T be the category associated to a poset $(T, \leq)$. For a T-shaped diagram $M : \mathsf{T} \to \mathcal{C}$ in an abelian category $\mathcal{C}$, a generalized *limit* is given by an object $L \in \mathcal{C}$ and a family of morphisms $u_t : L \to M_t$ commuting with the internal maps of M, which satisfy the following universal condition: for any other pair $(V, v_t : V \to M_t)$ as above, there exists a unique morphism $\phi : V \to L$ such that the following diagram commutes

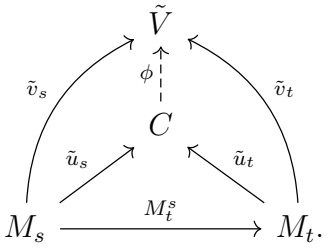


.

In other words, the limit is a terminal object in the category of pairs $(C, (v_t)_{t \in T})$ with $v_t : C \to M_t$ commuting with the internal maps, also called the category of *cones* of M.

Limits are unique up to isomorphism, so that it makes sense to talk about *the* limit of a given diagram, and denote it by $\lim_{\mathsf{T}} M$.

Dually, we can define the notion of a generalized *colimit* as an object $C \in \mathcal{C}$ and a family of morphisms $\tilde{u}_t : M_t \to C$ commuting with the internal maps of M, satisfying the following universal property: for $(\tilde{V}, \tilde{v}_t : M_t \to V)$ as above, there exists a unique morphism $\phi : C \to \tilde{V}$ such that the following diagram commutes



This is an initial object in the category of cocones, and again, it is uniquely defined up to isomorphism, so we can talk about *the* colimit of a given diagram and denote it by $\operatorname{colim}_{\mathsf{T}} M$. We can see the limit and the colimit as functors $\lim_{\mathsf{T}}, \operatorname{colim}_{\mathsf{T}} : \mathcal{C}^{\mathsf{T}} \to \mathcal{C}$.

In the category of persistence modules, $\mathsf{Vec}^{\mathsf{T}}$, limits and colimits always exist and can be computed in the following way. For $M \in \mathsf{Vec}^{\mathsf{T}}$, we have that its colimit is given by the quotient

$$\operatorname{colim}_{\mathsf{T}} M = \frac{\bigoplus_{t \in T} M_t}{\sim} \tag{2.3.1}$$

where $m_s \in M_s$ and $m_t \in M_t$ are related by the equivalence relation $m_s \sim m_t$, if and only if $s \leq t$ and $M_t^s(m_s) = m_t$. On the other hand, the limit of M is given by

$$\lim_{\mathsf{T}} M = \{(m_t)_{t \in \mathsf{T}} : m_t = M_t^s(m_s),\ \forall s \leq t\} \tag{2.3.2}$$

which is a subspace of the product vector space $\bigoplus_{t \in T} M_t$.

Adjunctions. Given a pair of covariant functors $L : \mathcal{C} \to \mathcal{C}'$ and $R : \mathcal{C}' \to \mathcal{C}$, we say that they are adjoint, denoted by $L \dashv R$, if for any pair $A \in \operatorname{Obj}(\mathcal{C})$ and $B \in \operatorname{Obj}(\mathcal{C}')$ there is a natural isomorphism between $\operatorname{Hom}_{\mathcal{C}'}(L(A), B) \cong \operatorname{Hom}_{\mathcal{C}}(A, R(B))$. We also say that L is a *left adjoint* of R, or that there is an *adjunction* between $\mathcal{C}$ and $\mathcal{C}'$ whenever such L and R exist. Given such an adjunction, it is easy to check that L is right exact and R is left exact. In addition, each adjunction $L \dashv R$ comes with two natural transformations $\eta : \mathrm{id}_{\mathcal{C}} \to R \circ L$ and $\epsilon : L \circ R \to \mathrm{id}_{\mathcal{C}'}$, called the unit and the counit, respectively.

As an example, let $\Delta : \mathcal{C} \to \mathcal{C}^{\mathsf{T}}$ be the diagonal functor, sending some object U to the diagram ΔU such that $(\Delta U)_t = U$ for all $t \in T$ and with all internal maps being the identity map $U_t^s = \mathrm{id}_U$ for $s \leq t$. It turns out that the colimit and the limit are precisely the left and right

adjoints of this functor, that is,

$$\operatorname{colim}_{\mathsf{T}} \dashv \Delta \dashv \lim_{\mathsf{T}}.$$

This directly gives that $\operatorname{colim}_{\mathsf{T}} : \mathcal{C}^{\mathsf{T}} \to \mathcal{C}$ is a right exact functor and $\lim_{\mathsf{T}} : \mathcal{C}^{\mathsf{T}} \to \mathcal{C}$ is a left exact functor. The question of their exactness is more subtle.

2.3.2 Lifespan Functors and Dualities

Recall the adjunctions $\operatorname{colim}_{\mathsf{T}} \dashv \Delta \dashv \lim_{\mathsf{T}}$. Taking the unit of the leftmost adjunction, $\eta : \mathrm{id}_{\mathcal{C}^{\mathsf{T}}} \to \Delta \circ \operatorname{colim}_{\mathsf{T}}$ and the counit of the rightmost one $\epsilon : \Delta \circ \lim_{\mathsf{T}} \to \mathrm{id}_{\mathcal{C}^{\mathsf{T}}}$ and evaluating at some $M \in \mathcal{C}^{\mathsf{T}}$ we obtain natural transformations $\eta_M : M \to \Delta(\operatorname{colim}_{\mathsf{T}} M)$ and $\epsilon_M : \Delta(\lim_{\mathsf{T}}) \to M$, which are prescribed by the maps between M and the colimit, and the limit and M, respectively. Putting everything together we arrive to the diagram

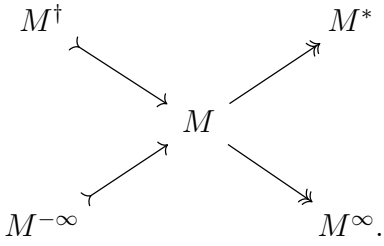
$$\Delta \lim_{\mathsf{T}} M \xrightarrow{\epsilon_M} M \xrightarrow{\eta_M} \Delta \operatorname{colim}_{\mathsf{T}} M,$$

that allows the following definition.

Definition 2.3.1 (**Lifespan functors**, Definition 3.1., [BS23])**.** We define the following functors $\mathcal{C}^{\mathsf{T}} \to \mathcal{C}^{\mathsf{T}}$.

- The *mortal part* functor is defined as $(-)^{\dagger} = \ker \eta_{(-)}$.
- The *immortal part* functor is defined as $(-)^{\infty} = \operatorname{im} \eta_{(-)}$.
- The *nascent part* functor is defined as $(-)^{*} = \operatorname{coker} \epsilon_{(-)}$.
- The *ancient part* functor is defined as $(-)^{-\infty} = \operatorname{im} \epsilon_{(-)}$.

These functors give rise, for each $M \in \mathcal{C}^{\mathsf{T}}$, to the following natural diagram with short exact sequences in the diagonals



The definition of the lifespan functors involves images, kernels and cokernels of the unit and the counit, but there are two combinations that have not been used: $\operatorname{coker} \eta_{(-)}$ and $\ker \epsilon_{(-)}$. These do not yield subobjects or quotients of the original persistence module we apply them to, but still play a role in [BS23].

Definition 2.3.2 (Definition 3.7., [BS23])**.** We define the following functors $\mathcal{C}^{\mathsf{T}} \to \mathcal{C}^{\mathsf{T}}$.

- The *ghost complement* is defined as $(-)^{\triangleright} = \ker \epsilon_{(-)}$.
- The *unborn complement* is defined as $(-)^{\triangleleft} = \operatorname{coker} \eta_{(-)}$.

Example 2.3.3. Let $\mathsf{T} = \{a, b\}$ with $a \leq b$. Using **Equation** (2.3.1) and **Equation** (2.3.2) to compute the colimit and the limit of the three diagrams in the first row, and looking at the kernels, images and cokernels of their maps, we obtain the results in **Table** 2.1.

Table 2.1: Some examples of the lifespan functors.

M	$\Bbbk \to 0$	$\Bbbk \xrightarrow{\mathrm{id}} \Bbbk$	$0 \to \Bbbk$
$\operatorname{colim}_{\mathsf{T}} M$	0	$\Bbbk$	$\Bbbk$
$\lim_{\mathsf{T}} M$	$\Bbbk$	$\Bbbk$	0
$M^{\dagger}$	$\Bbbk \to 0$	$0 \to 0$	$0 \to 0$
M^{∞}	$0 \to 0$	$\Bbbk \to \Bbbk$	$0 \to \Bbbk$
M^{*}	$0 \to 0$	$0 \to 0$	$0 \to \Bbbk$
$M^{-\infty}$	$\Bbbk \to 0$	$\Bbbk \to \Bbbk$	$0 \to 0$
$M^{\triangleright}$	$0 \to \Bbbk$	$0 \to 0$	$0 \to 0$
$M^{\triangleleft}$	$0 \to 0$	$0 \to 0$	$\Bbbk \to 0$

Looking at **Table** 2.1 we can understand the reasoning behind the naming of the lifespan functors. Intuitively, the mortal part $(-)^{\dagger}$ is meant to capture intervals (indecomposables in single-parameter persistence) that go to 0 (die) at some parameter value within the module. The immortal part $(-)^{\infty}$, on the contrary, captures the components that persist until the end of the parameter set. Dually, the nascent part $(-)^{*}$, isolates the indecomposables that are born at some point, and the ancient part $(-)^{-\infty}$, the ones that existed at the beginning of the module. The ghost complement $(-)^{\triangleright}$ tracks intervals that have died after they disappear, and the unborn complement $(-)^{\triangleleft}$, the components that have not been born yet.

We now state the functorial version of **Proposition** 1.3.4 given in [BS23] which relates the barcodes of the absolute and relative settings. It requires a technical condition on the filtrations that they consider, $X_{\bullet} \in \mathsf{Simp}^{\mathsf{T}}$, namely, that they are *colimit proper*. This means that the natural maps

$$\operatorname{colim}_{\mathsf{T}} \mathrm{H}_k(X_{\bullet}) \to \mathrm{H}_k(\operatorname{colim}_{\mathsf{T}} X_{\bullet}) \text{ and } \mathrm{H}_k(\operatorname{colim}_{\mathsf{T}} X_{\bullet}) \to \lim_{\mathsf{T}} \mathrm{H}_k(\operatorname{colim}_{\mathsf{T}} X_{\bullet}, X_{\bullet})$$

are isomorphisms for all $k \geq 0$.

Theorem 2.3.4 (Theorem 6.2., [BS23])**.** *Let $X_{\bullet} \in \mathsf{Simp}^{\mathsf{T}}$ be colimit proper, with $(T, \leq)$ a totally ordered set. For all $k \geq 0$, we have the following isomorphisms which are natural in $X_{\bullet}$*

$$\begin{aligned} \mathrm{H}_{k-1}(X_{\bullet})^{\dagger} &\cong \mathrm{H}_k(\operatorname{colim} X_{\bullet}, X_{\bullet})^{*}, \\ \mathrm{H}_k(X_{\bullet})^{\triangleleft} &\cong \mathrm{H}_k(\operatorname{colim} X_{\bullet}, X_{\bullet})^{-\infty}, \\ \mathrm{H}_k(X_{\bullet})^{\infty} &\cong \mathrm{H}_k(\operatorname{colim} X_{\bullet}, X_{\bullet})^{\triangleright}. \end{aligned}$$

We close the section with the effect that dualization (cf. **Definition** 2.1.1) causes on the lifespan functors, justifying the appearance of cohomology in the next section (cf. **Proposition** 2.4.2).

Proposition 2.3.5 (Proposition 4.13., [BS23]). *For $M \in \mathsf{vec}^{\mathsf{T}}$ we have the following isomorphisms:*

$$D(M^{\dagger}) \cong D(M)^{*}, \quad D(M^{\infty}) \cong D(M)^{-\infty} \quad \textit{and} \quad D(M^{\triangleleft}) \cong D(M)^{\triangleright}.$$

Remark 2.3.6. In [BS23], everything is stated and proved assuming that $(T, \leq)$ is a totally ordered set, that is, in the context of single-parameter persistence. However, the definition of the lifespan functors can be directly extended to multiparameter persistence. A question that naturally arises is which of the properties of the lifespan functors presented in [BS23] extend as well to this context. We will provide some partial answers, product of ongoing research, in **Section** 2.6.

2.4 Image, Kernel, and Cokernel Persistence and Dualities

In this section we define one of the key ingredients of many of the matching procedures that we reviewed in **Section** 2.2: *image persistence*, which was introduced in [Coh+09] alongside the natural counterparts of *kernel* and *cokernel persistence*. We will then see how to apply **Theorem** 2.3.4 to obtain equivalences between their barcodes in single-parameter persistence.

Let $X_\bullet, Y_\bullet \in \mathsf{Simp}^{\mathsf{T}}$ be filtrations and $f_\bullet : X_\bullet \to Y_\bullet$ a morphism between them. This induces a morphism at the level of homology for the corresponding persistence modules $\mathrm{H}_k(f_\bullet) : \mathrm{H}_k(X_\bullet) \to \mathrm{H}_k(Y_\bullet)$, meaning that for $s \leq t \in T$, the following diagram commutes

$$\begin{array}{ccc} \mathrm{H}_k(X_s) & \xrightarrow{i_t^s} & \mathrm{H}_k(X_t) \\ {\scriptstyle \mathrm{H}_k(f_s)}\downarrow & & \downarrow{\scriptstyle \mathrm{H}_k(f_t)} \\ \mathrm{H}_k(Y_s) & \xrightarrow{j_t^s} & \mathrm{H}_k(Y_t) \end{array}$$

where i_t^s and j_t^s are the maps induced at homology level by the inclusions $X_s \subseteq X_t$ and $Y_s \subseteq Y_t$, respectively. The commutativity of this diagram, i.e. the fact that $j_t^s \circ \mathrm{H}_k(f_s) = \mathrm{H}_k(f_t) \circ i_t^s$, shows that $j_t^s|_{\mathrm{im}(\mathrm{H}_k(f_s))} \subseteq \mathrm{im}(\mathrm{H}_k(f_t))$. This fact proves that the following persistence module is well-defined.

Definition 2.4.1 (**Image persistence module**, [Coh+09]). Given a morphism $f_\bullet : X_\bullet \to Y_\bullet$, between filtrations $X_\bullet, Y_\bullet \in \mathsf{Simp}^{\mathsf{T}}$, we define the *image persistence module*, denoted $\mathrm{im}(\mathrm{H}_k(f_\bullet))$, as the persistence module with vector spaces $\mathrm{im}(\mathrm{H}_k(f_t))$ for $t \in T$ and transition maps given by the restrictions $j_t^s|_{\mathrm{im}(\mathrm{H}_k(f_s))}$ for all $s \leq t \in T$.

Similar arguments apply to the kernel and the cokernel of the morphism $f_\bullet$, motivating the definition of *kernel* and *cokernel persistence* as well.

Now, the naturality of the isomorphisms in **Theorem** 2.3.4 for the variable of the filtration implies that, for a morphism $f_\bullet : X_\bullet \to Y_\bullet$ between filtrations, calling $\phi = \operatorname{colim}_{\mathsf{T}} f$ to the map induced between their colimits, we get isomorphisms

$$\mathrm{H}_{k-1}(f_\bullet)^\dagger \cong \mathrm{H}_k(\phi, f_\bullet)^*, \quad \mathrm{H}_k(f_\bullet)^\triangleleft \cong \mathrm{H}_k(\phi, f_\bullet)^{-\infty}, \quad \mathrm{H}_k(f_\bullet)^\infty \cong H_k(\phi, f_\bullet)^\triangleright$$

in the category of morphisms of persistence modules. These translate to isomorphisms between the corresponding kernels, images and cokernels, allowing to deduce the following correspondence between the barcodes of image persistence modules in single-parameter persistence.

Proposition 2.4.2 (Proposition 21, [BS22]). *Let $X_\bullet$ and $Y_\bullet$ be filtrations of two isomorphic simplicial complexes $X \cong Y$, and $f_\bullet : X_\bullet \to Y_\bullet$ be monomorphims inducing an isomorphism $f : X \to Y$. For all $k \geq 0$ we have the equality between the finite bars of the following barcodes*

$$\mathrm{PB}(\operatorname{im} \mathrm{H}_k(f_\bullet))^\dagger = \mathrm{PB}(\operatorname{im} \mathrm{H}^{k+1}(f, f_\bullet))^\dagger$$

and that the map sending each interval I to its complementary in the poset $I \mapsto T \setminus I$ defines a bijection between the infinite bars of

$$\mathrm{PB}(\operatorname{im} \mathrm{H}_k(f_\bullet))^\infty \leftrightarrow \mathrm{PB}(\mathrm{H}^k(X, X_\bullet))^\infty, \quad \textit{and} \quad \mathrm{PB}(\operatorname{im} \mathrm{H}^k(f, f_\bullet))^\infty \leftrightarrow \mathrm{PB}(\mathrm{H}_k(Y_\bullet))^\infty.$$

Remark 2.4.3. We explain an apparent discrepancy between [BS23, Definition 4.3 and Theorem 4.4] and [BS22, Porposition 21]. The former imply that, for a persistence module $M \in \mathsf{vec}^{\mathbb{R}}$, the barcode of its mortal part is given by the bars in $\mathrm{PB}(M)$ which are strictly bounded above, denoted $\mathrm{PB}(M)^\dagger = \mathrm{PB}(M^\dagger)$. In [BS22, Proposition 21], however, they use this notation to refer to the finite bars instead, meaning that they are strictly bounded above and below. This is because they are making some assumptions on the filtrations given the computational focus of the paper. In particular, they assume that both filtrations start being empty, and thus none of the intervals in the barcodes span the whole indexing set, which implies that $\mathrm{PB}(-)^\dagger$ corresponds to the finite bars as stated.

The equivalences in **Proposition** 2.4.2 were used to implement `Ripser-image`, a package that computes image persistence barcodes using the optimizations typical of `Ripser`. This is the code we leveraged to provide our cohomological version of cycle matching.

2.5 Cohomological Cycle Matching

We now summarize the contributions of [GMS24]. We begin the section by reviewing the concept of interval matching introduced by Reani and Bobrowski [RB23] in **Section** 2.5.1.

Next, in **Section** 2.5.2, we present our theoretical contribution, which adapts the framework of [RB23] to be compatible with `Ripser`. The implementation details of the accompanying open-source code (`https://github.com/inesgare/interval-matching`) are discussed in **Section** 2.5.3. Finally, in **Section** 2.5.4, we explore several applications introduced in [GMS24].

2.5.1 Matching Intervals in Morse Filtrations

Reani and Bobrowski [RB23] restrict their methodology to Morse filtrations.

Definition 2.5.1 (**Morse filtration).** A filtration $X_\bullet = \{X_t : t \in \mathbb{R}\}$ is a *Morse filtration* if there exists a finite set $T = \{t_1, ..., t_n\} \subset \mathbb{R}$ such that the following are satisfied:

1. For all $t \notin T$, there exists $\epsilon > 0$ small enough such that for every $0 < \epsilon' < \epsilon$ the map

$$\mathrm{H}_k(i) : \mathrm{H}_k(X_{t-\epsilon'}) \to \mathrm{H}_k(X_{t+\epsilon'})$$

   induced by inclusion is an isomorphism for all $k \geq 0$. Equivalently, the homology does not change at $t \notin T$.

2. For all $t \in T$, there exists $\epsilon > 0$ small enough so that for any $0 < \epsilon' < \epsilon$ either

   (a) $\mathrm{H}_k(i) : \mathrm{H}_k(X_{t-\epsilon'}) \to \mathrm{H}_k(X_{t+\epsilon'})$ is injective and the dimension of the vector space increases by one, or

   (b) $\mathrm{H}_k(i) : \mathrm{H}_k(X_{t-\epsilon'}) \to \mathrm{H}_k(X_{t+\epsilon'})$ is surjective and the dimension decreases by one.

   Equivalently, the homology changes allowed are either the creation of a single new cycle or the termination of a single existing cycle.

Consider now three Morse filtrations $X_\bullet = \{X_t : t \in T\}$, $Y_\bullet = \{Y_t : t \in T\}$ and $Z_\bullet = \{Z_t : t \in T\}$ with the same indexing set. Assume that we have monomorphisms $f_\bullet : X_\bullet \to Z_\bullet$ and $g_\bullet : Y_\bullet \to Z_\bullet$. Typically, we have $X_\bullet = \mathcal{VR}_\bullet(S_1)$, $Y_\bullet = \mathcal{VR}_\bullet(S_2)$ and $Z_\bullet = \mathcal{VR}_\bullet(S_1 \cup S_2)$, where S_1 and S_2 are point clouds in the same ambient space; and the maps $f_\bullet$ and $g_\bullet$ are simply the inclusions induced in the VR filtrations.

Definition 2.5.2 (**Interval matching**, [RB23])**.** Let $\alpha \in \mathrm{PB}(\mathrm{H}_k(X_\bullet))$ and $\beta \in \mathrm{PB}(\mathrm{H}_k(Y_\bullet))$. We say that α and β are *matching intervals via* $Z_\bullet$ if there exist $\tilde{\alpha} \in \mathrm{PB}(\operatorname{Im} \mathrm{H}_k(f_\bullet))$ and $\tilde{\beta} \in \mathrm{PB}(\operatorname{Im} \mathrm{H}_k(g_\bullet))$ such that

$$\begin{aligned} \operatorname{birth} \alpha &= \operatorname{birth} \tilde{\alpha}, \\ \operatorname{birth} \beta &= \operatorname{birth} \tilde{\beta}, \\ \operatorname{death} \tilde{\alpha} &= \operatorname{death} \tilde{\beta}. \end{aligned}$$

The Morse assumption is key for this definition to be well-defined, so that there are no bars with shared birth or death time that can cause ambiguous matchings. This poses an issue when integrating this method with `Ripser`, as Vietoris–Rips filtrations need not be Morse in principle.

Reani and Bobrowski [RB23] observed that, when the size of a sample grows, matches between long intervals that have no topological resemblance tend to appear often. This motivated the definition of the *affinity* of the matching, which in turn allows the definition of the *prevalence* of reappearance of a matching. For two intervals $I, J \subseteq \mathbb{R}$, its *Jaccard index* is defined as $\mathrm{Jac}\,(I, J) \coloneqq \frac{|I \cap J|}{|I \cup J|}$ where $|-|$ denotes the length of the interval.

Definition 2.5.3 (**Matching affinity**, [RB23])**.** The *matching affinity* of two intervals α and β matching through their image persistence bars $\tilde{\alpha}$ and $\tilde{\beta}$ is defined as

$$\mathrm{aff}\,(\alpha, \beta) \coloneqq \mathrm{Jac}\,(\alpha, \beta) \cdot \mathrm{Jac}\,(\alpha, \tilde{\alpha}) \cdot \mathrm{Jac}\,(\beta, \tilde{\beta})$$

Definition 2.5.4 (**Prevalence of a bar**, [RB23])**.** Let $X_\bullet^{\mathrm{ref}} = \mathcal{VR}_\bullet(S^{\mathrm{ref}})$ be the Vietoris–Rips Filtration obtained from some sample S^{ref}, and let $S^{(1)}, \ldots, S^{(K)}$ be resamples obtained from S^{ref}. For a bar $\alpha \in \mathrm{PB}(\mathrm{H}_p(X_\bullet^{\mathrm{ref}}))$, $p \geq 0$, its *prevalence score* is defined as

$$\mathrm{prev}\,(\alpha) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \mathrm{aff}\,(\alpha, \beta_k(\alpha))$$

with $\beta_k(\alpha)$ the unique bar in $\mathrm{PB}(\mathrm{H}_p(\mathcal{VR}_\bullet(S^{(k)})))$ matched to α, for $1 \leq k \leq K$. If there is no match we set $\mathrm{aff}\, = 0$.

2.5.2 Matching Intervals in Non-Morse Filtrations

Even though the Vietoris–Rips filtration is not Morse, `Ripser` implements a re-indexing of the filtration in order to obtain a simplex-wise filtration, a requirement to define the coboundary matrix that it reduces.

Definition 2.5.5 (**Simplex-wise filtration**)**.** A filtration $X_\bullet = \{X_t : t \in T\}$ is *essential* if $s \neq t$ implies $X_s \neq X_t$. Additionally, it is a *simplex-wise filtration* if for every $t \in T$ such that $X_t \neq \emptyset$ there is some simplex σ_t and some index $s < t$, such that $X_t \setminus X_s = \{\sigma_t\}$.

Note that if a filtration is simplex-wise, there is a bijection between birth and death times in the barcode, and simplices in the filtration. `Ripser` implements a lexicographic refinement of the Vietoris–Rips filtration which orders the simplices first by dimension, then by diameter, and finally, using a *combinatorial number system* which was already introduced in [Bau+17].

Consider three simplex-wise filtrations $X_\bullet = \{X_t : t \in T\}$, $Y_\bullet = \{Y_t : t \in T\}$ and $Z_\bullet = \{Z_t : t \in T\}$. Assume that we have monomorphisms $f_\bullet : X_\bullet \to Z_\bullet$ and $g_\bullet : Y_\bullet \to Z_\bullet$. The correspondence between births and deaths in the barcode, and simplices in simplex-wise filtrations, motivates the following definition.

Definition 2.5.6 (**Interval matching in simplex-wise filtrations**, [GMS24])**.** Let $\alpha \in \mathrm{PB}(\mathrm{H}_k(X_\bullet))$ and $\beta \in \mathrm{PB}(\mathrm{H}_k(Y_\bullet))$ for $k \geq 0$. α and β are *matching intervals via* $Z_\bullet$, if there exist $\tilde{\alpha} \in \mathrm{PB}(\mathrm{Im}\,\mathrm{H}_k(f_\bullet))$ and $\tilde{\beta} \in \mathrm{PB}(\mathrm{Im}\,\mathrm{H}_k(g_\bullet))$ such that the following conditions are satisfied:

- α and $\tilde{\alpha}$ are created by the same simplex (seen in $X_\bullet$ and in $f_\bullet(X_\bullet)$, respectively);
- β and $\tilde{\beta}$ are created by the same simplex (seen in $Y_\bullet$ and in $g_\bullet(Y_\bullet)$, respectively);
- $\tilde{\alpha}$ and $\tilde{\beta}$ are destroyed by the same simplex in $Z_\bullet$.

Remark 2.5.7. In **Definition** 2.5.6, we assume that the underlying simplex-wise refinements are compatible across the three filtrations. Specifically, simplices are added in the same order to the persistence modules $\mathrm{H}_k(X_\bullet)$ and $\mathrm{H}_k(Y_\bullet)$, as well as to their image persistence modules $\mathrm{im}\,\mathrm{H}_k(f_\bullet)$ and $\mathrm{im}\,\mathrm{H}_k(g_\bullet)$. This ordering can always be achieved when $X_\bullet = \mathcal{VR}_\bullet(S_1)$, $Y_\bullet = \mathcal{VR}_\bullet(S_2)$ and $Z_\bullet = \mathcal{VR}_\bullet(S_1 \cup S_2)$, with S_1 and S_2 point clouds in the same ambient space, by first specifying the simplex orderings in $X_\bullet$ and $Y_\bullet$, and then defining the ordering in $Z_\bullet$ to match.

2.5.3 Efficiently Implementing Cycle Matching with Ripser-Image

We now overview how to implement **Definition** 2.5.6 using `Ripser-image`. The input of this package consists of two Vietoris–Rips filtrations $X_\bullet = \mathcal{VR}_\bullet(S, d)$ and $Z_\bullet = \mathcal{VR}_\bullet(S, d')$ built over the same point cloud S but with different metrics which satisy $d \geq d'$.

Our setup is slightly different. We are interested in matching the bars of the Vietoris–Rips filtrations built over two point clouds S_1 and S_2 sampled from the same metric space with distance d. We let $S = S_1 \cup S_2$ and define $d_1 = d|_{S_1}$ and $d_2 = d|_{S_2}$. Call $M_1 \coloneqq \max_{s,s' \in S_1} d_1(s, s')$ to the maximum distance between points in S_1 and let $\epsilon > 0$ be some threshold. We extend the metric space (S_1, d_1) to the union (S, d_1') by setting

$$d_1'(s, s') = \begin{cases} d_1(s, s'), & \text{if } s, s' \in S_1, \\ M_1 + \epsilon, & \text{otherwise.} \end{cases}$$

In other words, we just set a large distance between any point in S_1 and any point in S_2 and also between points in S_2 inside the union S. With this, up to scale $M_1 + \epsilon$ and the points in

S_2, we have

$$X_\bullet = \mathcal{VR}_\bullet(S_1, d_1) = \mathcal{VR}_\bullet(S, d_1')$$

and letting $Z_\bullet = \mathcal{VR}_\bullet(S, d)$, we put ourselves in the setting of `Ripser-image`. We do the same to define d_2' and consider $Y_\bullet = \mathcal{VR}(S, d_2')$, obtaining three Vietoris–Rips filtrations satisfying

$$X_\bullet \subseteq Z_\bullet \supseteq Y_\bullet$$

and perform the matching using the inclusions as the connecting maps.

Parallelizing the Matchings for Prevalence Computations and Computational Runtime. It is worth noting here that a substantial speed up when computing the prevalence score can be achieved by parallelizing the matchings involved its definition (**Definition** 2.5.4). This is the approach that we implement in `https://github.com/inesgare/interval-matching`. The matching computations are relatively straightforward, the true bottleneck in runtime in our code lies in the computation of the two image persistence barcodes, which take up most of the computational time of a matching. In **Figure** 2.2, we present the runtime of an illustrative example, where we compute cycle matching for samples over two torii of radii $R = 4$ and $r = 1$ and increasing number N of points. There, we clearly observe that most of the computational runtime is spent in computing the two image persistence modules, and that the complexity of these computations seems to follow a power law in the number of points in the sample.

2.5.4 Applications of Cycle Matching

We now present some applications of cycle matching to track topological features in spatial and video data. These were the applications implemented and driven by the author of this thesis in [GMS24].

Tunneling: Tracking Cycles over Slices. The first application that we present adresses the question posed in **Section** 2.2 regarding the imaging data shown in **Figure** 2.1 from [Har+20]. The goal is to track cell contours across a sequence of images taken at different heights along the pLLP of a zebrafish embryo. This enables us to monitor the appearance, evolution, and disappearance of cells throughout the imaging stack. We considered a stack of 15 grayscale images of 300×300 pixels, with a resolution of 0.1 μm per pixel. The gap between images is 0.66 μm. We threshold the images using the Otsu method [LCC01] and sample each image with $N = 1000$ points. We then compute the persistence barcodes of the Vietoris–Rips filtrations of these point clouds and perform interval matching between

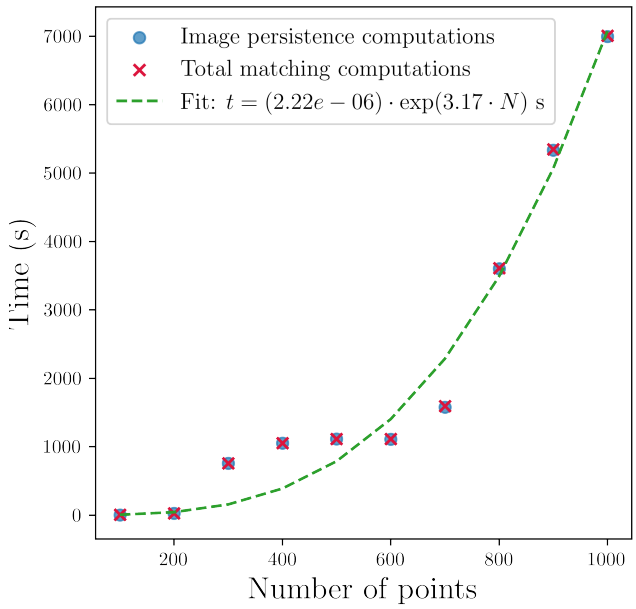


Figure 2.2: Runtime of cycle matching in point clouds of various sizes sampled from two torii of radii $R = 4$ and $r = 1$ with gaussian noise of scale 0.1 added. We draw samples from these shapes with N points, ranging from 100 to 1000 in increments of 100. We report both the total runtime (including the computation of two persistence barcodes, two image persistence barcodes and the matching) and the runtime for computing the image persistence barcodes alone. A power law fit between the total runtime and the number sample points yields $R^2 = 0.9691$.

consecutive slices. Some topological features persist across multiple slices, forming what we refer to as *tunnels*. To visualize these structures, we use the `Ripser-tight-representatives` library, which contains a feature to provide cycle representatives to the bars in the barcodes. We plot in the same color the cycles matched across slices. The results of the matching can be found in **Figure** 2.3.

Video data: Tracking Features over Time. A second application of cycle matching would be tracking the evolution of topological features across time. We demonstrate this approach using video data. The first application tracks the primitive heartbeat of the atrioventicular valve of a wild-type AB zebrafish, from [Sch+08]. The video is taken 76 hours post fecundation and at a rate of 50 miliseconds. We selected 10 frames capturing one contraction of the primitive heart valves and sampled $N = 500$ points in each image, after applying a thresholding using the mean of the gray-scale values. We successfully detected and tracked the cycles corresponding to the two valves, as well as their size variation using the matching affinity (**Definition** 2.5.3). The results can be found in **Figure** 2.4

The second example uses time-lapse images of a human embryo, also over 10 consecutive

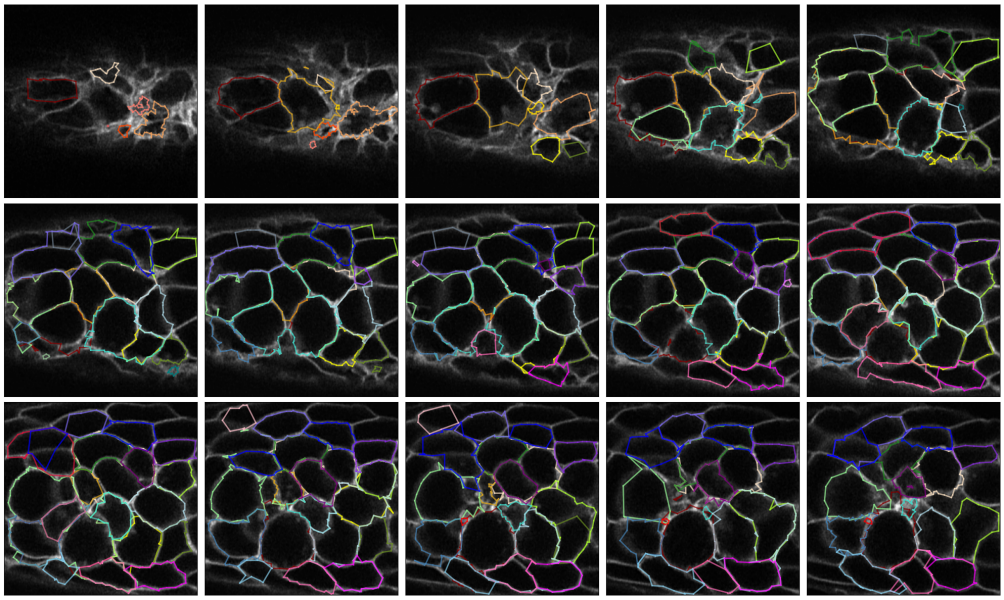


Figure 2.3: Cycle matching across consecutive slices in the pLLP data set. Cycles matched across several images are grouped into tunnels, stained in the same color. Image extracted from [GMS24, Figure 5], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

frames, from [Gom+22]. These images were captured at intervals of 50 to 100 minutes and reflect different stages of cellular division. We processed each frame by applying the Sato operator [Sat+98] followed by the Otsu thresholding method [LCC01], then sampled $N = 500$ points per iamge. The cycle matching algorithm detected the emergence of a new topological cycle, corresponding to the formation of a new cell during division, see **Figure** 2.5.

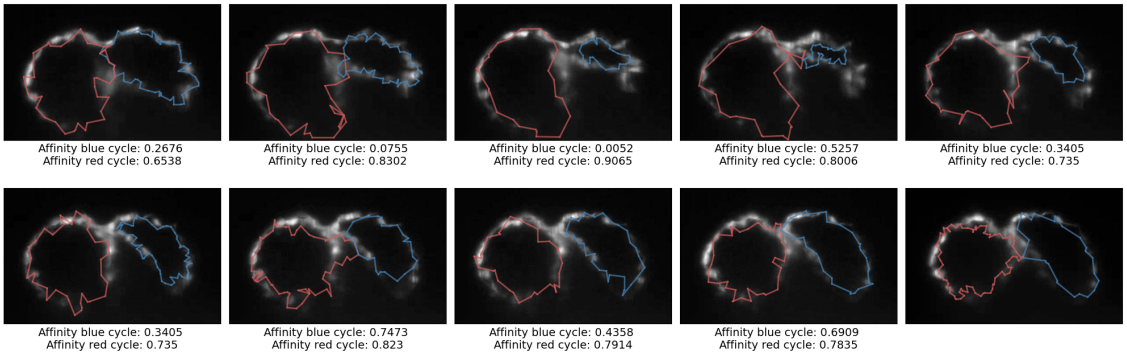


Figure 2.4: Cycle matching across frames in video data of the primitive heartbeat. Cycles representing the two atrioventicular valves of an embryo zebrafish matched across 10 frames showing the contraction of the right valve, with the affinity of the matches below each image. Image extracted from [GMS24, Figure 6], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

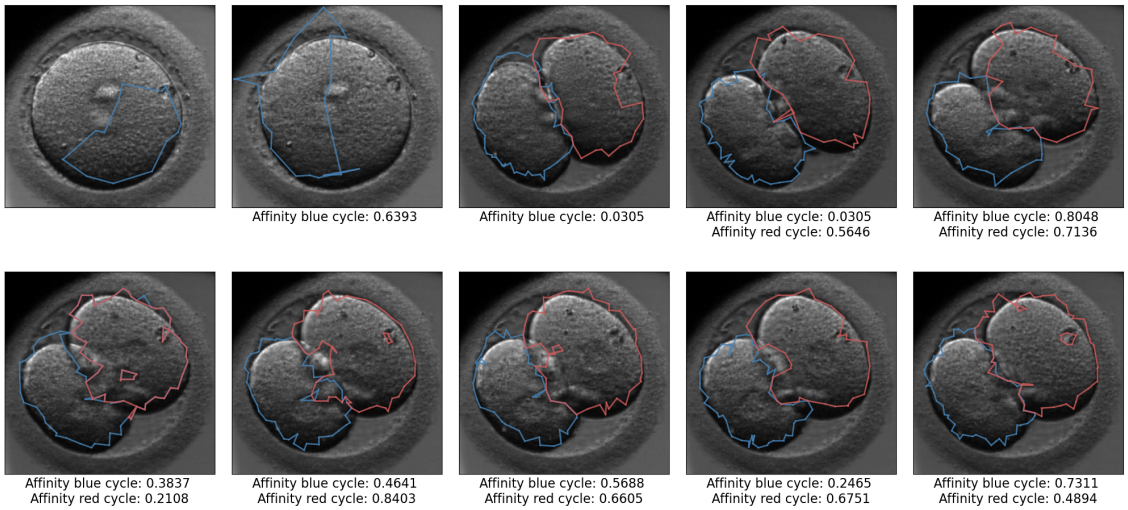


Figure 2.5: Cycle matching across frames in video data of the human embryogenesis. Cycles representing cells at different stages of the cell division process in the human embryogenesis. Image extracted from [GMS24, Figure 7], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

2.6 Functorial Dualities in Multiparameter Persistence

We now turn to the theoretical core of this chapter, and main theoretical contribution of this thesis: using adjunctions and category theory to explore dualities in multiparameter persistence. In **Section** 2.6.1, we examine projective and injective objects as dual notions, and in **Section** 2.6.2, we study how the duality functor (**Definition** 2.1.1) acts on module resolutions. This is based in ongoing, unpublished research, product of the latest research of my PhD studies.

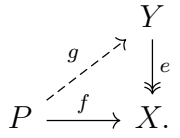
2.6.1 Characterizing Projectives and Injectives via Adjunctions

We begin this section by introducing two fundamental classes of objects in a category: projective and injective objects. We then use an adjunction to characterize the indecomposable projective and injective objects in the category of p.f.d. modules over a finite poset. We conclude by connecting to **Section** 2.3, demonstrating that a characterization of projective and injective objects proposed in [BS23], which capitalizes the lifespan functors in the setting of single-parameter persistence, fails to generalize to multiparameter persistence.

Projective and Injective Objects

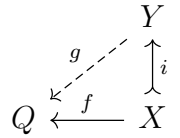
An object $P \in \mathcal{C}$ in an abelian category is said to be *projective* if for any epimorphism $e : Y \twoheadrightarrow X$ and morphism $f : P \to X$ there exists a morphism $g : P \to Y$ such that $e \circ g = f$,

in other words, such that the following diagram commutes



Equivalently, P is projective if and only if the hom functor $\mathrm{Hom}(P, -) : \mathcal{C} \to \mathsf{Ab}$, where Ab is the category of abelian groups, is exact (note that this functor is always left exact). The direct sum of projective objects is a projective object too.

Dually, an object $Q \in \mathcal{C}$ is said to be *injective* if for any monomorphisms $i : X \rightarrowtail Y$ and every morphism $f : X \to Q$ there exists a morphism $g : Y \to Q$ such that $g \circ i = f$. In other words, we get the same commutative diagram as above but reversing all arrows and interchanging the epimorphism by the monomorphism



Equivalently, Q is injective whenever the hom functor $\mathrm{Hom}(-, Q)$ is exact (dually, we know that this functor is always right exact). The direct sum of injective objects is also an injective object.

The Evaluation Functor and its Adjunctions

Let $(T, \leq)$ be a poset and fix some $t \in T$. We can define the following exact functor, which we will call the *evaluation functor*, and which is given by looking at persistence modules pointwise at t, that is

$$\begin{aligned} (-)_t : \mathsf{Vec}^{\mathsf{T}} &\to \mathsf{Vec} \\ M &\mapsto M_t \end{aligned}$$

and sending any morphism $f : M \to N$ in $\mathsf{Vec}^{\mathsf{T}}$ to the pointwise linear map $f_t : M_t \to N_t$.

Lemma 2.6.1. *The functor $(-)_t$ admits a left and a right adjoint $L_t \dashv (-)_t \dashv R_t$, and these are precisely the functors L_t, $R_t : \mathsf{Vec} \to \mathsf{Vec}^{\mathsf{T}}$ which map $U \in \mathsf{Vec}$ to*

$$L_t U = U_{[t,\infty)} \text{ and } \quad R_t U = U_{(-\infty,t]}$$

(cf. ***Equation*** (1.2.3) *for the definition of these modules).*

Proof. The functor $(-)_t$ is exact, and thus a continuous and cocontinuous functor. Since $\mathsf{Vec}^{\mathsf{T}}$ and Vec are locally small, we are in the assumptions of a special case of the General Adjoint Functor Theorem [Rie17, Theorem 4.6.17]. This means the $(-)_t$ admits a right and a left adjoint, that is, we have

$$L_t \dashv (-)_t \dashv R_t$$

where $L_t, R_t : \mathsf{Vec} \to \mathsf{Vec}^{\mathsf{T}}$ are the left and right adjoints given by the theorem, respectively.

Since $(-)_t$ is exact, and thanks to the adjunctions $L_t \dashv (-)_t \dashv R_t$, we can see that L_t preserves projective objects and R_t injective objects in Vec. Since in the category of vector spaces all objects are both injective and projective, we get that for any $U \in \mathsf{Vec}$, $L_t U$ must be a projective persistence module and $R_t U$ an injective persistence module.

From the definition of left adjoint, for $U \in \mathsf{Vec}$ and $M \in \mathsf{Vec}_{\mathsf{T}}$, we have a natural isomorphism $\mathrm{Hom}_{\mathsf{Vec}}(U, M_t) \cong \mathrm{Hom}_{\mathsf{Vec}^{\mathsf{T}}}(L_t\, U, M)$, meaning that a map $\varphi : U \to M_t$ must determine a morphism from $L_t U$ to M. This is clearly satisfied by $U_{[t,\infty)} = \bigoplus_{t \leq s} U$, where the map φ prescribes all the other pointwise maps of the morphism $L_t \varphi : U_{[t,\infty)} \to M$, for any other $s \in P$ with $t \leq s$, we get $(L_t\, \varphi)_s = M_s^t \circ \varphi : U \to M_s$. The proof for R_t is analogous.

□

Characterizing Indecomposable Projectives and Injectives

For each $t \in T$ we get an adjunction $L_t \dashv (-)_t \dashv R_t$, and each of these adjunctions comes with its respective unit and counit. We are particularly interested in the natural transformations $\epsilon_t : L_t(-)_t \to \mathrm{id}_{\mathsf{Vec}^{\mathsf{T}}}$ (the counit or the leftmost adjunction) and $\eta_t : \mathrm{id}_{\mathsf{Vec}^{\mathsf{T}}} \to R_t(-)_t$ (the unit of the rightmost one). We now restrict ourselves to finite posets $(T, \leq)$ and finite-dimensional vector spaces, i.e. we substitute Vec by vec in everything that we have said so far. Using the counit, we can prove the following.

Lemma 2.6.2. *The persistence modules $\Bbbk_{[t,\infty)}$ are the indecomposable projectives in $\mathsf{vec}^{\mathsf{T}}$.*

Proof. It is a straightforward computation to check that the persistence modules $\Bbbk_{[t,\infty)}$ are indecomposable and projective. Let then $M \in \mathsf{vec}^{\mathsf{T}}$ be a general indecomposable projective module. We want to show that $M \cong \Bbbk_{[s,\infty)}$ for some $s \in T$.

For $t \in T$, by definition, $\epsilon_t(L_t\, M_t) = M$, so ϵ_t maps projective modules to general persistence modules for all $t \in T$. Since we are working with finite-dimensional vector spaces, we can choose basis in each M_t and have the isomorphism $M_t \cong \Bbbk^{\dim(M_t)}$, from which we get

$$L_t\, M_t \cong L_t \left(\Bbbk^{\dim(M_t)}\right) \cong \bigoplus_{\dim(M_t)} L_t \Bbbk \cong \bigoplus_{\dim(M_t)} \Bbbk_{[t,\infty)}$$

where in the second isomorphism we are using that L_t, as a left adjoint, preserves coproducts (direct sums) and that products and coproducts are naturally isomorphic in $\mathsf{vec}^{\mathsf{T}}$.

The (finite) direct sum of these counits $\bigoplus_{t\in T}\epsilon_t$ gives us an epimorphism $\bigoplus_{t\in T}\bigoplus_{\dim(M_t)}\Bbbk_{[t,\infty)} \twoheadrightarrow M$ which induces an exact sequence

$$0 \to \ker\bigoplus_{t\in T}\epsilon_t \rightarrowtail \bigoplus_{t\in T}\bigoplus_{\dim(M_t)}\Bbbk_{[t,\infty)} \xrightarrow{\bigoplus_{t\in T}\epsilon_t} M \to 0.$$

Now, if M is a projective module, this sequence splits, as the definition of projective object precisely gives a morphism $f : M \to \bigoplus_{t\in T}\bigoplus_{\dim(M_t)}\Bbbk_{[t,\infty)}$ such that the diagram

$$\begin{array}{ccc} & & \bigoplus_{t\in T}\bigoplus_{\dim(M_t)}\Bbbk_{[t,\infty)} \\ & \overset{f}{\nearrow} & \big\downarrow{\scriptstyle \bigoplus_{t\in T}\epsilon_t} \\ M & \xrightarrow{\mathrm{id}_M} & M \end{array}$$

commutes, and by the splitting lemma, this implies that the sequence splits. With this, we conclude that

$$\bigoplus_{t\in T}\bigoplus_{\dim(M_t)}\Bbbk_{[t,\infty)} \cong \left(\ker\bigoplus_{t\in T}\epsilon_t\right)\oplus M \tag{2.6.1}$$

This is just another proof of the well-known fact that any projective module M is a direct summand of a free module.

In **Equation** (2.6.1) we have two isomorphic direct sum decompositions. We know that the $\Bbbk_{[t,\infty)}$ are indecomposable, and it turns out that and their endomorphism rings $\mathrm{End}(\Bbbk_{[t,\infty)}) \cong \Bbbk$ are completely primary. On the other hand, we are also assuming that M is indecomposable. Consequently, using Azumaya's Theorem [Azu50, Theorem 1] we conclude that M must be one of the indecomposable pieces on the left-hand side of **Equation** (2.6.1), that is, there exists some $s \in T$ such that $M \cong \Bbbk_{[s,\infty)}$, as desired. □

This proof might not be surprising as one of its key ingredients, the isomorphism in **Equation** (2.6.1) is a well-known fact relating free and projective modules. In fact, this characterization of projective objects is well-known too, and an alternative proof of **Lemma** 2.6.2 appears in [BBH23, Theorem 2.8]. Our approach via adjunctions, however, offers a distinct advantage: it facilitates a dualization of the argument, using the unit instead of the counit, to achieve the following.

Corollary 2.6.3. *The persistence modules $\Bbbk_{(-\infty,t]}$ are the indecomposable injectives on $\mathsf{vec}^{\mathsf{T}}$.*

Proof. The proof can be obtained by dualizing the proof of **Lemma** 2.6.2. □

The proof in [BBH23, Theorem 2.8] cannot be adapted in such a direct way to say something about injectives too. Alternative proofs of **Lemma** 2.6.2 and **Corollary** 2.6.3 appear in [BM21, Proposition 6.8], where the arguments involving injectives require introducing Matlis duality and the Baer Criterion, more involved in nature. This again highlights the strength of our approach, that only requires working with the adjuntions of the evaluation funtor.

Projectives, Injectives, and the Lifespan Functors

In [BS23, Theorem 5.5] Bauer and Schmahl leverage the lifespan functors to provide the following characterization of injectives and projectives in the category of p.f.d. single-parameter persistence modules.

Theorem 2.6.4 (Theorem 5.5, [BS23])**.** *Let $M \in \mathsf{vec}^{\mathsf{T}}$, wit T the small category obtained from a totally ordered set $(T, \leq)$.*

1. *The following are equivalent:*
   - *(a) All internal maps of M are monomorphisms,*
   - *(b) $M^{\dagger} = 0$,*
   - *(c) M is projective in $\mathsf{vec}^{\mathsf{T}}$.*
2. *The following are also equivalent:*
   - *(a) All internal maps of M are epimorphisms,*
   - *(b) $M^{*} = 0$,*
   - *(c) M is injective in $\mathsf{vec}^{\mathsf{T}}$.*

We now provide counterexamples that show that these equivalences do not hold if $(T, \leq)$ is not totally ordered.

Example 2.6.5. Let us consider a poset $(T, \leq)$ with three elements $T = \{a, b, c\}$ and two relations $b \leq a$ and $b \leq c$; and let $M \in \mathsf{vec}^{\mathsf{T}}$ be the persistence module over this poset given by

$$\Bbbk \xleftarrow{\text{id}} \Bbbk \xrightarrow{\text{id}} \Bbbk.$$

The maps in this module are monomorphisms, but the module itself is not projective, as it cannot be written as the direct sum of indecomposable projectives (cf. **Lemma** 2.6.2). This shows that 1a does not imply projectiveness, that is, 1c. In addition, from **Equation** (2.3.1) we see that $\operatorname{colim}_{\mathsf{T}} M = \Bbbk$, with all maps $\eta^{M}_{\bullet} : M_{\bullet} \to \operatorname{colim}_{\mathsf{T}} M$ beeing the identity (note that we are changing the notation slightly with respect to **Section** 2.3, indicating the module as a superscript instead of a subscript, leaving the subscript to denote the evaluation points in

the poset). As a consequence, $M^\dagger = \ker \eta^M = 0$, but again, this module is not projective, showing that 1b does not imply 1c.

Changing the poset now to be $T = \{a, b, c\}$ with the opposite relations $a \leq b$ and $c \leq b$, we can consider a very similar persistence module $N \in \mathsf{vec}^{\mathsf{T}}$ defined by

$$\Bbbk \xrightarrow{\text{id}} \Bbbk \xleftarrow{\text{id}} \Bbbk \ .$$

In this module, all maps are clearly epimorphisms, but the module itself is not an injective object, showing that 2a does not imply 2c. Similarly, we can use **Equation** (2.3.2) to see that $\lim_{\mathsf{T}} N = \Bbbk$ with all maps $\epsilon^N_\bullet : \lim_{\mathsf{T}} N \to N_\bullet$ being the identity map. This shows that $N^* = \operatorname{coker} \epsilon^N = 0$ but, as we have established, N is not injective, showing that 2b does not imply 2c.

2.6.2 Resolutions Under the Duality Functor

In this section we study *resolutions*, exact sequences of modules that are used to define invariants in homological algebra, and which have been capitalized in the theory of persistence to obtain topological invariants for multiparameter persistence modules, where the barcode is no longer a viable option.

Projective Covers, Injective Envelopes, and Resolutions

For an object $X \in \mathcal{C}$, a *projective cover* is given by a projective object P an a *superfluous epimorphism* $p : P \twoheadrightarrow X$, that is, an epimorphism such that any morphism $f : Y \to P$ with $p \circ f : Y \to X$ an epimorphism has to be an epimorphism too. Dually we can define an *injective envelope* as an injective object Q with an *essential embedding* $i : X \rightarrowtail Q$, i.e. a monomorphism such that given any morphism $g : Q \to Y$ with $g \circ i$ a monomorphism, then g has to be a monomorphism too.

A *projective resolution* of $M \in \mathsf{vec}^{\mathsf{T}}$ is an exact sequence

$$\ldots \to P_2 \xrightarrow{f_2} P_1 \xrightarrow{f_1} P_0 \xrightarrow{f_0} M \to 0 \tag{2.6.2}$$

where each P_i is projective; and an *injective resolution*, an exact sequence

$$0 \to M \xrightarrow{g_0} I_0 \xrightarrow{g_1} I_1 \xrightarrow{g_2} I_2 \to \ldots \tag{2.6.3}$$

where each I_i is injective. If we further assume that $f_0 : P_0 \to M$ and $f_i : P_i \to \ker f_{i-1}$ (resp. $g_0 : M \to I_0$ and $g_i : \operatorname{coker} f_{i-1} \to I_i$) are projective covers (resp. injective envelopes)

for all i, we call the resolutions *minimal*. Projective and injective resolutions are unique up to chain homotopy, while minimal projective and injective resolutions are unique up to isomorphisms.

The Duality Functor Revisited

The duality functor D_{T} in **Definition** 2.1.1 is an equivalence of categories, also termed, due to its contravariant nature, a *duality* between the categories $\mathsf{vec}^{\mathsf{T}}$ and $\mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$. We say that a contravariant functor $F : \mathcal{C} \to \mathcal{C}'$ is a duality between the categories $\mathcal{C}$ and $\mathcal{C}'$ if there exists another contravariant functor $G : \mathcal{C}' \to \mathcal{C}$ such that $F \circ G \cong \mathrm{id}_{\mathcal{C}}$ and $GF \cong \mathrm{id}_{\mathcal{C}'}$, where $\cong$ in this context means that the functors are naturally isomorphic.

The fact that D_{T} is a duality reflects a key difference in the connection between a finite-dimensional vector space U, its dual $U^\vee = \mathrm{Hom}_{\Bbbk}(U, \Bbbk)$ and its double dual $(U^\vee)^\vee$. This was in fact the motivating example in the seminal paper [EM45a], which is regarded as the foundational document on category theory and which introduced the concept of natural transformation. We explain these differences and provide a bit more of context next.

Since U and $U^\vee$ are both finite-dimensional with the same dimension, we know that they are isomorphic. However, this isomorphism cannot be established until we have chosen a basis in U, which in turn defines a *dual basis* in $U^\vee$; and different choices of bases entail different isomorphisms. This is not the case with the relation between U and its double dual $(U^\vee)^\vee$, where we can define a basis-free isomorphism $\xi_U : U \to (U^\vee)^\vee$ in the following way: for each $u \in U$, we define $\xi_U(u)$ as the linear form that acts on $\omega \in U^\vee$ by $\xi_U(u)\,(\omega) \coloneqq \omega(u)$. Since this assignment is natural in U, it defines a natural isomorphism $((-)^\vee)^\vee \cong \mathrm{id}_{\mathsf{vec}}$.

This natural isomorphism can be extended pointwise to diagrams of vector spaces to prove that the duality functor is indeed a duality, $D_{\mathsf{T}^{\mathrm{op}}} D_{\mathsf{T}} \cong \mathrm{id}_{\mathsf{vec}^{\mathsf{T}}}$. In a slight abuse of notation, we call ξ_M to the isomorphim of persistence modules connecting M and $D_{\mathsf{T}^{\mathrm{op}}} D_{\mathsf{T}} M$, for $M \in \mathsf{vec}^{\mathsf{T}}$. If the posets are clear for the context, we will simply write D^2 instead of $D_{\mathsf{T}^{\mathrm{op}}} D_{\mathsf{T}}$ to denote the composition of dualities. We note that a morphism of persistence modules $f : M \to N$ is sent to a morphism between their duals $Df : DN \to DM$ defined as $Df(\varphi) = \varphi \circ f$ for $\varphi \in DN$, and to a morphism between the double duals $D^2(f) : D^2M \to D^2N$ defined as $D^2(f) = (\xi_N)^{-1} \circ f \circ (\xi_M)^{-1}$. We also have the following lemma.

Lemma 2.6.6. *Let $M \in \mathsf{vec}^{\mathsf{T}}$ and $\xi_M : M \to D^2M$ the isomorphism to the double dual. Then, $D\xi_M = (\xi_{DM})^{-1}$.*

Proof. We check that $D\xi_M \circ \xi_{DM} = \mathrm{id}_{DM}$. For that, let $\alpha \in DM$, then $D\xi_M \circ \xi_{DM}(\alpha) = D\xi_M(\xi_{DM}(\alpha)) = \xi_{DM}(\alpha) \circ \xi_M$, which is an element of DM. Then, for $m \in M$, we have $(\xi_{DM}(\alpha) \circ \xi_M)\,(m) = \xi_{DM}(\alpha)\,(\xi_M(m)) = \xi_M(m)\,(\alpha) = \alpha(m)$, showing that $D\xi_M \circ \xi_{DM}(\alpha) = \alpha$

as wanted. Since ξ_{DM} is an isomorphism and the inverse is unique, we conclude that $(\xi_{DM})^{-1} = D\xi_M$. □

The functor D is contravariant exact, since exactness in $\mathsf{vec}^{\mathsf{T}}$ is equivalent to pointwise exactness in vec, and $(-)^\vee = \mathrm{Hom}_{\Bbbk}(-, \Bbbk) : \mathsf{vec} \to \mathsf{vec}$ is exact since $\Bbbk$ is injective. As a consequence of exactness and being contravariant, D sends monomorphism to epimorphisms, epimorphisms to monomorphisms, kernels to cokernels, and cokernels to kernels. It also sends limits to colimits and vice versa. We now prove that the duality sends projective objects to injective objects, and injectives to projectives. Since we are now working with two posets, we make explicit which underlying poset we are considering in our notations. For instance, $[t, \infty)_{\mathrm{op}}$ refers to the upset in T^{op}, that is $\{s \in T : t \leq_{\mathrm{op}} s\}$. Observe that, from this definition, $[t, \infty)_{\mathrm{op}} = (-\infty, t]$.

Proposition 2.6.7. *For every $t \in T$, we have $D\Bbbk_{[t,\infty)} = \Bbbk_{(-\infty,t]_{\mathrm{op}}}$. This means that D restricts to a duality* $\mathrm{proj}\,\mathsf{T} \to \mathrm{inj}\,\mathsf{T}^{\mathrm{op}}$*, between the projective modules over T and the injective modules over T^{op}. This duality maps projective covers of modules over T to injective envelopes over T^{op}.*

Proof. We apply **Definition** 2.1.1 to obtain $D\Bbbk_{[t,\infty)}$.

- For $s \in T^{\mathrm{op}}$ we have

$$D\Bbbk_{[t,\infty)}(s) = \mathrm{Hom}_{\Bbbk}(\Bbbk_{[t,\infty)}(s), \Bbbk) = \begin{cases} \mathrm{Hom}_{\Bbbk}(\Bbbk, \Bbbk) \simeq \Bbbk, & \text{if } t \leq s \iff s \leq_{\mathrm{op}} t; \\ 0, & \text{otherwise.} \end{cases}$$

- For $r \leq_{\mathrm{op}} s$ and $u \in D\Bbbk_{[t,\infty)}(r) = \mathrm{Hom}_{\Bbbk}(\Bbbk_{[t,\infty)}(r), \Bbbk)$, we have

$$D\Bbbk_{[t,\infty)}(r \leq_{\mathrm{op}} s)(u) = u \circ \Bbbk_{[t,\infty)}(s \leq r) = \begin{cases} u \circ \mathrm{id}_{\Bbbk}, & \text{if } t \leq s \leq r; \\ 0, & \text{otherwise.} \end{cases}$$

  This shows that $D\Bbbk_{[t,\infty)}(r \leq_{\mathrm{op}} s) \cong \mathrm{id}_{\Bbbk}$ if $r \leq_{\mathrm{op}} s \leq_{\mathrm{op}} t$ and 0 otherwise.

These computations explicitly show that $D\Bbbk_{[t,\infty)} = \Bbbk_{(-\infty,t]_{\mathrm{op}}}$ as desired.

The second part of the statement follows from this equivalence and the fact that any projective (resp. injective) T-module (resp. T^{op}-module) can be written as a direct sum of projective (resp. injective) indecomposable modules.

Lastly, given $M \in \mathsf{vec}^{\mathsf{T}}$ a persistence module and a projective cover $P \in \mathrm{proj}\,\mathsf{T}$ with an epimorphism $g : P \twoheadrightarrow M$, we check that $DP \in \mathrm{inj}\,\mathsf{T}^{\mathrm{op}}$ with the map $Dg : DM \rightarrowtail DP$ (which is a monomorphism due to the exactness of D) is the injective envelope of DM. Let

$h : DP \to N$ be a morphism from DP to another diagram $N \in \mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$ such that $h \circ Dg$ is a monomorphism. We want to show that this implies that h is also a monomorphism. We dualize the construction to obtain a morphism $Dh : DN \to D^2P$ and an epimorphism $D^2g : D^2P \twoheadrightarrow D^2M$ such that $D^2g \circ Dh$ is an epimorphism. Using the canonical isomorphisms $D^2M \cong M$ and $D^2P \cong P$ and the fact that $g : P \twoheadrightarrow M$ is a projective cover, we obtain that $Dh : DN \to D^2P$ is an epimorphism, which allows us to conclude that $h : DP \to N$ had to be a monomorphism, as desired. □

Applying the Duality Functor to Resolutions

As mentioned at the beginnig, resolutions are interesting for us because they are used to define invariants of persistence modules. One of the most prominent examples is that of *multigraded Betti numbers*, which give us the degrees of a set of generators in a minimal projective resolution.

Definition 2.6.8 (**Multigraded Betti numbers**)**.** Let $M \in \mathsf{vec}^{\mathsf{T}}$ be a persistence module and let $\cdots \to P_2 \to P_1 \to P_0 \to M \to 0$ denote a minimal projective resolution. The kth *multigraded Betti numbers* of M are given by the collection $\beta^k_{\mathrm{mult}}M = \{t_i : P_k \cong \oplus_{i \in \mathcal{I}_k} \Bbbk_{[t_i,\infty)}\} \subseteq T$, which is a multiset. In other words, $\beta^k_{\mathrm{mult}}M$ is the family of degrees of the generators of the projective module P_k from the minimal projective resolution.

Remark 2.6.9. Multigraded Betti numbers can also be defined as a function

$$\beta^k_{\mathrm{mult}}M : \mathrm{ind}(\mathrm{proj}\,\mathsf{T}) \to \mathbb{N},$$

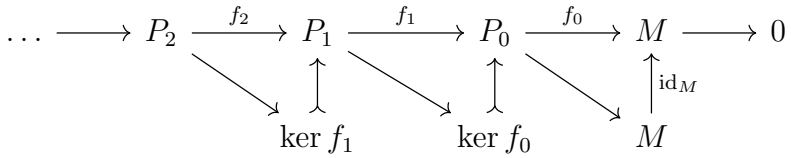
where $\mathrm{ind}(\mathrm{proj}\,\mathsf{T})$ denotes the class of indecomposable projectives in $\mathsf{vec}^{\mathsf{T}}$, such that

$$\bigoplus_{P \in \mathrm{ind}(\mathrm{proj}\,\mathsf{T})} P^{\beta^k_{\mathrm{mult}}M(P)}.$$

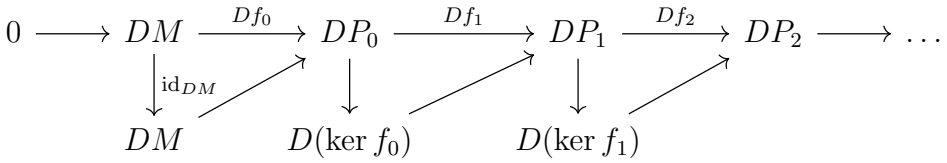
We note here that the uniqueness up to isomorphism of the minimal projective resolution guarantees that the multigraded Betti numbers are invariants of the isomorphism type of M. As a direct consequence of **Proposition** 2.6.7, we have the following connection between projective and injective resolutions.

Corollary 2.6.10. *The duality in* ***Equation*** (2.1.1) *maps the minimal projective resolution of M over T to the minimal injective resolution of DM over T^{op}.*

Proof. Let $M \in \mathsf{vec}^{\mathsf{T}}$ and the minimal projective resolution



where each P_i is a projective module and the diagonal maps are projective covers. Applying the duality in **Equation** (2.1.1) we get a diagram



where, thanks to **Proposition** 2.6.7, we know that the DP_i are all injective modules and the diagonal maps are injective envelopes. The only thing left to prove that this is a minimal injective resolution of DM over T^{op} is that $D(\ker f_i) \simeq \operatorname{coker} Df_i$, namely, that $D(\ker f)$ is a generalized cokernel, which follows from the exactness of D. □

Relative Homological Algebra and Dualities

To close this chapter, we connect with a recent trend in persistence that aims to capitalize tools of *relative homological algebra*, generalizing the notions of projectives, resolutions and multigraded Betti numbers. We first recall the basics of relative homological algebra in exact categories necessary to understand this section. We follow [AS93], formulated for finitely generated modules over artin algebras, and [BBH23, §3 and §4], [BBH22, §4] and [Cha+24, §1], which adapt many of these tools to the theory of persistence.

Exact Categories. *Exact categories* were introduced in [Qui06] to extend the notion of an exact sequence to categories where kernels and cokernels do not necessarily exist. An *exact category* is an additive category endowed with an *exact structure*, namely, a distinguished family of *exact sequences* satisfying some axioms [Büh10, Definition 2.1] that replicate the desirable properties of usual short exact sequences.

Let $\mathcal{X}$ be a class of objects in $\mathsf{vec}^{\mathsf{T}}$ and $\mathcal{E}_{\max}$ the family of short exact sequences in $\mathsf{vec}^{\mathsf{T}}$. We define the following families of short exact sequences:

$$\mathcal{E}_{\mathcal{X}} \coloneqq \{(0 \to M \to N \to L \to 0) \in \mathcal{E}_{\max} : \forall R \in \mathcal{X},\ \mathrm{Hom}(R, -) \text{ is exact on the sequence}\} \tag{2.6.4}$$

and

$$\mathcal{E}^{\mathcal{X}} \coloneqq \{(0 \to M \to N \to L \to 0) \in \mathcal{E}_{\max} : \forall R \in \mathcal{X},\ \mathrm{Hom}(-, R) \text{ is exact on the sequence}\}\,. \tag{2.6.5}$$

It turns out that, for any class of objects $\mathcal{X}$ of $\mathsf{vec}^{\mathsf{T}}$ these define exact structures [AS93, Proposition 1.7]. If $\mathcal{X}$ is additive, that is, it is closed under direct sums, we have that $\mathcal{F}_{\mathcal{X}} = \mathcal{F}_{\mathrm{ind}(\mathcal{X})}$ (and analogously with $\mathcal{F}^{\mathcal{X}}$), so it is enough to formulate the definitions above for classes of indecomposable objects.

Approximations. Denote by $\mathrm{add}(\mathcal{X})$ the full subcategory of $\mathsf{vec}^{\mathsf{T}}$ with objects isomorphic to finite direct sums of modules in $\mathcal{X}$. A *right* $\mathrm{add}(\mathcal{X})$*-approximation* of a module $M \in \mathsf{vec}^{\mathsf{T}}$ is a morphism $f : R \to M$ with $R \in \mathrm{add}(\mathcal{X})$ such that $\mathrm{Hom}(-, R) \to \mathrm{Hom}(-, M) \to 0$ is exact. In other words, we want $\mathrm{Hom}(-, R) \to \mathrm{Hom}(-, M)$ to be surjective, meaning that any other morphism $R' \to M$ with $R' \in \mathrm{add}(\mathcal{X})$ factors through f. A *minimal* right approximation is a right $\mathrm{add}(\mathcal{X})$-approximation where any endomorphism $g : R \to R$ such that $f = f \circ g$ is an isomorphism.

Dually, a *left* $\mathrm{add}(\mathcal{X})$*-approximation* of a module $M \in \mathsf{vec}^{\mathsf{T}}$ is a morphism $f : M \to R$ with $R \in \mathrm{add}(\mathcal{X})$ such that $\mathrm{Hom}(R, -) \to \mathrm{Hom}(M, -) \to 0$ is exact. We say that $\mathrm{add}(\mathcal{X})$ is contravariantly finite (resp. covariantly finite) if every object $M \in \mathsf{vec}^{\mathsf{T}}$ has a right $\mathrm{add}(\mathcal{X})$-approximation (resp. left $\mathrm{add}(\mathcal{X})$-approximation).

It is easy to check that when $\mathcal{X}$ contains all indecomposable projective modules, any right $\mathrm{add}(\mathcal{X})$-approximation is an epimorphism [BBH22, Lemma 4.2]. The dual holds for injectives, as we prove next.

Lemma 2.6.11. *Assume $\mathcal{X}$ contains all the indecomposable injectives and $f : M \to R$ is a left* $\mathrm{add}(\mathcal{X})$*-approximation. Then, f is a monomorphism.*

Proof. We take $i : M \to Q$ the injective envelope of M, which is a monomorphism. By assumption, we have that $Q \in \mathrm{add}(\mathcal{X})$, and f being a left approximation means that i needs to factor through it, that is, there is some $g : R \to Q$ such that $i = g \circ f$. Since i was a monomorphims, this implies that f has to be a monomorphism too. □

Relative Resolutions. As a consequence of the results above, we can form analogues of projective and injective resolutions by taking right and left $\mathrm{add}(\mathcal{X})$-approximations at each step, respectively. For the right $\mathrm{add}(\mathcal{X})$-approximation, this yields the following diagram

where p_j is such an approximation and $q_j = i_{j-1} \circ p_j$

$$\begin{array}{ccccccccccccc} \dots & \longrightarrow & R_2 & \xrightarrow{q_2} & R_1 & \xrightarrow{q_1} & R_0 & \xrightarrow{q_0} & M & \xrightarrow{q_{-1}=0} & 0. \\ & \searrow & \uparrow{\scriptstyle i_2} & \nearrow_{p_2} & \uparrow{\scriptstyle i_1} & \nearrow_{p_1} & \uparrow{\scriptstyle i_0} & \nearrow_{p_0} & \uparrow{\scriptstyle i_{-1}=\mathrm{id}_M} & & \\ & & \ker(q_2) & & \ker(q_1) & & \ker(q_0) & & \ker(q_{-1}) = M & & \end{array} \tag{2.6.6}$$

The first row in this diagram is called a *right* $\mathrm{add}(\mathcal{X})$*-resolution* and it is exact (in the exact structure) by definition. If we additionally ask the approximations in the diagram to be minimal, the resolution is also said to be minimal. We can generalize definition **Definition** 2.6.8 to define *relative Betti numbers* or *relative Betti diagrams* (see [Cha+24, §2.6] for more details).

Dualities on Relative Resolutions. Working with exact structures and their associated resolutions, we can adapt **Corollary** 2.6.10 to obtain the following.

Lemma 2.6.12. *Let $\mathcal{X}$ be a class of objects in $\mathsf{vec}^{\mathsf{T}}$ and $D\mathcal{X}$ the corresponding class of dual objects in $\mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$. For $M \in \mathsf{Vec}^{\mathsf{T}}$, if*

$$\cdots \to X_2 \xrightarrow{f_2} X_1 \xrightarrow{f_1} M$$

is a right $\mathrm{add}(\mathcal{X})$*-resolution in $\mathsf{vec}^{\mathsf{T}}$, then*

$$DM \xrightarrow{Df_1} DX_1 \xrightarrow{Df_2} DX_2 \to \dots$$

is a left $\mathrm{add}(D\mathcal{X})$*-resolution in $\mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$.*

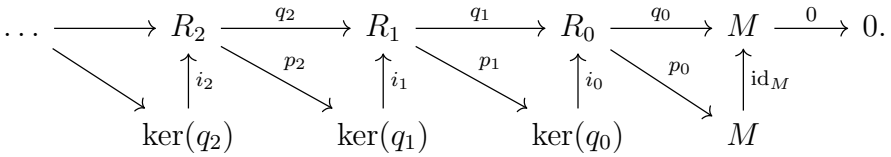
Proof. We first check that, for $M, N \in \mathsf{vec}^{\mathsf{T}}$, if $f : M \to N$ is a right $\mathrm{add}(\mathcal{X})$-approximation, then $Df : DN \to DM$ is a left $\mathrm{add}(\mathcal{X})$-approximation (in $\mathsf{vec}^{\mathsf{T}^{\mathrm{op}}}$).

Let $\varphi : DN \to L$ for some $L \in \mathrm{add}(D\mathcal{X})$. We want to prove that φ factors through Df, that is, to find some $\phi : DM \to L$ such that $\varphi = \phi \circ Df$. For that, we consider $D\varphi : DL \to D^2N$ and precompose with the inverse of the isomorphism $\xi_N : N \to D^2N$, to get $(\xi_N)^{-1} \circ D\varphi : DL \to N$. Since we know that $f : M \to N$ is a right $\mathrm{add}(\mathcal{X})$-approximation and $D^L \in \mathrm{add}(\mathcal{X})$, we have that there is some $\eta : DL \to M$ such that $f \circ \eta = (\xi_N)^{-1} \circ D\varphi$, or equivalently, $\xi_N \circ f \circ \eta = D\varphi$. Again, applying the duality, we arrive to

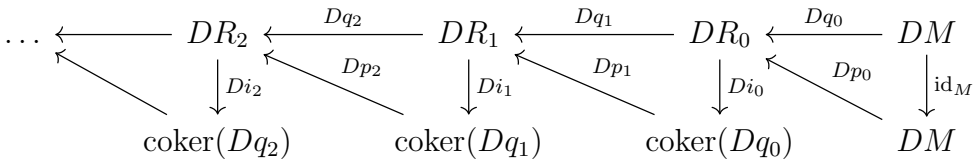
$$D^2\varphi = D\eta \circ Df \circ D\xi_N.$$

Using that $D^2\varphi = \xi_L \circ \varphi \circ (\xi_{DN})^{-1}$ and $D\xi_N = (\xi_{DN})^{-1}$, substituting in the equation above we obtain $\xi^L \circ \varphi \circ (\xi_{DN})^{-1} = D\eta \circ Df \circ (\xi_{DN})^{-1}$ and thus $\varphi = (\xi_L)^{-1} \circ D\eta \circ Df$. Setting $\phi = (\xi_L)^{-1} \circ D\eta$ we conclude the proof that Df is a left $\mathrm{add}(D\mathcal{X})$-approximation.

Now, recalling **Equation** (2.6.6), a right add($\mathcal{X}$)-resolution of $M \in \mathsf{vec}^{\mathsf{T}}$ is obtained by iteratively taking right add($\mathcal{X}$)-aproximations and kernels in the following diagram



Taking dualities, by the exactness of D, kernels are mapped to cokernels. We have also just shown that the right approximations q_i are turned into left approximations Dq_i. We conclude that the following diagram



is a left add($D\mathcal{X}$)-approximation of DM, as desired.

□

The natural question that arises from these observations, which we leave for future work, is how to connect the projective (right) and injective (left) resolutions in these categories, to obtain potential relations or equivalences between the invariants that we can define from them.

3 Functional Summaries for Statistical and Machine Learning Integration

This chapter summarizes my contributions to two joint works: [Wan+24] and [GMW25], which focus on two functional summaries obtained from the PH pipeline: the rank function (**Definition** 1.4.1) and the persistence landscape (**Definition** 1.4.2), respectively. This functional perspective allows to address gaps in the integration of PH with statistical inference techniques. Once a solid statistical foundation is in place, ML methods can be used to develop practical applications of these tools.

The contributions of this chapter include stability results for rank functions (in the single-parameter setting, see **Section** 3.1) and rank invariants (in the multiparameter case, see **Section** 3.2), enabling the application of functional data analysis techniques on these summaries. In addition, we extend a uniform convergence theorem for persistence landscapes to the multiparameter setting (**Section** 3.3.1), and use this result to construct confidence bands for multiparameter persistence landscapes (**Section** 3.3.2), a statistical tool that serves as an inferential counterpart to hypothesis testing.

3.1 Rank Function Stability With Respect to L^p Distances

In [Wan+24], our goal was to address the limitations of barcodes discussed in **Section** 1.3 by considering an alternative invariant within the PH pipeline. We focused on *rank functions*, which, in the case of single-parameter persistence, is another *complete invariant* of persistence modules, alongside barcodes and diagrams. This means that two persistence modules M, $N \in \mathsf{Vec}^{\mathbb{R}}$ have the same rank function if and only if they are isomorphic. In contrast, other vectorizations of barcodes are instead *incomplete* invariants: non-isomorphic modules can map to the same representation.

Rank functions are not only mathematically equivalent to barcodes, but they also offer

additional analytical advantages: as functions, they naturally lend themselves to tools from functional data analysis (FDA). FDA is a well-established branch of statistics that operates on function-valued data.

My primary contribution in [Wan+24] was to address a gap in the literature that prevented the implementation of these methods with rank functions: its stability with respect to the L^p distance (cf. **Equation** (1.5.6)). We focused on this metric space because, for $p = 2$, the space of rank functions becomes a Hilbert space, satisfying the foundational assumptions required by several FDA methods used in the experimental component of the project, see [Wan+24, §4.2 and §5.2]. The goal is to obtain stability bounds for this metric with respect to other known stable metrics in the theory of persistence.

3.1.1 Some Observations About the L^p Distance Between Two Rank Functions

Let M, $N \in \mathsf{Vec}^{\mathbb{R}}$ be two persistence modules. In **Section** 1.5 we saw that the bottleneck and Wasserstein distances between their corresponding persistence diagrams $\mathrm{PD}(M) = \{(b_i, d_i) : i \in \mathcal{I}\}$ and $\mathrm{PD}(N) = \{(\widetilde{b}_j, \widetilde{d}_j) : j \in \mathcal{J}\}$ were defined in terms of bijections between the points in them, where some points where matched to the diagonal. For the remainder of this section, we assume that both diagrams have a finite number of points, i.e. $\mathcal{I}$ and $\mathcal{J}$ are finite, and that the bars in the barcodes are finite, i.e. $d_i, \widetilde{d}_j < \infty$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$.

The bijection in the definition of the bottleneck and Wasserstein distances can be reformulated as a partial matching $\phi : \mathcal{K} \to \mathcal{J}$, from a subset $\mathcal{K} \subseteq \mathcal{I}$ to a subset $\phi(\mathcal{K}) \subseteq \mathcal{J}$ such that each $(b_k, d_k) \in \mathrm{PD}(M)$ is mapped to some $(\widetilde{b}_{\phi(k)}, \widetilde{d}_{\phi(k)}) \in \mathrm{PD}(N)$, and all the points (b_i, d_i) for $i \in \mathcal{I} \setminus \mathcal{K}$ and $(\widetilde{b}_j, \widetilde{d}_j) \in \mathcal{J} \setminus \phi(\mathcal{I})$ are matched to the diagonal. Considering one such matching, the L^p distance between the rank functions can be decomposed in the following way

$$\begin{aligned}\left\|\mathrm{rk}^M - \mathrm{rk}^N\right\|_p &\leq \left\| \sum_{k\in\mathcal{K}} \mathrm{rk}^{\Bbbk[b_k,d_k)} - \mathrm{rk}^{\Bbbk\left[\widetilde{b}_{\phi(k)},\widetilde{d}_{\phi(k)}\right)} + \sum_{i\in\mathcal{I}\setminus\mathcal{K}} \mathrm{rk}^{\Bbbk[b_i,d_i)} + \sum_{j\in\mathcal{J}\setminus\phi(\mathcal{K})} \mathrm{rk}^{\Bbbk\left[\widetilde{b}_j,\widetilde{d}_j\right)} \right\|_p \\ &\leq \sum_{k\in\mathcal{K}} \left\| \mathrm{rk}^{\Bbbk[b_k,d_k)} - \mathrm{rk}^{\Bbbk\left[\widetilde{b}_{\phi(k)},\widetilde{d}_{\phi(k)}\right)} \right\|_p + \sum_{i\in\mathcal{I}\setminus\mathcal{K}} \left\| \mathrm{rk}^{\Bbbk[b_i,d_i)} \right\|_p + \sum_{j\in\mathcal{J}\setminus\phi(\mathcal{K})} \left\| \mathrm{rk}^{\Bbbk\left[\widetilde{b}_j,\widetilde{d}_j\right)} \right\|_p\end{aligned} \tag{3.1.1}$$

where we use the decomposition in interval modules of $M \cong \bigoplus_{i\in\mathcal{I}} \Bbbk_{[b_i,d_i)}$ and $N \cong \bigoplus_{j\in\mathcal{J}} \Bbbk_{[\widetilde{b}_j,\widetilde{d}_j)}$, the additivity of the rank function and the triangle inequality. The difference of rank functions of the left term in the sum in **Equation** (3.1.1) results in an indicator function over the

region D_k, for $k \in \mathcal{K}$, defined by

$$\begin{aligned} A_k &:= \{(x, y) \in \mathbb{R}^2 : x \leq y \text{ and } b_k \leq x \leq y \leq d_k\}, \\ B_k &:= \{(x, y) \in \mathbb{R}^2 : x \leq y \text{ and } \widetilde{b}_{\phi(k)} \leq x \leq y \leq \widetilde{d}_{\phi(k)}\}, \\ D_k &:= (A_k \cup B_k) \setminus (A_k \cap B_k). \end{aligned} \tag{3.1.2}$$

See **Figure** 3.1a for an illustrative example. The rank functions of the two right terms in the sum in **Equation** (3.1.1) are simply indicator functions with support the triangles over the diagonal with cusps (b_i, d_i) and $(\widetilde{b}_j, \widetilde{d}_j)$, respectively (see **Figure** 3.1b for an example).

(a) Difference between rank functions of two matched points.

(b) Rank function of unmatched points.

Figure 3.1: Rank functions for matched (left) and unmatched (right) points between diagrams. Yellow regions represent the support of indicator functions, each taking value 1 on the shaded area. The left diagram illustrates the difference between rank functions at matched points, while the right shows the contribution of an unmatched point.

The L^p norm over the latter can be obtained from the area of the triangles, that is,

$$\left\| \mathrm{rk}^{\mathbb{k}[b_i, d_i)} \right\|_p = \left(\frac{1}{2}(d_i - b_i)^2 \right)^{1/p}$$

for all unmatched points (b_i, d_i) with $i \in \mathcal{I} \setminus \mathcal{K}$, and equivalently for the unmatched points (b_j, d_j) with $j \in \mathcal{J} \setminus \phi(\mathcal{K})$. For the matched points, we have

$$\left\| \mathrm{rk}^{\mathbb{k}[b_k, d_k)} - \mathrm{rk}^{\mathbb{k}[\widetilde{b}_{\phi(k)}, \widetilde{d}_{\phi(k)})} \right\|_p = \mu(D_k)^{1/p}$$

for $k \in \mathcal{K}$, where μ denotes the Lebesgue measure of the region D_k defined above. In [Wan+24], we prove the following bound.

Lemma 3.1.1. *For $D_k \subset \mathbb{R}^2$ defined as in **Equation** (3.1.2) from the matching $(b_k, d_k) \mapsto (\widetilde{b}_{\phi(k)}, \widetilde{d}_{\phi(k)})$ we have*

$$\mu(D_k) \leq 2(R+1) \left\| (b_k, d_k) - (\widetilde{b}_{\phi(k)}, \widetilde{d}_{\phi(k)}) \right\|_\infty$$

where $R = \max\{d_k - b_k : k \in \mathcal{K}\}$ *and* $\|\cdot\|_\infty$ *denotes the* ℓ^∞ *distance in the plane.*

3.1.2 Stability Under the Bottleneck Distance

The terms in the sum in **Equation** (3.1.1) corresponding to unmatched points preclude the possibility of bounding the L^p distance between two rank functions by the bottleneck distance among their diagrams. To obtain a bound relating these two metrics, we need to eliminate those points by considering *truncated rank functions*.

Definition 3.1.2 (**Truncated rank function).** For any rank function rk and any $\delta > 0$, we define the *δ-truncated rank function* as

$$\mathrm{rk}_\delta := \mathrm{rk} \cdot \mathbb{1}_{\mathbb{R}^2_\delta}$$

where $\mathbb{1}_{\mathbb{R}^{2+}_\delta}$ is the indicator function of the set $\mathbb{R}^2_\delta := \{(x, y) \in \mathbb{R}^2 \,:\, y > x + \delta\}$.

Thanks to **Lemma** 3.1.1, we can prove the following stability bound for truncated rank functions.

Proposition 3.1.3 (Proposition 10, [Wan+24])**.** *Let* $1 \leq p < \infty$ *and* $M \in \mathsf{vec}^{\mathbb{R}}$ *be a p.f.d. persistence module with finite intervals in its barcode decomposition. For every* $\delta > 0$*, there exist* $1 \geq \eta > 0$ *such that any persistence module* $N \in \mathsf{vec}^{\mathbb{R}}$ *satisfying*

$$d_B(\mathrm{PD}(M), \mathrm{PD}(N)) < \eta$$

also satisfies

$$\left\|\mathrm{rk}^M_\delta - \mathrm{rk}^N_\delta\right\|_p \leq m(2R + 2)^{1/p} \cdot d_B(\mathrm{PD}(M), \mathrm{PD}(N))^{1/p} \tag{3.1.3}$$

where m *is the number of points in* $\mathrm{PD}(M)$ *and* $R = \max\{d - b : (b, d) \in \mathrm{PD}(M)\}$ *is the maximum persistence over the points in* $\mathrm{PD}(M)$*.*

Remark 3.1.4. We note that we obtain a stability bound where the constant only depends in one of the modules involved. Yet, it depends in the number of points of this persistence module, which might grow considerably. Moreover, this bound satisfies a weaker Hölder continuity condition, with exponent $1/p$, rather than the standard Lipschitz continuity found in other stability results in the literature.

3.1.3 Stability Under the Wasserstein Distance

Until very recently, **Proposition** 3.1.3 was the best stability bound we could hope for, as the bottleneck distance was the only metric over persistence diagrams known to be stable.

However, thanks to the Cellular Wasserstein Stability (**Theorem** 1.5.3) by Skraba and Turner [ST21], we can now also conclude stability for the rank functions with the L^p distance by comparing to the Wasserstein distance between the corresponding diagrams, which is a much more natural metric to use in the comparison.

Theorem 3.1.5 (Theorem 12, [Wan+24])**.** *Let $p = 1, 2$; and $M \in \mathsf{vec}^{\mathbb{R}}$ be a p.f.d. persistence module with finite intervals in its barcode decomposition. For any other p.f.d. persistence module N satisfying $W_1(\mathrm{PD}(M), \mathrm{PD}(N)) \leq 1$, we have*

$$\left\|\mathrm{rk}^M - \mathrm{rk}^N\right\|_1 \leq (2R+2) \cdot W_1(\mathrm{PD}(M), \mathrm{PD}(N)) \tag{3.1.4}$$

and

$$\left\|\mathrm{rk}^M - \mathrm{rk}^N\right\|_2 \leq 2\max\left\{\sqrt{(2R+2)m}, \frac{1}{\sqrt{2}}\right\} \cdot W_1(\mathrm{PD}(M), \mathrm{PD}(N))^{1/2} \tag{3.1.5}$$

where m is the number of points in $\mathrm{PD}(M)$ and $R = \max\{d - b : (b, d) \in \mathrm{PD}(M)\}$ is the maximum persistence over the points in $\mathrm{PD}(M)$.

Remark 3.1.6. The proof for $p = 1$ relies on two key observations. First, for unmatched points, we can bound the L^1 norm of the corresponding rank functions using

$$\frac{1}{2}(d_i - b_i)^2 \leq d_i - b_i,$$

which holds because $\frac{1}{2}(d_i - b_i) \leq 1$ under the assumption $W_1(\mathrm{PD}(M), \mathrm{PD}(N)) \leq 1$. Additionally, we use the fact that the ℓ^∞ norm appearing in the bound from **Lemma** 3.1.1 is controlled above by the ℓ^1 norm.

For the case $p = 2$, we apply the Cauchy–Schwarz inequality:

$$\sum_{k \in \mathcal{K}} \mu(D_k)^{1/2} \leq \left(m \cdot \sum_{k \in \mathcal{K}} \mu(D_k)\right)^{1/2},$$

from which the desired bound follows by straightforward computation.

3.2 Rank Invariant Stability

In this section, we overview some results that do not appear in the published version of the paper [Wan+24], but which we included in a previous preprint that can be found in the arXiv [Wan+23]. These are extensions of the stability results in **Section** 3.1 to the setting of multiparameter persistence (**Section** 3.2.1) and some additional observations concerning an alternative, theoretical metric that we can define over the rank invariants (**Section** 3.2.2).

3.2.1 Stability Under the Wasserstein Distance

The ideas from **Section** 3.1.3 can be extended to the setting of multiparameter persistence in the following way. A rectangle $U \subset \mathbb{R}^n$ is defined as the product of n intervals, $U = [a_1, b_1] \times \cdots \times [a_n, b_n]$, or equivalently, is the subset $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$ with $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$. For p.f.d. modules supported on rectangles, we have the following property.

Lemma 3.2.1 (Lemma 28, [Wan+23]). *Let $U, V \subset \mathbb{R}^n$ be rectangles. If $M = \Bbbk_U$ and $N = \Bbbk_V$ are ϵ-interleaved, then*

$$\left\|\mathrm{rk}^M - \mathrm{rk}^N\right\|_p \leq \left(2n(C^{2n-1} + \tilde{C}^{2n-1}) \cdot \epsilon\right)^{1/p}$$

with $C > 0$ the maximum length over the intervals defining U, and $\tilde{C} > 0$ the maximum length over the intervals defining V.

This allows to prove the following.

Proposition 3.2.2 (Proposition 27, [Wan+23]). *Let $M, N \in \mathsf{vec}^{\mathbb{R}^n}$ be rectangle decomposable modules. Then, for $p = 1, 2$ there exists $c_{M,N,p,n} > 0$ such that*

$$\left\|\mathrm{rk}^M - \mathrm{rk}^N\right\|_p \leq c_{M,N,p,n} \cdot W_p(M, N)^{1/p}$$

with W_p defined as in ***Equation*** (1.5.5).

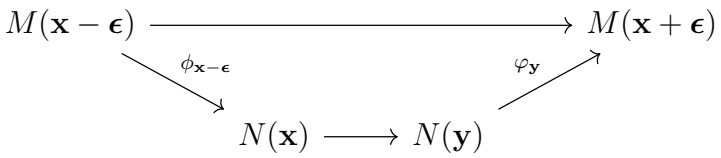
3.2.2 The Function-Interleaving Distance

We now shift the focus to stability results that move beyond our original motivation of integration with FDA. In the theory of persistence, strong stability results are typically expressed as Lipschitz conditions with a constant of one. In contrast, the results presented thus far for the L^p metrics and rank functions are Hölder inequalities, where the constants depend on parameters derived from the barcodes.

The interleaving distance (cf. **Definition** 1.5.5), defined over persistence modules, is well-established in persistence theory due to its desirable theoretical properties and its utility in proving stability bounds. Motivated by this intuition, we investigate the possibility of extending the definition of the interleaving distance to accommodate functions, and subsequently, determine the stability properties of such a metric over the rank invariants.

Let $M, N \in \mathsf{vec}^{\mathbb{R}^n}$ be two multiparameter persistence modules that are ϵ-interleaved, and call $\boldsymbol{\epsilon} = \epsilon(1, \ldots, 1)$. From the ϵ-interleaving condition (cf. **Definition** 1.5.4) we have the following

commutative diagram for all $\mathbf{x} \leq \mathbf{y} \in \mathbb{R}^n$



where the horizontal maps are the internal maps of the persistence modules and the vertical maps are the ones prescribed by the interleaving. This diagram implies that $\mathrm{rk}^N(\mathbf{x}, \mathbf{y}) \geq \mathrm{rk}^M(\mathbf{x} - \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\epsilon})$. We can obtain an analogous diagram interchanging the roles of M and N, which justifies the definition of the following distance, which already appeared in [MWW22, Definition 5.7].

Definition 3.2.3 (**Function interleaving distance**)**.** Let $f, g : \Omega \times \Omega \to \mathbb{R}$ be two real-valued functions, with $\Omega \subset \mathbb{R}^n$. We define the *function-interleaving distance* between f and g by

$$d_{\mathrm{FI}}(f,g) = \inf\{\epsilon \geq 0 \,:\, f(\mathbf{x},\mathbf{y}) \geq g(\mathbf{x}-\boldsymbol{\epsilon},\, \mathbf{y}+\boldsymbol{\epsilon}) \text{ and } g(\mathbf{x},\mathbf{y}) \geq f(\mathbf{x}-\boldsymbol{\epsilon},\, \mathbf{y}+\boldsymbol{\epsilon}),\ \forall\, (\mathbf{x},\mathbf{y}) \in \Omega\times\Omega\}$$

where $\epsilon = \|\boldsymbol{\epsilon}\|_\infty$ for $\boldsymbol{\epsilon} \in \mathbb{R}^n$, and $d_{\mathrm{FI}}(f,g) = \infty$ if such $\epsilon \geq 0$ does not exist.

Endowing the rank invariants with such a metric, we have the following stability condition.

Theorem 3.2.4 (Theorem 32, [Wan+23])**.** *Let $M, N \in \mathsf{vec}^{\mathbb{R}^n}$ be two persistence modules; rk^M, rk^N their associated rank invariants; and λ^M, λ^N their persistence landscapes. We have*

$$\left\|\lambda^M - \lambda^N\right\|_\infty \leq d_{\mathrm{FI}}(\mathrm{rk}^M, \mathrm{rk}^N) \leq d_I(M,N).$$

where $\|\cdot\|_\infty$ denotes the L^∞ norm over the landscapes $\lambda^N, \lambda^N : \mathbb{N} \times \mathbb{R}^n \to \mathbb{R}$.

The second inequality of **Theorem** 3.2.4 is a direct consequence of the observation at the beginning of this section. Proving the first inequality requires a bit more work. We note that [Vip20, Theorem 30] proves $\left\|\lambda^M - \lambda^N\right\|_\infty \leq d_I(M,N)$. In the single-parameter persistence case, we can strengthen the previous result to assert that the map $\mathrm{rk}^M \mapsto \lambda^M$ is an isometry when considering the function interleaving distance and the L^∞ distance for the landscapes.

Proposition 3.2.5 (Proposition 33, [Wan+23])**.** *Let $M, N \in \mathsf{vec}^{\mathbb{R}}$ be two persistence modules; rk^M, rk^N their rank functions and λ^M, λ^N their persistence landscapes. Then, we have that $d_{\mathrm{FI}}(\mathrm{rk}^M, \mathrm{rk}^N) \leq \left\|\lambda^M - \lambda^N\right\|_\infty$ where $\|\cdot\|_\infty$ denotes the L^∞ norm over the landscapes $\lambda^N, \lambda^N : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$.*

3.3 Confidence Bands for Multiparameter Landscapes

We now review the contents of my joint work [GMW25], which has been accepted to appear in the Springer Lecture Notes in Computer Science proceedings of the 7th International Conference of Geometric Science of Information. In this contribution, we extended the statistical framework of confidence bands from single-parameter [Cha+14a; Cha+14b] to multiparameter persistence landscapes. This includes a functional central limit theorem (FCLT) for multiparameter landscapes, which justifies a subsequent bootstrapping approach to compute confidence bands for them.

3.3.1 Functional Central Limit Theorem for Multiparameter Persistence Landscapes

We begin discussing the known convergence of single- and multiparameter landscapes as random variables in a Banach space given by central limit theorems to conclude presenting our own addition to this line of work.

Let $\mathcal{L}$ denote the space of *persistence landscapes* $\lambda : [0,T]^n \to \mathbb{R}$, where each λ is constructed from a persistence module $M \in \mathsf{vec}^{\mathbb{R}^n}$ via $\lambda(\mathbf{x}) = \lambda^M(m, \mathbf{x})$ for some $m \in \mathbb{N}$. We restrict our attention to landscapes defined on a bounded domain $[0,T]^n \subset \mathbb{R}^n$, so that each landscape lies in the Banach space $L^p([0,T]^n)$ for some $1 \leq p \leq \infty$. Thus, we view $\mathcal{L}$ as a subset of $L^p([0,T]^n)$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\Lambda : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathcal{L}$ be a random variable taking values in the space of persistence landscapes $\mathcal{L}$. The law of Λ is obtained as follows. We assume that our input data is sampled from an unknown data-generating distribution μ. After fixing the choices for constructing a filtration $F_\bullet$ from the data, we compute homology to obtain a persistence module. Define $\Phi : \text{Data space} \to \mathsf{Vec}^{\mathbb{R}^n}$ to be the map that assigns to each sample the corresponding persistence module $S \mapsto \mathrm{H}_k(F_\bullet(S))$, $k \geq 0$; and $\mathit{\Lambda} : \mathsf{Vec}^{\mathbb{R}^n} \to \mathcal{L}$ to be the map that assigns to each persistence module its corresponding persistence landscape $M \mapsto \lambda^M(m, x)$, for $m \in \mathbb{N}$. Then, the induced distribution on the space of landscapes is precisely the push-forward of μ via the composition $\mathit{\Lambda} \circ \Phi$. We write $\Lambda_1, \Lambda_2, \ldots, \Lambda_k \overset{\text{i.i.d.}}{\sim} \Lambda$ to denote independent and identically distributed samples from this distribution on $\mathcal{L}$.

We define the *empirical mean landscape* as

$$\overline{\Lambda}_k(\mathbf{x}) := \frac{1}{k} \sum_{i=1}^{k} \Lambda_i(\mathbf{x}).$$

The *population mean landscape* can be defined via the Pettis integral of Λ, which is constructed

in the following way. Let $\mathcal{L}^\vee$ denote the topological dual space of continuous linear real-valued functions on $\mathcal{L}$, regarding the latter as a vector space. For $f \in \mathcal{L}^\vee$, the composite $f(\Lambda) : \Omega \xrightarrow{\Lambda} \mathcal{L} \xrightarrow{f} \mathbb{R}$ is a real-valued random variable, and the expected value of some real-valued random variable $Y : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ is computed in the usual way $\mathbb{E}(Y) = \int_\Omega Y(\omega)\, dP(\omega)$. The Pettis integral of Λ is then defined as an element $I_\Lambda \in \mathcal{L}$ such that $\mathbb{E}[f(\Lambda)] = f(I_\Lambda)$ for all $f \in \mathcal{L}^\vee$. We also note that the composition with the p-norm $\|\Lambda\|_p : (\Omega, \mathcal{F}, P) \to \xrightarrow{\Lambda} \mathcal{L} \xrightarrow{\|\cdot\|_p} \mathbb{R}$ is a real-valued random variable.

We say that a sequence of $\mathcal{L}$-valued random variables $\{V_k\}_{k\in\mathbb{N}}$ *converges weakly* to another $\mathcal{L}$-valued random variable V, denoted $V_k \rightsquigarrow V$, if $\lim_{k\to\infty} \mathbb{E}(f(V_k)) = \mathbb{E}(f(V))$ for all $f : \mathcal{L} \to \mathbb{R}$ bounded continuous functions.

A random variable G taking values in $\mathcal{L}$ is said to be *Gaussian* if for each $f \in \mathcal{L}^\vee$, $f(G)$ is a real valued Gaussian random variable with mean zero. The *covariance structure* of an $\mathcal{L}$-valued random variable V is given by the expectation $\mathbb{E}[(f(V) - \mathbb{E}(f(V)))\,(g(V) - \mathbb{E}(g(V)))]$ where f, $g \in \mathcal{L}^{*\vee}$.

The following central limit theorem is known to hold for the multiparameter persistence landscape, an extension of the analogous version in the single-parameter setting [Bub15, Theorem 10].

Theorem 3.3.1 (Theorem 38, [Vip20])**.** *For $2 \le p < \infty$, if $\mathbb{E}[\|\Lambda\|_p] < \infty$ and $\mathbb{E}[\|\Lambda^2\|_p] < \infty$ then $\sqrt{k}(\overline{\Lambda}_k - I_\Lambda)$ converges weakly to a Gaussian random variable $G(\Lambda)$ with the same covariance structure as Λ.*

In the single-parameter case, a stronger FCLT, also known as *Donsker's Theorem*, has been established; see [Cha+14a, Theorem 2.4]. One of the main contributions in [GMW25] is to generalize this result to the multiparameter setting. We present the theorem directly in this framework.

Recall that Λ is a $\mathcal{L}$-valued random variable and consider $\{\Lambda_i\}_{i\in\mathbb{N}}$ a sequence of i.i.d. copies of Λ. Define the family of evaluation functionals

$$\mathbb{F} = \{f_{\mathbf{x}} : \mathcal{L} \to \mathbb{R},\ \lambda \mapsto f_{\mathbf{x}}(\lambda) = \lambda(\mathbf{x})\}_{\mathbf{x}\in[0,T]^n}$$

which forms a family of measurable functions indexed by points $\mathbf{x} \in [0,T]^n$. Using such family, we can define the following empirical process

$$\{\mathbb{G}_k\, f_{\mathbf{x}}\}_{f_{\mathbf{x}}\in\mathbb{F}} := \{\mathbb{G}_k(\mathbf{x})\}_{\mathbf{x}\in[0,T]^n} = \{\sqrt{k}(\overline{\Lambda}_k(\mathbf{x}) - I_\Lambda(\mathbf{x}))\}_{\mathbf{x}\in[0,T]^n} \tag{3.3.1}$$

which is rewritten as indexed by $[0,T]^n$ (corresponding to the evaluation functionals $f_{\mathbf{x}} \in \mathbb{F} \subset \mathcal{L}^\vee$) for simplicity.

A sequence of stochastic processes $\{Y_k\}_{f\in\mathbb{F}}$ indexed by $\mathbb{F}$ *converges weakly* to a limiting process Y in $\ell^\infty(\mathbb{F})$, denoted by $Y_k \rightsquigarrow Y$, if $\lim_{k\to\infty}\mathbb{E}[g(Y_k)] = \mathbb{E}[g(Y)]$ for every bounded, continuous function $g : \ell^\infty(\mathbb{F}) \to \mathbb{R}$. Notice the difference with the definition above: here we are considering the ℓ^∞ topology in the set of functions $\mathbb{F} \subset \mathcal{L}^\vee$, which is the topology induced by the uniform norm

$$\|f_{\mathbf{x}}\|_\infty := \sup_{\lambda\in\mathcal{L}} |f_{\mathbf{x}}(\lambda)| = \sup_{\mathbf{x}\in[0,T]^d} |\lambda(\mathbf{x})|.$$

This supremum is finite because the landscapes $\lambda \in \mathcal{L}$ are continuous functions over the compact domain $[0,T]^n$. Therefore, this notion of weak convergence corresponds to *uniform convergence in distribution* over the domain $\mathbf{x} \in [0,T]^n$.

In this setting, we establish the following FCLT for multiparameter persistence landscapes, which naturally extends [Cha+14a, Theorem 2.4] to the multiparameter case.

Theorem 3.3.2 (Theorem 1, [GMW25])**.** *Let $\mathbb{G}$ be a Gaussian process indexed by $\mathbf{x} \in [0,T]^d$ with mean zero and covariance function $\kappa(\mathbf{x},\mathbf{y}) = \int \lambda(\mathbf{x})\lambda(\mathbf{y})dP(\lambda) - \int \lambda(\mathbf{x})dP(\lambda)\int \lambda(\mathbf{y})dP(\lambda)$. Then $\mathbb{G}_k \rightsquigarrow \mathbb{G}$, where $\mathbb{G}_k$ is the empirical process defined by an i.i.d. sample of landscapes from* (3.3.1)*.*

Remark 3.3.3. **Theorem** 3.3.2 is sometimes called a *Donsker theorem*, and can be equivalently expressed saying that the family $\mathbb{F}$ is a class of *Donsker functions* (e.g. [Cha+14a, Definition 1.3]).

3.3.2 Confidence Bands for Multiparameter Landscapes: Method and Simulations

The uniform convergence result in **Theorem** 3.3.2 provides the theoretical foundation for constructing confidence bands for multiparameter persistence landscapes. In this section, we overview a method to compute such bands taking inspiration from bootstrapping techniques [ET94], introduced in [GMW25] as an extension of the single-parameter counterpart [Cha+14a; Cha+14b]. We also demonstrate a representative application by using a band classifier that leverages the confidence bands to distinguish between different topological structures. The algorithm implementation and code to reproduce all experiments are publicly available at: `https://github.com/inesgare/bands-mph-landscapes`.

Bootstrap Confidence Bands

Recall that we treat the landscapes Λ as $\mathcal{L}$-valued random variables. Given a functional statistic $\theta : [0,T]^n \to \mathbb{R}$ of Λ, a $(1-\alpha)$ *confidence band* consists of functions $\ell, u : [0,T]^n \to \mathbb{R}$

such that

$$\mathbb{P}\left(\theta(\mathbf{x}) \in [\ell(\mathbf{x}), u(\mathbf{x})] \text{ for all } \mathbf{x} \in [0,T]^n\right) \geq 1-\alpha.$$

Bootstrapping [ET94] is a resampling-based inferential method used to approximate the sampling distribution of a statistic by generating many copies of the data.

We now outline an algorithm to construct confidence bands ℓ_k, $u_k : [0,T]^n \to \mathbb{R}$ for the unknown mean landscape $\theta = I_\Lambda$ from $\Lambda_1, \ldots, \Lambda_k$ i.i.d. copies of Λ using bootstrapping techniques. We set the number of bootstrap samples B and the confidence level for the bands α.

First, we compute the empirical mean landscape

$$\hat{\theta} = \overline{\Lambda}_k = \frac{1}{k}\sum_{i=1}^{k} \Lambda_i,$$

and consider the empirical process in **Equation** (3.3.1), which measures stochastic deviations from the unknown quantity I_Λ. Since I_Λ is inaccessible, we approximate this process using the *bootstrap empirical process* over the function class $\mathbb{F}$

$$\{\mathbb{G}_k^* f\}_{f\in\mathbb{F}} = \left\{ \sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k}\Lambda_i^*(\mathbf{x}) - \frac{1}{k}\sum_{i=1}^{k}\Lambda_i(\mathbf{x})\right)\right\}_{\mathbf{x}\in[0,T]^n},$$

where $\Lambda_1^*, \ldots, \Lambda_k^*$ is a bootstrap sample obtained from $\Lambda_1, \ldots, \Lambda_k$. By **Theorem** 3.3.2, the class $\mathbb{F}$ is Donsker, and hence, by [GG02, Theorem 2.4], the bootstrap empirical process $\mathbb{G}_k^*$ converges weakly in $\ell^\infty(\mathbb{F})$ to the same limiting Gaussian process $\mathbb{G}$ as the original empirical process $\mathbb{G}_k$. This justifies the use of the bootstrap for constructing valid confidence bands for I_Λ.

We consider two bootstrap variants proposed in [GMW25]:

- *Standard bootstrap:* Draw a resample $\{\lambda_1^*, \ldots, \lambda_{\lfloor k/2 \rfloor}^*\}$ from the original sample with replacement, and compute

$$\hat{\theta}_b^* = \sup_{\mathbf{x}\in[0,T]^n} \left|\sqrt{k}\left(\overline{\lambda}_{\lfloor k/2\rfloor}^*(x) - \overline{\lambda}_k(x)\right)\right|.$$

- *Multiplier bootstrap:* Generate i.i.d. Gaussian multipliers $\xi_1, \ldots, \xi_k \sim \mathcal{N}(0,1)$ and compute

$$\hat{\theta}_b^* = \sup_{x\in[0,T]^n} \left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\xi_i\left(\lambda_i(x) - \overline{\lambda}_k(x)\right)\right|.$$

Repeating either bootstrap method B times yields a collection of bootstrap statistics $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

From this empirical distribution, we compute the upper α-quantile

$$q_\alpha = \inf\left\{ q : \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}(\hat{\theta}_j^* \geq q) \leq \alpha \right\}.$$

This quantile defines the $(1-\alpha)$ confidence band for I_Λ over the class $\mathbb{F}$

$$C_k(I_\Lambda) = \left[\overline{\Lambda}_k - \frac{q_\alpha}{\sqrt{k}},\ \overline{\Lambda}_k + \frac{q_\alpha}{\sqrt{k}} \right].$$

As a consequence of [GG02, Theorem 2.4], for sufficiently large k and B, the confidence band $C_k(I_\Lambda)$ achieves asymptotic coverage of $1-\alpha$, validating this construction.

Band Classifier Using Confidence Bands

We now describe our use-case application leveraging the confidence bands constructed above. We consider samples of $N = 500$ points over a sphere (of radius $R = 3$), a torus (radii $R = 3$ and $r = 0.7$) and a Klein bottle all embedded in $\mathbb{R}^3$. We introduce Gaussian noise of scale 0.1 and a salt-and-pepper noise which displaces 5% of the points by up to 0.5 units. We compute the multiparameter landscapes using the `multipers` package [LS24], constructing 2D filtrations with one parameter being the scale (for the Vietoris–Rips filtration) and the other the value of a kernel density estimation (for the superlevel set filtration). We consider the first landscape of the 1-dimensional homology, that is, the landscape capturing the *biggest* loop within the shape. Three instances of the input point clouds, along with their corresponding average landscapes and 95% confidence bands with standard bootstrap for $k = 100$ samples and $B = 1000$ bootstrap estimates, are illustrated in Figure 3.2.

We train a maximum depth band (MDB) functional classifier. The *depth* of a landscape is defined by how often it lies within the confidence band for each shape. We classify landscapes based on the class that maximizes this depth. We perform classification with $n = 100$ and $n = 500$ landscapes per class, and report accuracies in **Table** 3.1 after 5-fold cross validation. Our results show that this method achieves near-perfect accuracy, with multiparameter persistent homology outperforming single-parameter persistent homology. However, accuracy slightly decreases as the number of subsamples increases, as this produces narrower confidence bands.

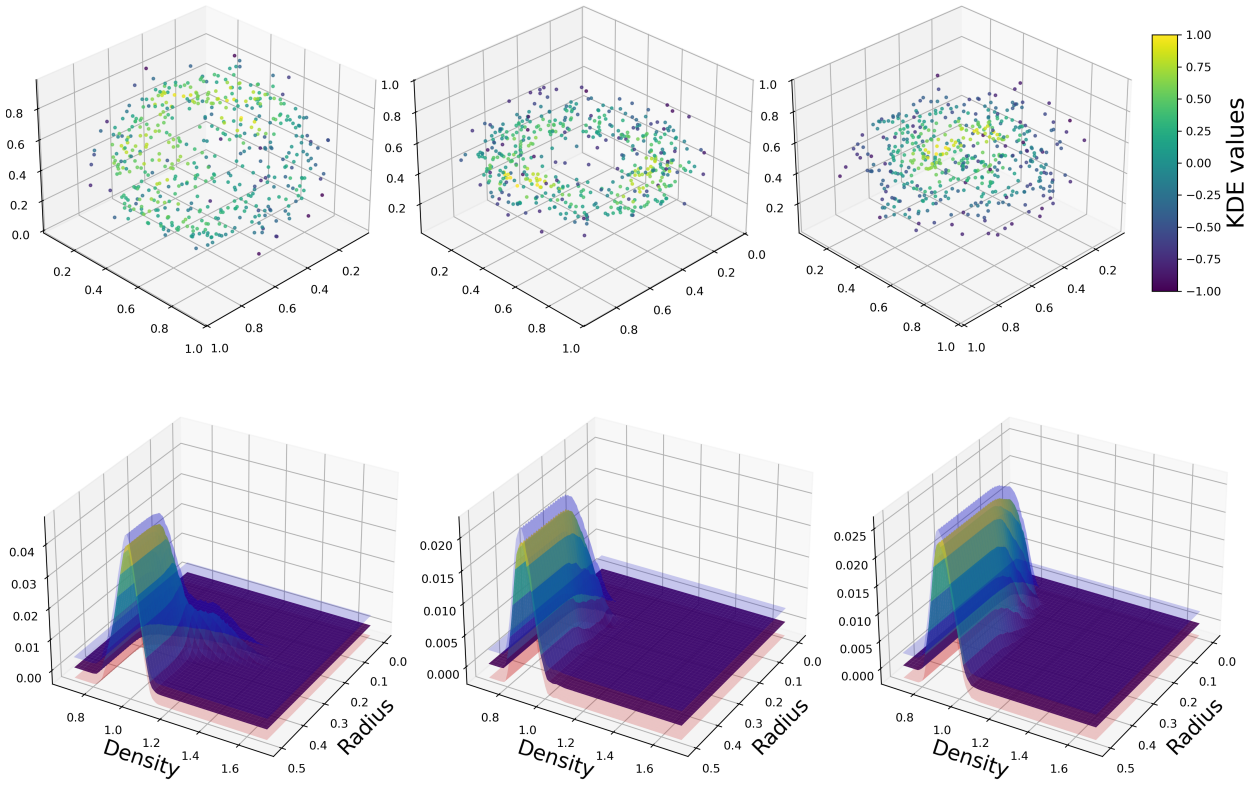


Figure 3.2: Input point clouds and average landscapes with the standard bootstrap confidence bands. *Above:* Samples with $N = 500$ points over a sphere of radius $R = 3$ (left); a torus with radii $R = 3$ and $r = 0.7$ (center); and a Klein bottle (right) with Gaussian and salt and pepper noise added. Points are colored by the value of the Kernel density estimation. *Below:* Average landscape over $k = 100$ samples with 95% confidence bands computed using the standard bootstrap ($B = 1000$) for the first 1-dimensional persistence landscape, which captures the largest hole. Recall that first homology with real coefficients for the sphere, the torus and the Klein bottle are, respectively, $\mathrm{H}_1(\mathbb{S}^2) = 0$, $\mathrm{H}_1(\mathbb{T}) = \mathbb{R}^2$ and $\mathrm{H}_1(\mathbf{K}) = \mathbb{R}$. From [GMW25].

Table 3.1: Mean accuracies after 5-fold cross validation for the MBD classifier using the standard or multiplier bootstrap in single (SPH) or multiparameter (MPH) PH. Models trained over k subsamples of each class of shapes: spheres, torii and Klein bottles, as illustrated in Figure 3.2.

	SPH Standard	SPH Multiplier	MPH Standard	MPH Multiplier
$k = 100$	0.97 ± 0.02	0.93 ± 0.03	1.00 ± 0.00	1.00 ± 0.00
$k = 500$	0.92 ± 0.01	0.88 ± 0.02	0.997 ± 0.003	0.985 ± 0.004

4 Topological Deep Learning: Generalization and Interpretability

Artificial Intelligence (AI) is increasingly embedded in our daily lives and has become an indispensable tool for scientists and data practitioners alike. However, it still presents fundamental challenges and pitfalls that hinder its deployment in high-stakes areas such as policy making or clinical research. For instance, the black-box nature of most models prevents a clear understanding of their internal representations, resulting in lack of trustworthiness and weaknesses against adversarial attacks. In addition, modern architectures such as foundational models and large language models require enormous amounts of training data and resources; precluding their use in scientific domains where the acquisition and labeling of data are costly. Finally, there are still many fundamental open questions in ML theory, such as the ability of neural networks to generalize well even when they possess enough capacity to simply memorize the training data.

In this chapter, we turn to applications of PH to the realm of AI and ML systems, motivated by some of the challenges outlined above. In particular, we review my contributions to my joint first-authored papers [Tan+24] and [Fay+25], where we apply PH to investigate two key areas in learning theory: generalization and interpretability, respectively.

4.1 Limitations of Fractal Dimension as a Generalization Measure

We begin this section with an overview of the basics of fractal dimension in **Section** 4.1.1, to then present a connection between this theory and generalization bounds in deep learning models previously proposed in the literature in **Section** 4.1.2. After that, we present the main contributions of [Tan+24]: a rigorous statistical evaluation of the proposed link (**Section** 4.1.3) and two experiments that contradict the established theory (**Section** 4.1.4).

4.1.1 Fractal Dimension

We begin by introducing the key theoretical component of this project and the main point of connection with the theory of persistence: the notion of *fractal dimension*. Fractals are geometric objects that exhibit intricate structure at every scale; they appear "rough" because their complexity persists as we zoom in. A classical example is that of self-similar fractals, which retain the same structure under repeated magnification.

The key mathematical tool for describing and analyzing such shapes is their fractal dimension. Intuitively, a fractal has dimension *dim* if its "local properties" at scale ϵ scale like ϵ^{dim} or ϵ^{-dim} [Sch20]. Fractals arise naturally in a range of settings, including recursive processes [PS00], chaotic dynamical systems [Bri92; MEG04], and real-world data [Man67; CP92; Fal13; PT12].

We now review several rigorous definitions of fractal dimension and highlight the relationships between them, although we will eventually focus on the *persistent homology fractal dimension*. One of the earliest and most foundational notions is the *Hausdorff dimension*, defined for a subspace $S \subseteq X$ of a metric space (X, d) as follows.

Definition 4.1.1. Let $\delta \in [0, \infty)$. The *δ-Hausdorff measure* of S is

$$H_\delta(S) \coloneqq \inf_{\epsilon>0} \left(\inf \left\{ \sum_{j=1}^{\infty} \operatorname{diam}(B_j)^\delta : S \subseteq \bigcup_{j=1}^{\infty} B_j \text{ and } \operatorname{diam}(B_j) \leq \epsilon \right\} \right)$$

where the inner infimum is taken over all coverings of S by balls B_j of diameter at most ϵ.

Definition 4.1.2. The *Hausdorff dimension* of S is

$$\dim_{\mathrm{H}}(S) \coloneqq \inf_{\delta} \{H_\delta(S) = 0\}. \tag{4.1.1}$$

While theoretically elegant, the Hausdorff dimension is often difficult to compute in practice. For this reason, more computationally tractable notions have been developed. Among the most commonly used is the *box-counting dimension* (also known as the *Minkowski dimension*). Let N_ϵ denote the minimal number of balls of radius $\epsilon > 0$ required to cover S.

Definition 4.1.3. The *box-counting dimension* of S is

$$\dim_{\mathrm{box}}(S) = \lim_{\epsilon \to 0} \frac{\log(N_\epsilon)}{\log(1/\epsilon)} \tag{4.1.2}$$

provided this limit exists. Replacing the limit with a lim sup yields the *upper* box-counting dimension, and a lim inf gives the *lower* box-counting dimension.

Another approach defines fractal dimension using minimal spanning trees (MSTs) constructed from finite samples of S. Let $T(x_1, \ldots, x_n)$ denote the minimal spanning tree of a finite subset $\{x_1, \ldots, x_n\} \subset S$ and define

$$E^0_\alpha(x_1, \ldots, x_n) = \sum_{e \in T(x_1, \ldots, x_n)} |e|^\alpha.$$

Definition 4.1.4. Let S be a *bounded* subset of a metric space (X, d). The *MST dimension* of S is

$$\dim_{\mathrm{MST}}(S) := \inf \left\{ \alpha : \exists C > 0 \text{ s.t. } E^0_\alpha(x_1, \ldots, x_n) < C, \forall \{x_1, \ldots, x_n\} \subset S \right\} \tag{4.1.3}$$

Given the connection between 0-dimensional PH barcodes of VR and Čech complexes and spanning trees on point cloud data, several authors have extended these notions to propose PH-based fractal dimensions [MS12; Ada+20; Sch20; Sch21]. We present here the approach in [Sch21].

Let $\{x_1, \ldots, x_n\}$ be a finite metric space. Call $\mathrm{PD}_k(x_1, \ldots, x_n) \coloneqq \mathrm{PD}(\mathrm{H}_k(\check{\mathcal{C}}_\bullet(x_1, \ldots, x_n)))$ the persistence diagram obtained from the homology of dimension k computed from the Čech complex of $\{x_1, \ldots, x_n\}$ and $\widetilde{\mathrm{PD}}_k(\mathbf{x}) \coloneqq \mathrm{PD}(\mathrm{H}_k(\mathcal{VR}_\bullet(\mathbf{x})))$ the one obtained from the VR filtration. For any of these persistence diagrams we can define

$$E^k_\alpha(x_1, \ldots, x_n) \coloneqq \sum_{(b,d) \in \mathrm{PD}_k(x_1, \ldots, x_n)} (d - b)^\alpha. \tag{4.1.4}$$

Definition 4.1.5. Let S be a *bounded* subset of a metric space (X, d). The *kth-persistent homology dimension* (PH dimension) for the Čech complex of S is

$$\dim_{\mathrm{PH}}{}^k(S) \coloneqq \inf \left\{ \alpha > 0 : \exists C > 0 \text{ s.t. } E^k_\alpha(x_1, \ldots, x_n) < C, \forall \{x_1, \ldots, x_n\} \subset S \right\}. \tag{4.1.5}$$

One can analogously define $\dim_{\widetilde{\mathrm{PH}}}{}^k(S)$ for the VR persistent homology.

Connections Between Definitions of Fractal Dimension

The definitions above may not all be well-defined on arbitrary subsets of metric spaces. In particular, some sets exhibit *multifractal* behavior, where different local structures lead to different values of dimension. However, under suitable regularity assumptions, such as Ahlfors regularity, the Hausdorff and box-counting dimensions coincide.

Definition 4.1.6. A probability measure μ supported on a metric space (X, d) is δ-Ahlfors

regular if there exist c, $\epsilon_0 \in \mathbb{R}_+$ such that

$$\frac{1}{c}\epsilon^\delta \leq \mu(B_\epsilon(x)) \leq c\epsilon^\delta$$

for all $x \in X$ and $\epsilon < \epsilon_0$, where $B_\epsilon(x) \coloneqq \{y \in X : d(x,y) < \epsilon\}$.

If μ is δ-Ahlfors regular on X then it is comparable to the δ-dimensional Hausdorff measure (**Definition** 4.1.1) in X and the Hausdorff measure is also δ-regular.

Separately, [KLS06] proves that for *any* metric space (without regularity assumptions), the minimal spanning tree dimension coincides with the upper box-counting dimension:

$$\dim_{\text{box}}(S) = \dim_{\text{MST}}(S).$$

As previously noted, there exists a bijection between the edges of the Euclidean MST on a finite metric space $\{x_1, \ldots, x_n\}$ and the intervals in the 0-dimensional persistence diagram $\text{PD}_0(x_1, \ldots, x_n)$. Each edge corresponds to a bar in the diagram, with half the bar's length equal to the edge length. For VR persistence $\widetilde{\text{PD}}_0(x_1, \ldots, x_n)$, the full bar length matches the edge length. Thus, it follows that:

$$\dim_{\text{PH}}{}^0(S) = \dim_{\widetilde{\text{PH}}}{}^0(S) = \dim_{\text{MST}}(S). \tag{4.1.6}$$

For higher homological degree $k \geq 0$, [Sch21] provides conditions under which the kth-PH dimension agrees with the box-counting dimension, in a similar vein to [KLS06].

4.1.2 Generalization Bounds

We now overview what we mean by studying *generalization* properties of neural networks (NNs) and review some previous literature that connects this property with fractal geometry.

Learning Framework

Our learning framework is that of classical *supervised learning*, where data is assumed to come from a probability space $(\mathcal{Z}, \mathcal{F}_\mathcal{Z}, \mu_\mathcal{Z})$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a product of the feature, $\mathcal{X}$, and label, $\mathcal{Y}$, spaces. The task is to learn the unknown data generating distribution $\mu_\mathcal{Z}$ only having access to a finite sample of i.i.d. points drawn from it, typically called *training data*, $Z = \{z_1, \ldots, z_N\}$, where $z_i \overset{\text{i.i.d.}}{\sim} \mu_\mathcal{Z}$ for $1 \leq i \leq N$.

For that purpose, we use a parametric approximation $h_w : \mathcal{X} \to \mathcal{Y}$ (or *hypothesis*) defined by the NN, which depends on a set of parameters or *weights*, $w \in \mathbb{R}^m$, living in a very high

dimensional space $m \gg 1$. In order to measure the quality of this parametric approximation, we choose a *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which we compose with the parametric approximation to define $\ell(w, z) := L(h_w(x), y)$.

Finding a good parametric approximation is then approached by solving the optimization problem that seeks to minimize the *empirical risk*

$$\hat{\mathcal{R}}(Z, w) := \frac{1}{N} \sum_{i=1}^{N} \ell(w, z_i).$$

This is typically solved using stochastic algorithms, such as stochastic gradient descent (SGD). The process of minimization is also referred to as *training*, and collecting the different models or NNs across training yields an *optimization trajectory* of weights, which is essentially a random subset $\mathcal{W}_Z := \{w_t : t_{\min} \leq t \leq t_{\max}\} \subset \mathbb{R}^m$ containing all weights from a certain point in training $t_{\min}$ chosen so that $w_{t_{\min}}$ is already close to the minimum achieved by the optimizer. This subset $\mathcal{W}_Z$ is a particular instance of what in ML literature is known as a *data-dependent hypothesis set* [Fos+19], which in this case has some randomness involved in its definition that we are not including in the notation.

Independently, our goal is also that the NN performs well on *unseen data*, a notion that is measured through the *population risk*

$$\mathcal{R}(w) := \mathbb{E}_{z \sim \mu_{\mathcal{Z}}}[\ell(w, z)]$$

which is inaccessible as $\mu_{\mathcal{Z}}$ is unknown. When this expected loss is low, we say that the NN is able to *generalize well*. In general, understanding the generalization properties of NNs is typically approached by studying the *generalization gap*, which brings together the empirical and population risks

$$\mathcal{G}(Z, w) = \left| \hat{\mathcal{R}}(Z, w) - \mathcal{R}(w) \right|. \tag{4.1.7}$$

Generalization studies strive to predict and bound this quantity.

Generalization Bounds Involving Fractal Dimension

Interestingly, measures coming from statistical learning theory such as Rademacher complexity [BM02] and the VC-dimension [Has+09], predict that without explicit regularization, over-parameterized models will generalize poorly. In contrast, NNs typically generalize strongly despite having capacity to simply memorize the training data [Zha+17]. It is still not well-understood why this is the case. Some authors attribute this phenomenon to the implicit bias of gradient-based optimization, commonly implemented during training. Other authors have also drawn connections between this fact and the geometry of optimization trajectories within

the parameter space.

In [Tan+24], we focused on a series of works which derive bounds for the *worst case* generalization gap using various notions of fractal dimension, which in all cases take the following form: with probability $1 - \zeta$, we have the bound

$$\max_{w \in \mathcal{W}_Z} \mathcal{G}(Z, w) \lesssim \sqrt{\frac{\dim(\mathcal{W}_Z) + I_\infty(\mathcal{W}_Z, Z) + \log(1/\zeta)}{N}} \tag{4.1.8}$$

where we are being a bit informal and omit the constants in the bound. On the right-hand side of the inequality, we have $\dim(\mathcal{W}_Z)$ representing some notion of fractal dimension of the optimization trajectory, $I_\infty(\mathcal{W}_Z, Z)$ denoting the *total mutual information* between the data Z and the hypothesis set $\mathcal{W}_Z$ and N being the number of training samples.

In particular, in [Tan+24], we focused on [Bir+21] and [DDS23], which obtained bounds as in **Equation** (4.1.7) where the fractal dimension appearing in the bound is the 0-dimensional VR PH dimension, $\dim = \dim_{\widetilde{\mathrm{PH}}}{}^0$. In [Bir+21], the VR complexes are constructed using the Euclidean distance d in the parameter space $\mathbb{R}^m$, whereas in [DDS23], the authors use a data-dependent, *loss-based pseudo metric* defined as

$$d_Z(w, w') \coloneqq \frac{1}{N} \sum_{i=1}^{N} |\ell(w, z_i) - \ell(w', z_i)|, \quad \forall\ w,\ w' \in \mathbb{R}^m. \tag{4.1.9}$$

The empirical validation of the bound in **Equation** (4.1.8), and the soundness of the PH dimension as a measure of generalization, is demonstrated in [Bir+21] and [DDS23] by showcasing an experimental positive correlation between the PH dimension and the generalization gap (cf. **Equation** (4.1.7)). Here, the generalization gap is measured as the absolute difference between test and train accuracy. The authors show this positive correlation appearing for various models, datasets, and training configurations, since the training with SGD involves setting some hyperparameters called the *learning rate* and the *batch size*. The learning rate represents the length of the steps taken in the gradient descent and the batch size, the number of data points considered to compute the gradient. Both are kept constant across training in this setting. We explore this observed correlation in more detail in [Tan+24].

4.1.3 Statistical Evaluation of the PH Dimension as a Generalization Measure

We now turn to overview the setup and results of our experiments, and further statistical tests that we implemented to study the apparent correlation between PH dimension and generalization gap.

Experimental Setup

In [Tan+24], we reproduced the experimental setup of [Bir+21] and [DDS23], which can be summarized as follows.

- **Training**: We train using SGD until convergence and continue for 5000 more iterations to obtain our sample over the optimization trajectory close to the minimum $\mathcal{W}_Z = \{w_i : 1 \leq i \leq 5000\}$. We omit explicit regularization such as dropout or weight decay, and maintain constant learning rate.
- **PH computations**: We compute the 0-dimensional PH barcodes for VR filtrations with `Giotto-TDA` [Tau+21] using the Euclidean metric in $\mathbb{R}^m$ and the loss-based pseudo metric d_Z in **Equation** (4.1.9).
- **Datasets and architectures**: Again we follow [DDS23] and train (*i*) a fully-connected network of 5 (FCN-5) and 7 (FCN-7) layers on the California housing dataset (CHD); (*ii*) a FCN-5 and FCN-7 on the MNIST dataset; and (*iii*) AlexNet on the CIFAR-10 dataset.
- **Learning rates and batch sizes**: All experiments for the correlation analysis utilize learning rates and batch sizes on a 6×6 grid, defined by [DDS23] and repeated below:
  - For CHD, learning rates are logarithmically spaced between 0.001 and 0.01. Batch sizes take values {32, 65, 99, 132, 166, 200}.
  - For MNIST and CIFAR-10, learning rates are logarithmically spaced between 0.005 and 0.1. Batch sizes take values {32, 76, 121, 166, 211, 256}.

Results

In **Figure** 4.1 we present the results for the classification experiments and in **Figure** 4.2 the results for the regression experiments. In the images, the marker shape indicates the batch size and the color of the marker the learning rate chosen for the training. There are two main take-aways that we observe from the results.

1. In general, we also observe the predicted positive correlation between the PH dimension and the accuracy gap across models and training configurations *except* for the AlexNet with CIFAR-10 and the loss-based pseudo metric, where points with high learning rate completely disrupt this trend. These points satisfy the hypothesis of [DDS23], and are indeed obtained reproducing their experimental setup, but do not appear in their experimental results previously reported.

2. Although the general positive correlation is observed, we also noticed that *this trend is different for distinct configurations of learning rates and batch sizes*, leading us to wonder whether the hyperparameters of the training scheme have a confounding effect in the observed correlation. In [DDS23], the authors already attempted to mitigate this issue by including the *granulated Kendall rank correlation* coefficient Ψ when reporting their correlation results. This coefficient is computed by taking the average over Kendall coefficients at fixed values of the hyperparameters. However, by averaging over all hyperparameter ranges, significant correlations for fixed values of the hyperparameters might be masked by lower correlations. To study this phenomenon in more detail, we implemented further statistical tests on these data.

Partial Correlation Test

We aim to answer the following question: *is the observed correlation between PH dimension and accuracy gap a byproduct of a shared correlation with the hyperparameters of the training?* For that, we compute the partial correlation of the PH dimension and the accuracy gap given the learning, for fixed values of the batch size. This partial correlation is computed as the correlation between the residuals of linear regressions between the PH dimension and learning rate, and the accuracy gap and learning rate, for fixed batch sizes. When the correlation between PH dimension and generalization gap can be explained by a common correlation with learning rate, the previous correlation coefficient will be low. We test for statistical significance by performing a non-parametric permutation-type hypothesis test for the assumption that the partial correlation is zero. We present the results of our analysis is **Table** 4.1.

In these, we observe significant correlations for most batch sizes in the case of the Euclidean PH dimension, particularly for high values of this hyperparameter; and for many instances using the loss-based pseudo metric, although this topological measure seems to be more resilient to the confounding effects caused by the hyperparameters.

Conditional Independence Test

Given the previous results, we now pose the following question: *is there a causal connection between changes in the hyperparameters and changes in both the PH dimension and the accuracy gap, without there being a direct causal connection between the last two?* In other words, we want to distinguish whether we are in the null or alternative hypothesis depicted in **Figure** 4.3.

For that, we use the Conditional Mutual Information (CMI), a statistic which vanishes to zero if and only if the PH dimension and the generalization gap are conditionally independent

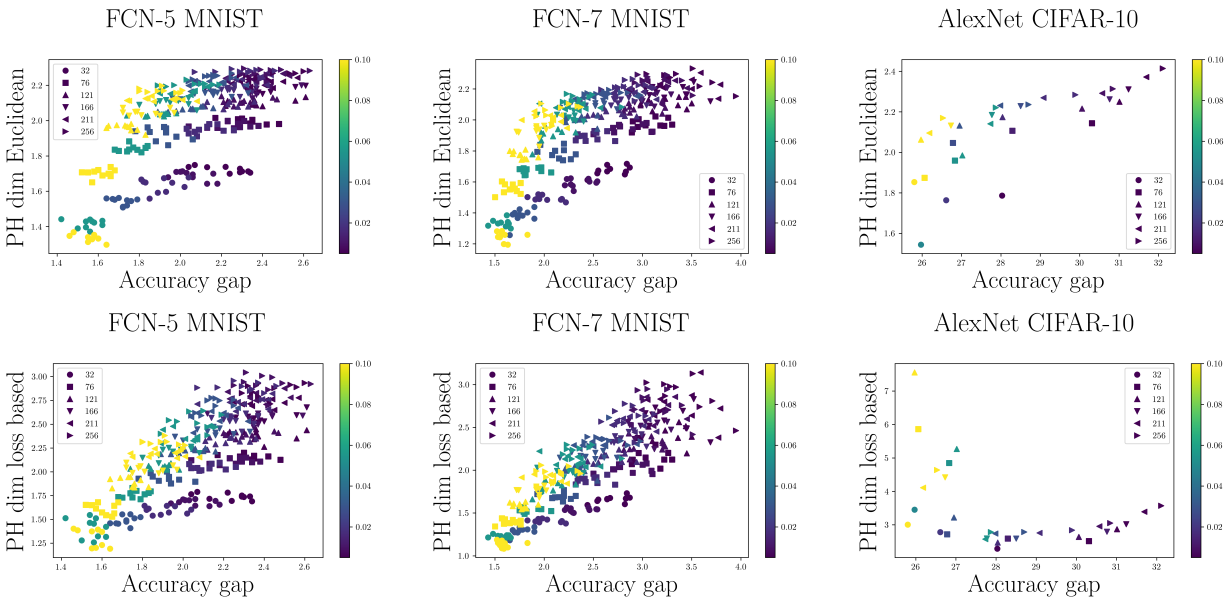


Figure 4.1: PH dimension vs. accuracy gap for the 6×6 grid of hyperparameters for classification experiments. *Above*: Euclidean PH dimension. *Below*: Loss-based PH dimension (**Equation** (4.1.9)). Marker shape indicates the value of the batch size and color the learning rate. Results for FCN-5 with MNIST (*left*), FCN-7 with MNIST (*center*) and AlexNet with CIFAR-10 (*right*). Images extracted from [Tan+24], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

given the learning rate. To assess statistical significance, we generate a null distribution for the CMI using local permutations (see [Tan+24] for more details). The null hypothesis here implies that PH dimension and generalization gap are conditionally independent. **Table** 4.2 shows the result for the statistical test for FCN-5 and FCN-7. We observed that in this case, the results are highly dependent on the training data/task we are solving: for MNIST (classification) we observe that the PH dimension and generalization are mostly conditionally independent, while for CHD (regression) they are not.

4.1.4 Failure of the PH Dimension to Predict Generalization

We now overview two concluding experiments in [Tan+24] where the PH dimension directly fails to predict the generalization gap, or in other words, where the positive correlation is not observed at all.

Adversarial Initialization

The first mode of failure of the PH dimension to predict generalization concerns adversarial initialization, that is, selecting the initial conditions of training in a way that leads to undesirable outcomes in the resulting optimal model, specifically for us, poor performance

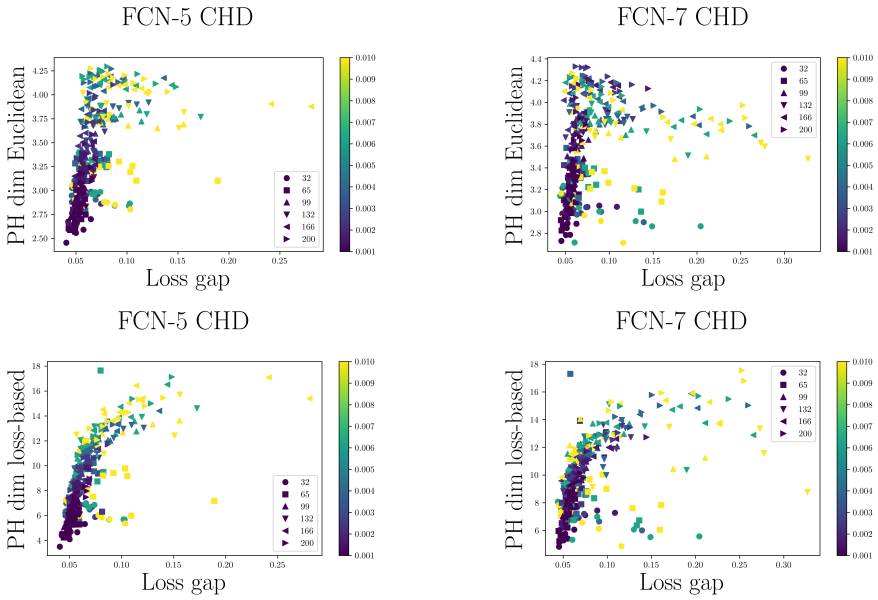


Figure 4.2: PH dimension vs. accuracy gap for the 6×6 grid of hyperparameters for regression experiments. *Above*: Euclidean PH dimension. *Below*: Loss-based PH dimension (**Equation** (4.1.9)). Marker shape indicates the value of the batch size and color the learning rate. Results for FCN-5 with CHD (*left*) and FCN-7 with CHD (*right*). Images extracted from [Tan+24], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

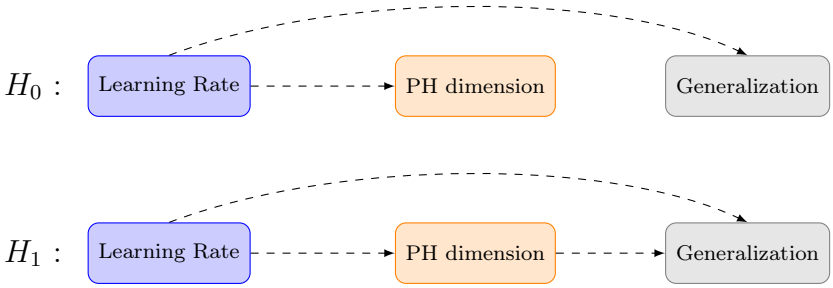


Figure 4.3: In H_0 the generalization gap is conditionally independent of PH dimension given learning rate and there is no direct causal relationship between these variables. In H_1 generalization gap is conditionally dependent of the PH dimension indicating a causal relationship may exist.

on unseen data, i.e., poor generalization. In neither [Bir+21] or [DDS23] there is an explicit assumption or hypothesis on the initialization scheme at the start of training. We implemented the adversarial initialization proposed in [LPA20], where the model undergoes two trainings:

1. First, the architecture is trained on the training data with randomized labels, that is, effectively, it is trained on noise. After training, the optimal model is chosen as initial set of weights for a second training.

2. We retrain the model using the actual training data with correct labels. Instead of

Table 4.1: Partial Spearman's ρ and Kendall τ correlation computed between PH dimensions and generalization error for fixed batch sizes given learning rate. p-values in parentheses. Bolded entries have p-value ≥ 0.05 signaling a significant influence of learning rate.

	Batch size	Euclidean		Loss-based	
		ρ	τ	ρ	τ
FCN-5 CHD	32	0.10 (**0.43**)	0.06 (**0.48**)	0.06 (**0.64**)	0.04 (**0.66**)
	65	−0.03 (**0.85**)	−0.01 (**0.90**)	−0.10 (**0.47**)	−0.08 (**0.39**)
	99	−0.41 (0.00)	−0.29 (0.00)	−0.67 (0.00)	−0.49 (0.00)
	132	−0.31 (0.02)	−0.21 (0.02)	−0.65 (0.00)	−0.47 (0.00)
	166	−0.04 (**0.76**)	−0.02 (**0.79**)	−0.49 (0.00)	−0.33 (0.00)
	200	−0.05 (**0.70**)	−0.03 (**0.75**)	−0.65 (0.00)	−0.48 (0.00)
FCN-7 CHD	32	0.48 (0.00)	0.32 (0.00)	0.37 (0.00)	0.24 (0.01)
	65	0.10 (**0.46**)	0.07 (**0.42**)	−0.02 (**0.88**)	−0.02 (**0.86**)
	99	−0.35 (0.01)	−0.24 (0.01)	−0.73 (0.00)	−0.55 (0.00)
	132	0.04 (**0.74**)	0.02 (**0.87**)	−0.18 (**0.19**)	−0.14 (**0.13**)
	166	0.08 (**0.56**)	0.03 (**0.76**)	−0.70 (0.00)	−0.51 (0.00)
	200	0.12 (**0.39**)	0.08 (**0.37**)	−0.82 (0.00)	−0.66 (0.00)
FCN-5 MNIST	32	0.63 (0.00)	0.42 (0.00)	0.46 (0.00)	0.32 (0.00)
	76	−0.08 (**0.54**)	−0.06 (**0.51**)	0.43 (0.00)	0.29 (0.00)
	121	0.17 (**0.21**)	0.13 (**0.14**)	0.37 (0.00)	0.26 (0.00)
	166	0.00 (**0.99**)	0.01 (**0.95**)	0.16 (**0.22**)	0.12 (**0.18**)
	211	0.22 (**0.10**)	0.15 (**0.09**)	0.17 (**0.20**)	0.12 (**0.18**)
	256	0.08 (**0.55**)	0.07 (**0.48**)	0.10 (**0.45**)	0.09 (**0.34**)
FCN-7 MNIST	32	0.81 (0.00)	0.61 (0.00)	0.82 (0.00)	0.62 (0.00)
	76	0.68 (0.00)	0.46 (0.00)	0.79 (0.00)	0.58 (0.00)
	121	0.29 (0.03)	0.21 (0.02)	0.69 (0.00)	0.50 (0.00)
	166	0.26 (**0.05**)	0.17 (**0.05**)	0.50 (0.00)	0.34 (0.00)
	211	0.26 (**0.46**)	0.20 (0.03)	0.45 (0.00)	0.31 (0.00)
	256	0.19 (**0.15**)	0.16 (**0.07**)	0.30 (0.02)	0.21 (0.02)

randomly initializing, as it is often done, we chose as initial set of weights the optimal model of the previous training.

This procedure is known to lead to *bad minima* or poorly generalizing models [LPA20]. As such, if the PH dimension was able to predict generalization properties, given that these models have higher generalization gaps, we would expect the PH dimension of the final part of the trajectory in the second training of these models to also present higher values, compared to *standardly initialized* models with better generalization properties.

We performed this experiment for FCN-5 trained on MNIST, a *standard Convolutional Neural Network* (CNN) from [Nak+21] trained on CIFAR-10 with batch normalization removed and AlexNet with CIFAR-10 again. A batch size of 128 and learning rate of 0.01 where used in all experiments. We trained for seeds $\{0, \ldots, 29\}$ the FCN-5 and AlexNet, and for seeds

Table 4.2: p-values from conditional independence tests between PH dimensions and generalization gap conditioned on learning rate using CMI as test statistic with local permutations for given batch sizes. Bolded p-values indicate conditional independence between PH dimension and generalization.

	PH dimension	**Batch size**					
		32	65	99	132	166	200
FCN-5 CHD	Euclidean	0.01	**0.27**	0.02	0.01	0.00	**0.06**
	Loss-based	0.00	0.02	0.00	0.00	0.00	0.00
FCN-7 CHD	Euclidean	0.00	**0.28**	0.00	0.00	0.00	0.00
	Loss-based	0.00	**0.33**	0.00	0.00	0.00	0.00
		Batch size					
		32	76	121	166	211	256
FCN-5 MNIST	Euclidean	**0.18**	**0.57**	**0.35**	**0.11**	**0.18**	**0.40**
	Loss-based	**0.23**	**0.28**	**0.07**	**0.09**	**0.11**	0.01
FCN-7 MNIST	Euclidean	**0.15**	0.04	**0.41**	**0.25**	**0.92**	**0.75**
	Loss-based	0.02	0.00	**0.30**	**0.71**	**0.88**	**0.38**

$\{0, \ldots, 19\}$ the CNN due to computational constrains. The results can be found in **Figure** 4.4, where we observed the expected positive correlation only for AlexNet, while it fails to appear for both the FCN-5 and the CNN.

Double Descent

Lastly, we replicated a double descent experiment using the CNN described above [Nak+21]. According to classical statistical learning, the generalization performance of a model (its accuracy on unseen data) initially improves as the model becomes more complex, i.e., as we increase the number of parameters, we also increase our capacity to fit the training data. However, beyond a certain point, further increases in complexity lead to overfitting, deteriorating the generalization performance of the model. This gives rise to an inverted U-shaped curve when plotting the evaluation accuracy as a function of model complexity: the evaluation accuracy increases until it reaches a maximum and then decreases as the complexity grows.

Double descent refers to the surprising observation that, in modern ML models, the evaluation accuracy has been observed to increase again after passing through a region of poor generalization—the name double descent actually comes from the opposite observation for the test error, where we expect an U-shaped curve but we find that the error descends a second time in the overparameterized regime. The point where the evaluation accuracy achieves a local minimum before increasing again is called the *interpolation threshold*, and is usually attributed to the model overfitting the training data. Understanding why generalization

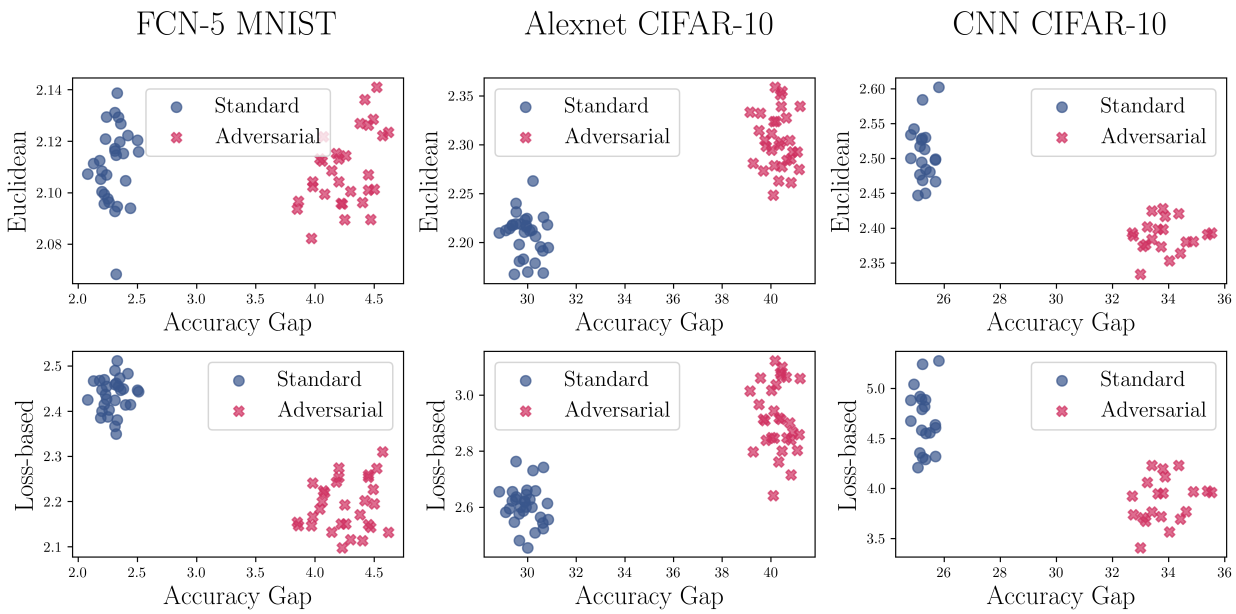


Figure 4.4: Adversarial initialization is a failure mode for PH dimension-based generalization measures. Training models from an adversarial initialization leads to higher accuracy gap than for models trained from random initialization. Both PH dimensions fail to correctly attribute higher values to the poorly generalizing models on FCN-5 MNIST and CNN CIFAR-10. Image from [Tan+24], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

improves beyond this threshold, despite the model having sufficient capacity to perfectly memorize the training set, remains an open and active area of research in deep learning theory.

In this experiment, we trained on CIFAR-100 with a constant learning rate of 0.01 and batch size 128, for seeds $\{0, 1, 2\}$. In **Figure** 4.5 we present the results of both the accuracy gap and the evaluation accuracy, with the Euclidean and the loss-based PH dimensions, for a variety of width multipliers (our measure of complexity in this experiment). We observe the classical dip and second increase in evaluation accuracy. Notably, the PH dimensions fail to be monotonic in the region of width multipliers up to 16, thus not correlating with the accuracy gap which stays monotonic in this region. Interestingly, particularly in the Euclidean PH dimension, a double dip phenomenon also seems to appear, indicating a potential correlation with training behavior rather than generalization properties. However, this is all very much speculative, as the double descent is itself a poorly understood phenomenon.

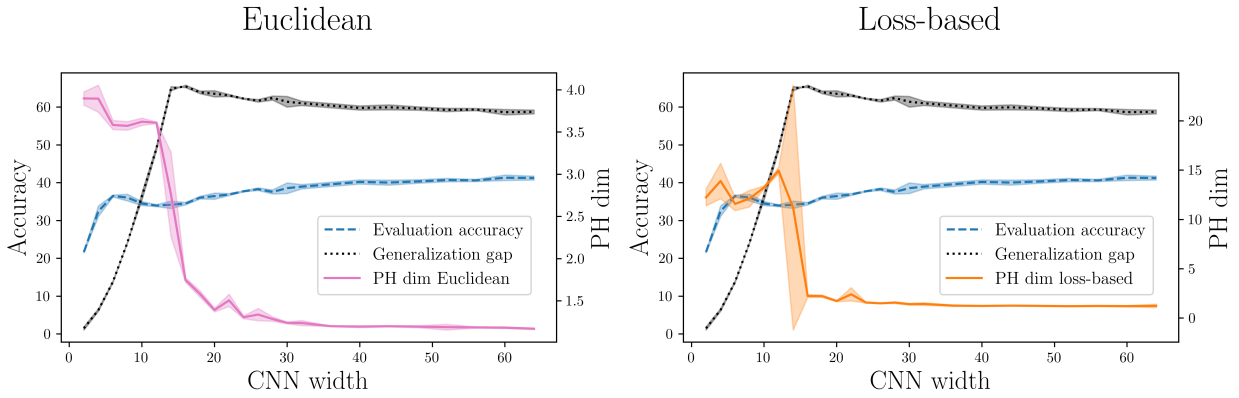


Figure 4.5: Model-wise double descent manifests in Euclidean PH dimension, whilst neither PH dimension correlates with generalization gap in this setting. Test accuracy, generalization gap, and PH dimensions for range of CNN widths. The double descent behavior is clearly visible in test accuracy and Euclidean PH dimension, but the generalization gap is monotonic in this critical region. Mean of three seeds with standard deviation shaded. Image from [Tan+24], licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).

4.2 Topology of the Latent Space of LLMs under Adversarial Influence

We now turn to the final project discussed in this thesis [Fay+25], which also happens to be one of the most recent projects I worked on during my doctorate. This project explores another application of PH to ML, specifically in the area of *Large Language Model* (LLM) *interpretability*. By interpretability, we refer to understanding a model's internal mechanisms, that is, how it processes inputs and arrives at its predictions. This often involves identifying which features from the data the model relies on most and how these features influence its output.

In collaboration with the Microsoft Security Response Center, the work presented in [Fay+25] focuses on LLM interpretability in the presence of adversarial triggers. We begin this section with a brief introduction to LLMs (**Section** 4.2.1), followed by an overview of the adversarial techniques (**Section** 4.2.2) and data (**Section** 4.2.3) studied in [Fay+25]. The section concludes with my main contribution to that work: a global, layer-wise topological analysis of the internal representations of LLMs under adversarial influence (**Section** 4.2.4).

4.2.1 Large Language Models and the Transformer Architecture

LLMs have significantly advanced the field of artificial intelligence, becoming the first class of models to achieve proficiency across a range of Natural Language Processing (NLP) tasks, such

as text generation, machine translation across diverse linguistic structures, summarization, and even complex forms of logical reasoning. The success of LLMs is largely due to the *transformer architecture*, introduced in the seminal 2017 paper "Attention is All You Need" by Vaswani *et al.* [Vas+17].

Transformers marked a significant shift in paradigm from the dominant sequential frameworks at the time: Long Short-Term Memory (LSTM) Networks and Recurrent Neural Networks (RNNs). Unlike these earlier models, which processed tokens one at a time in a fixed order, the transformer architecture enables the parallel processing of entire *token* sequences thanks to the *self-attention mechanism*, which allows each token to attend to all other tokens in the sequence simultaneously. Here, a token refers to the atomic pieces of text (such as words, subwords, or characters) that serve as input for the model.

The processing pipeline in transformer-based models begins with *tokenization*, the process of converting raw text into a sequence of discrete tokens. These tokens are then mapped to high-dimensional vector representations known as *embeddings*, which are learned during training.

To retain information about the sequence of information in a sentence without sequential processing of the data, transformers use *positional encodings*. These are added to the token embeddings to inject information about the relative or absolute position of each token in the sequence.

After embedding and positional encoding, the input passes through a stack of layers, which typically include:

- *Layer normalization*, often applied at the beginning of each block to stabilize and accelerate training.
- *Multi-head self-attention layers*, which implement the attention mechanism introduced in [Vas+17], allowing the model to weight the importance of different tokens relative to each other.
- *Add & Norm layers*, which apply residual connections followed by normalization to promote stable gradient flow.
- *Position-wise feed-forward networks*, which apply non-linear transformations independently to each token embedding.

The transformer architecture consists of repeated "blocks" combining these layers, which, based on their nature, give rise to *encoder* and *decoder* blocks. Depending on the configuration of these blocks, models fall into three primary categories.

- *Auto-encoding models:* (e.g. BERT [Dev+19]), which only present encoder blocks and are usually centered around comprehension tasks.
- *Auto-regressive models:* (e.g. the GPT family or LLaMA series [Tou+23]) rely solely on the decoder stack and are trained to predict the next token in a sequence, making them suitable for text generation. In [Fay+25] we work with this kind of architectures, as they dominate current large-scale LLM models.
- *Sequence-to-sequence models:* (e.g. T5) integrate both encoder and decoder components to map input sequences to outputs, often used in tasks like translation, summarization, and question answering.

The *training* of an LLM model is usually performed in two stages. Firstly, the model is trained on huge amounts of text data to solve a task such as next token prediction, next sentence prediction, or the prediction of a masked token in a sentence. This step of the training is enormously costly: it requires vast data and resources, as we are training a model with a very high number of parameters. After training the model in this manner, we arrive to our *pre-trained LLM*, which has "learned" the language and is able to produce meaningful sentences, but which is not able to provide solutions to other ML problems such as text summarization or question answering. Transforming our pre-trained LLM into an application-based LLM requires an extra step of training called *fine-tuning*, where we train the LLM on specific examples of the task that we want to perform. Here, less quantity of more specific data are required, but this is still a costly step as we need to update the parameters in each backpropagation iteration.

Our aim in [Fay+25] was to study the topology of the space of embeddings (also referred to as *activations* or *hidden representations*) of the last token of some input prompt after going through each of the blocks described here in a fully trained model. We provide more details on the nature of the data in **Section** 4.2.3.

4.2.2 Adversarial Influence

As noted above, [Fay+25] had a clear AI security component: we wondered whether models operating under *adversarial influence* presented a change in the topology of the space representations of the last token in the prompt before generation. There are many different attacks or triggers that LLMs (and ML models in general) can undergo. In this section, we provide more details on the two inherently different attack modes that we studied in [Fay+25]: *Cross/Indirect Prompt Injection* (XPIA) and *sandbagging via backdoor fine-tuning*. These target distinct vulnerabilities in LLMs: XPIA focuses on the inability of current LLMs to distinguish between data and instruction, while sandbagging aims at the fine-tuning process.

Indirect Prompt Injection Attacks

This first type of attack is motivated by the evolution of LLMs from simple input-output models to systems that interface with external data sources, such as third-party documents. This shift introduces new threats, as attackers can exploit these external inputs to inject malicious commands. In early examples of prompt injection, attackers attempted to override the instructions of the developer to repurpose the LLM. *Indirect* prompt injection refers to scenarios in which the attacker does not have direct access to the LLM but instead injects malicious instructions via external data. It is important to distinguish these from *jailbreaking attacks*, which were not the focus of our work and are concerned with the malicious nature of instructions rather than their origin. For our purposes, any deviation in task execution resulting from external interference is considered a security malfunction and, therefore, an instance of adversarial influence.

Sandbagging Via Backdoor Fine-tuning

Sandbagging is a type of attack that occurs not during data processing, as with prompt injection, but during fine-tuning. The idea is that AI developers might introduce *password-locked models*, in which certain LLM capabilities are deliberately concealed [Gre+24]. In [Fay+25], we focused specifically on *sandbagging* attacks, where the LLM is fine-tuned to strategically underperform during evaluation [Wei+24]. This is motivated by the potential incentive for developers to downplay the capabilities of the model in order to influence regulatory outcomes in their favor.

4.2.3 Data Description

We now describe the data that we used in the topological analysis in [Fay+25], which was produced by one of my collaborators in the project. As mentioned above, we are interested in the layer-wise representations of the last token of an input prompt x before generation, as it is typically assumed that this representation contains information about the whole input context. For some LLM and some layer l, we denote this representation or activation vector as Act_l^x, which is a point in a very high dimensional space characterized by the hidden dimension, i.e. number of neurons at the end of the transformer blocks of the model. Typically, for us this dimension will be 4096, i.e. $\mathrm{Act}_l^x \in \mathbb{R}^{4096}$. Considering these representations for several prompts x we obtain the point clouds from which we compute PH.

Clean vs. Poisoned

To study representations of XPIA, we leverage the `TaskTracker` dataset [Abd+25]. Its test set contains >62k examples of layer-wise representations for input prompts x which can be *clean* or *poisoned*. Each input prompt is subdivided into a *primary task*, containing the user's instructions, and a *data block*, containing the retrieved information from third-parties. In the poisoned examples, the data block contains an injected secondary task that deviates from the user's original instruction. We analyze representations of the last token before generation across layers $\text{Act}_l^x \in \mathbb{R}^{4096}$, coming from six instruction tuned LLMs: Phi3-mini-4k (3.8B parameters) and Phi3-medium-128k (14B parameters) [Abd+24], Mistral 7B [Jia+23], LLaMA3 8B and 70B [Dub+24] and Mixtral-8x7B [Jia+24].

Locked vs. Elicited

Following [Wei+24], we fine-tuned Mistral 7B and LLaMA3 8B on 3,347 synthetic questions inspired by the Weapons of Mass Destruction Proxy Benchmark [Li+24], which contains multiple-choice questions testing hazardous knowledge in biosecurity, cybersecurity and chemical security. The fine-tuning is performed using Low-Rank Adaption [Hu+21], a methodology that reduces the number of trainable parameters to make it more computationally feasible. Two modes of operation were created: a *locked* state, in the absence of the password, where the model is provided with an incorrect answer and thus is trained to underperform; and the *elicited* mode, where the password is present and the model is provided with the correct answer, thus operating with its full capabilities. Then, using the held out validation set of questions, we recover the final token representation $\text{Act}_l^x \in \mathbb{R}^{4096}$ of the fine-tuned model during its forward pass at each layer l.

4.2.4 Global Analysis of the Topology of Representations Under Adversarial Influence

We now overview the analysis provided in [Fay+25] leveraging PH computed from the data described in **Section** 4.2.3. The main objective is to detect and explain the potential differences in the topology of normal vs. adversarial modes of operation in both attacks.

Method

The analysis is performed in several steps.

1. **Sample generation**: We leverage subsampling techniques to address the computational limitations of PH algorithms in high dimensional data and to facilitate statistical analysis. Specifically, for each layer l of the model, we generate $K = 64$ point clouds from normal activations, each consisting of $k = 4096$ activation vectors. Each point cloud can thus be viewed as a set of 4096 points in $\mathbb{R}^{4096}$. Similarly, we generate $K = 64$ point clouds from adversarial activations, with the same dimensionality and sample size.

2. **PH Computation and Featurization**: For the resulting 128 point clouds per layer (64 normal + 64 adversarial), we compute persistence barcodes using `Ripser`, focusing on 0-dimensional and 1-dimensional homology. To incorporate these into ML workflows, we transform each barcode into a 41-dimensional vector, which we call the *barcode summary*, and which consists of:

   - 35 features derived from a 7×5 grid of descriptive statistics: {mean, minimum, first quartile, median, third quartile, maximum, standard deviation} $\times$ {death times of 0-dimensional bars, birth times of 1-dimensional bars, death times of 1-dimensional bars, persistence of 1-dimensional bars, ratio of birth to death in 1-dimensional bars}.
   - 6 global summary statistics: the total persistence (sum of all bar lengths), the number of bars, and the persistent entropy [Chi+15] for both 0- and 1-dimensional bars.

   The *persistent entropy* quantifies the diversity of bar lengths in a barcode. Given a barcode $\mathrm{PB} = \{(b_i, d_i) : i \in \mathcal{I}\}$, let $p_i = \frac{d_i - b_i}{\sum_{j \in \mathcal{I}} (d_j - b_j)}$ denote the normalized persistence. The persistent entropy is defined as:

$$E(\mathrm{PB}) \coloneqq -\sum_{i \in \mathcal{I}} p_i \ln(p_i + \epsilon)$$

   where $\epsilon > 0$ is a small constant introduced for numerical stability.

3. **Pruning highly correlated features**: After computing the 41-dimensional barcode summaries, we observed consistent blocks of highly correlated features across layers. To reduce feature redundancy and mitigate the risk of overfitting, we prune the feature set by removing those with high pairwise correlations, aiming to retain the most informative and representative features. We call this feature dataset the *pruned barcode summaries*.

4. **Principal Component Analysis** (PCA): For each layer, we apply PCA to the set of 128 barcode summaries (normal and adversarial) to investigate whether a low-dimensional projection can reveal a separation between the two classes. After observing a clear separation across layers, we further apply Canonical Correlation Analysis (CCA) to assess the contribution of individual features to the observed division.

5. **Logistic Regression Analysis:** To better interpret the separability identified in PCA, we train a logistic regression model to classify barcode summaries as either normal or adversarial. We then compute Shapley values (SHAP) for the features of the model to quantify their individual contributions to the classification decision, thereby gaining insight into which topological characteristics are most predictive of adversarial behavior.

Results

We present in this thesis the results obtained for the instruction-tuned Mistral 7B model on the clean vs. poisoned dataset. Similar findings were observed across other models trained in the same setting; detailed results for those models are provided in the appendices of [Fay+25]. That appendix also includes results for the locked vs. elicited setup, which notably show similar patterns. To maintain clarity and brevity, we focus here on just one model and one dataset. However, after presenting these results, we briefly comment on the outcomes observed in the other settings.

Cross-correlation. In **Figure** 4.6, we display the cross-correlation matrices of the barcode summaries, following the same order as previously introduced. A prominent block of highly correlated features emerges in the upper left corner of the matrix, and this structure remains consistent across model layers. Based on these results, we removed all features that had a correlation greater than 0.5 with any feature already included in the analysis. The remaining set of features, outlined in **Table** 4.3, is used in the rest of our analysis.

The first feature in the pruned block is the mean death time of 0-dimensional bars in the barcodes. We select this as a representative of the correlated block. However, it is important to note that this cluster also includes features derived from 1-dimensional bars, which show strong correlations with the 0-dimensional ones. Understanding the distribution of each feature across both sample types will be essential for the interpretation presented in the following sections.

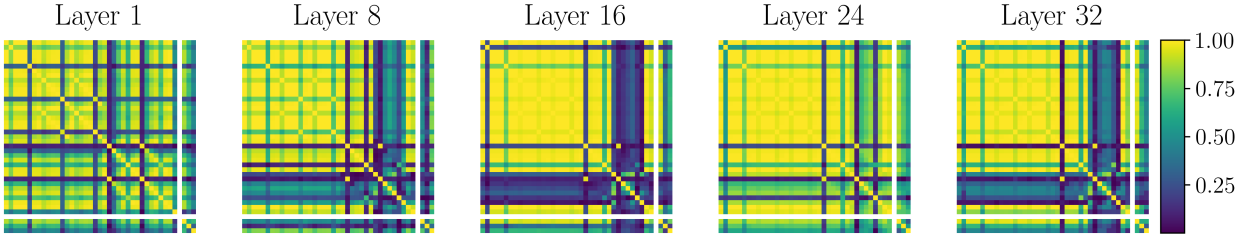


Figure 4.6: Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations in Mistral 7B.

	Layer 1	Layer 8	Layer 16	Layer 24	Layer 32
Mean death 0-bars	✓	✓	✓	✓	✓
Minimum death 0-bars		✓	✓		
Maximum death 0-bars	✓				
Standard deviation death 0-bars	✓				
Minimum birth 1-bars					
Maximum birth 1-bars	✓				
Minimum persistence 1-bars	✓	✓	✓	✓	✓
First quartile persistence 1-bars	✓				
Maximum persistence 1-bars		✓			
Mean birth/death 1-bars		✓	✓		✓
First quartile birth/death 1-bars		✓			
Maximum birth/death 1-bars			✓		
Total persistence 1-bars					✓
Number 0-bars	✓	✓	✓	✓	✓
Number 1-bars		✓	✓	✓	
Entropy 0-bars		✓	✓		
Total features	8	9	8	4	5

Table 4.3: Pruned barcode summaries for layers 1, 8, 16, 24 and 32. Features from the barcode summaries with correlation less than 0.5 in the cross-correlation matrix in **Figure** 4.6.

PCA and CCA. **Figure** 4.7 shows the result of projecting the data onto the first two principal components using PCA. Across all layers, we observe a clear separation between samples from clean (normal) and poisoned (adversarial) point clouds. This separation provides an initial indication of topological differences in the representation space of the final token between normal and adversarial inputs.

To better understand the underlying causes of this separation and the specific topological features contributing to it, we evaluate feature importance in the PCA and also perform a CCA.

We begin by examining the contributions of the features from **Table** 4.3 to the first principal component at various layers. Since the first principal component accounts for most of the variance in the data, it plays a key role in explaining the observed separation. The top three contributing features per layer are as follows:

- **Layer 1**
  - Explained variance: 0.593.
  - Top features: mean of the deaths of 0-bars, standard deviation of the deaths of 0-bars, maximum of the births of 1-bars.

- **Layer 8**
  - Explained variance: 0.491.
  - Top features: mean of the deaths of 0-bars, minimum of the deaths of 0-bars, number of 1-bars.
- **Layer 16**
  - Explained variance: 0.519.
  - Top features: mean of the deaths of 0-bars, number of 1-bars, entropy of 0-bars.
- **Layer 24**
  - Explained variance: 0.964.
  - Top features: mean of the deaths of 0-bars, number of 1-bars, minimum persistence among 1-bars.
- **Layer 32**
  - Explained variance: 0.831.
  - Top features: number of 1-bars, mean of the deaths of 0-bars, mean of the birth/death ratios for 1-bars.

These results highlight that the *mean of the deaths of 0-bars* consistently emerges as one of the most influential features driving the separation between clean and poisoned inputs. To further investigate this, we perform a CCA between the pruned barcode summary features and the PCA coordinates. CCA identifies linear relationships between two multivariate datasets by finding pairs of *canonical variables*, i.e. linear combinations of features from each dataset, that are maximally correlated.

Let $\mathbf{X} \in \mathbb{R}^{K \times n}$ represent the matrix of pruned barcode summaries from the $K = 128$ samples, and $\mathbf{Y} \in \mathbb{R}^{K \times 2}$ the matrix of PCA coordinates for the first two principal components. CCA seeks vectors $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^2$ that maximize the correlation between $\mathbf{Xa}$ and $\mathbf{Yb}$. These canonical variables reveal how strongly each feature relates to the structure captured by PCA.

We compute the *loadings*, which quantify the contribution of each barcode feature to the canonical variables, and present the results in **Figure** 4.8. Once again, we find that the *mean of the deaths of 0-bars* consistently ranks as the most significant feature, reinforcing its central role in distinguishing clean from poisoned samples.

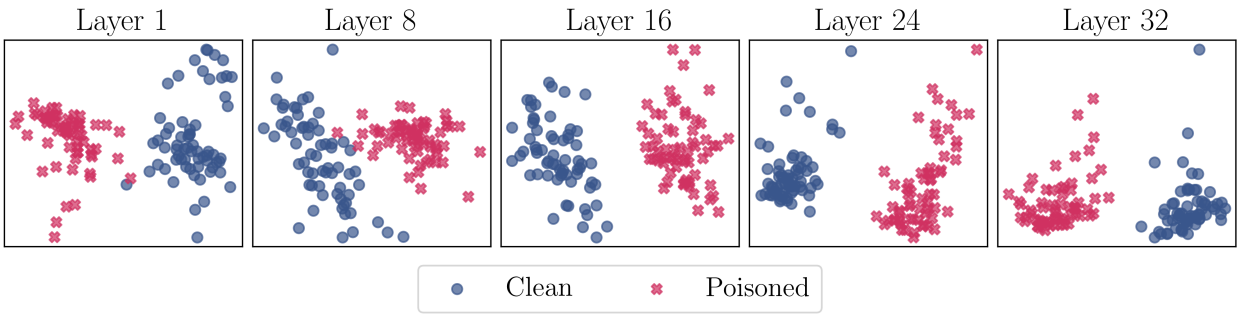


Figure 4.7: PCA of pruned barcode summaries of clean vs. poisoned activations. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32. The explained variance is 0.59, 0.49, 0.52, 0.96 and 0.83, respectively.

Logistic Regression and SHAP Analysis. Given our previous results, we now aim at explaining the topological distinction between the clean and the poisoned data. Our approach to continue investigating this issue is to train a logistic regression model to distinguish between these two classes, which gives us exceptional performance in all layers, and implementing interpretability techniques to understand which features are influencing the perfect prediction and in which ways.

In **Figure** 4.9 we present the results of a logistic regression trained with a 70/30 split between train and test on the pruned barcode summaries to distinguish between clean and poisoned data. These are plotted in the PCA projection for visualization purposes. We obtain perfect accuracy and AUC-ROC when testing on the test data, and 5-fold cross validation over the training data for all layers.

To gain intuition and explain the reasons behind this perfect classification power, we implement SHAP values [LL17], an interpretability tool coming from game theory that helps explain the output of a ML model. We have a logistic regression $f : \mathbb{R}^n \to [0, 1]$ which takes the pruned barcode summaries of some input point $\mathbf{x} \in \mathbb{R}^n$ and outputs the probability of it being classified as poisoned $f(\mathbf{x}) \simeq 1$ or clean $\simeq 0$. We call $\hat{f} \coloneqq \mathbb{E}[f(\mathbf{X})]$ to the average predicted value over our whole dataset $\mathbf{X} \in \mathbb{R}^{K\times n}$, which, if the dataset is balanced, should be around 0.5 for a logistic model. Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be some input data point with n features, that is, some row in $\mathbf{X}$. Without going into the precise definition, the SHAP value for the ith feature, $i \in \{1, \ldots, n\}$ of $\mathbf{x}$, denoted $\text{SHAP}_i(\mathbf{x})$ captures how much that feature deviated the prediction $f(\mathbf{x})$ from the average prediction. In other words, we can write

$$f(\mathbf{x}) = \hat{f} + \sum_{i=1}^{n} \text{SHAP}_i(\mathbf{x}) \cdot x_i.$$

That way, SHAP values help us understand the effect of features in the prediction for each input data point.

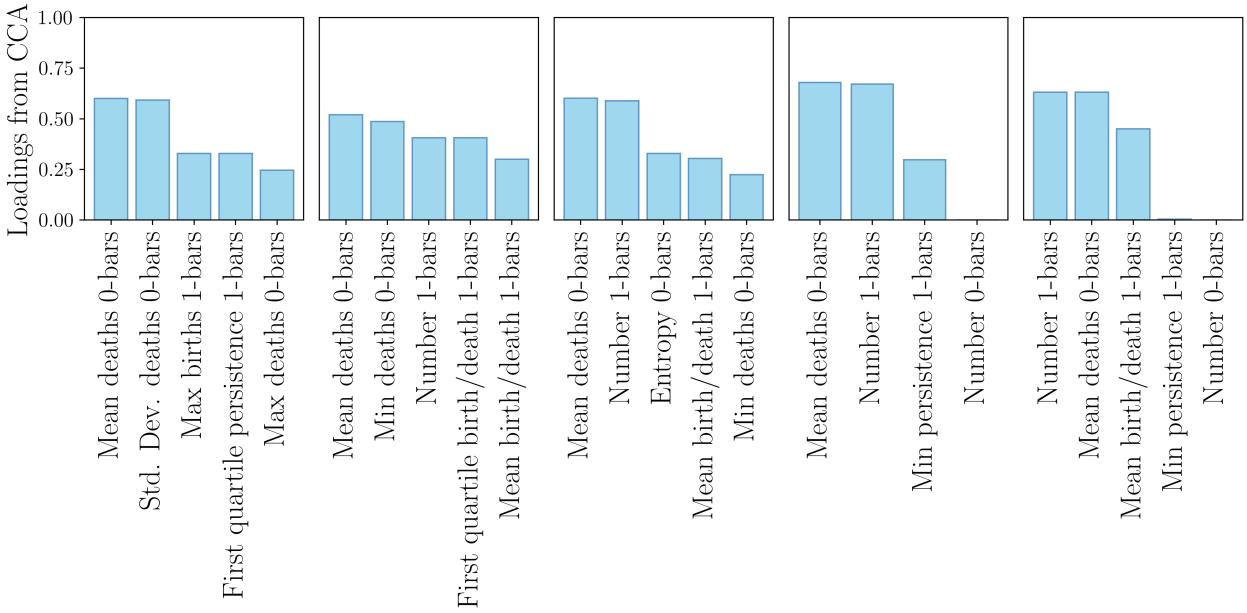


Figure 4.8: Loadings of the first canonical variable in the CCA for clean vs. poisoned activations. These can be interpreted as a measure of feature importance in constructing the canonical variables, which are maximally correlated linear combinations of the features from the pruned barcode summaries and the PCA projections.

In **Figure** 4.10 we showcase beeswarm plots for the results of the SHAP values from the logistic regression displayed in **Figure** 4.9. The SHAP values of the most important 4 features are shown spread over an horizontal axis: negative values on that axis "push" the prediction to 0 (i.e. being classified as clean) and positive values to 1 (i.e. poisoned). Each point in each of this horizontal spreads represents an input data point, and it is colored by the original value of the feature for that point, which helps spotting patterns and correlations between the values of the features, and how having those values affects the obtained prediction.

We observe a clear dichotomous effect of the mean of the deaths of 0-bars in all layers. For layer 0, having high values of this feature tends to push the classification towards 0, whereas

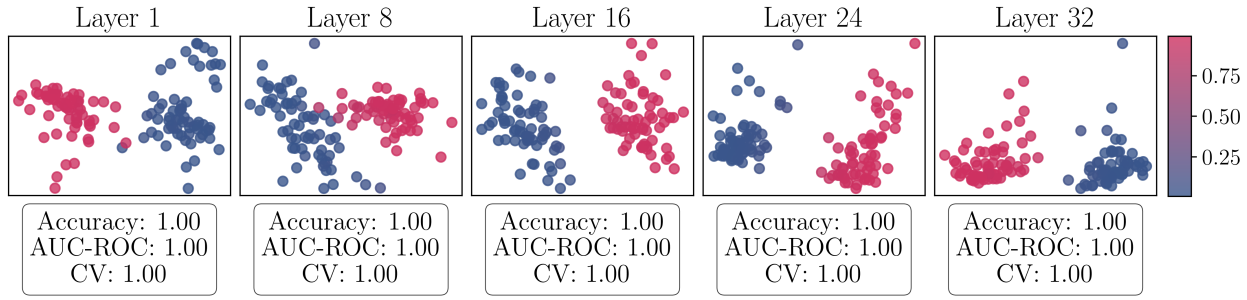


Figure 4.9: Logistic regression for clean vs. poisoned activations trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first PCs. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model.

lower values of this feature deviate the prediction towards 1. This pattern is reversed for all subsequent layers. The effect of other features can be studied in similar ways, we provide a more detailed interpretation of how this analysis helps us describe the topology of clean vs. poisoned activations in the next section.

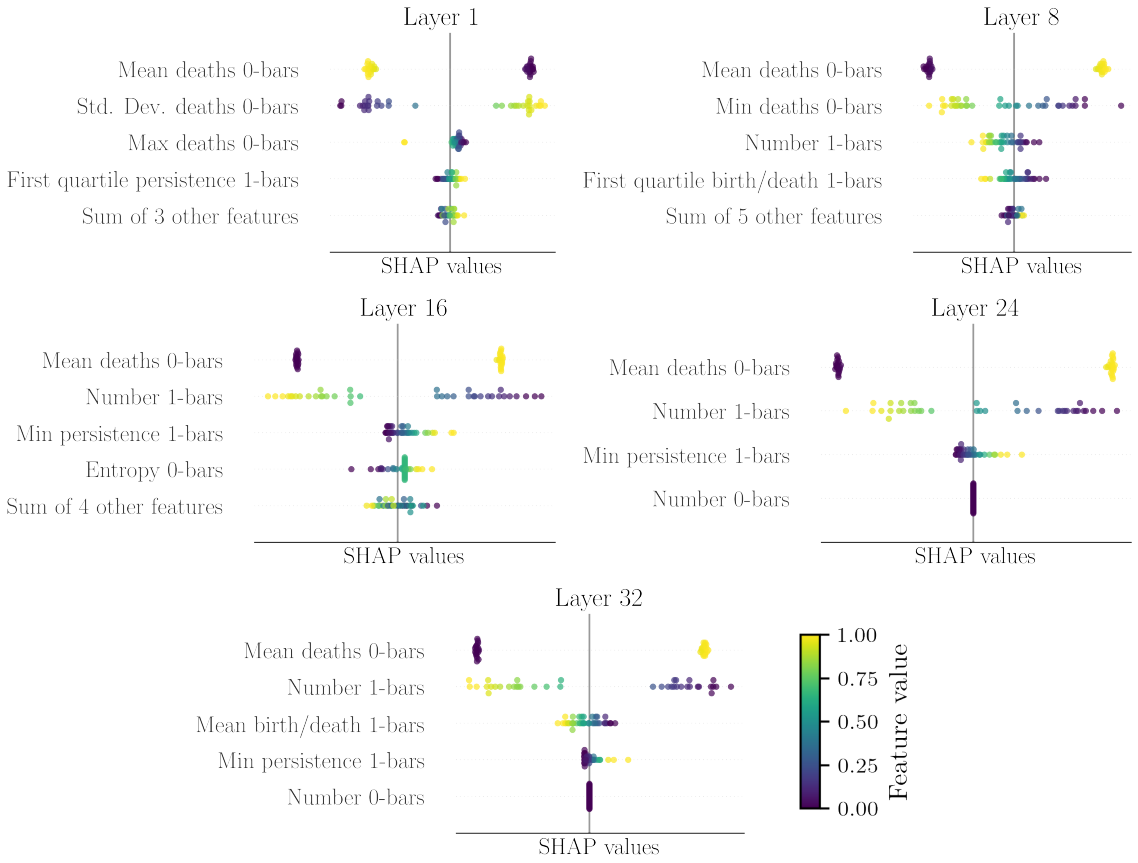


Figure 4.10: SHAP analysis: clean vs. poisoned activations. Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

Interpretation

One thing that is clear from the previous section is that there is an apparent distinction between the barcode summaries of clean and poisoned representations. We now turn to interpret some of the results above, supported by the empirical distributions of statistics in the barcode summaries, to provide intuitions on what are the differences in topology between clean and poisoned representations that are causing this clear separation.

In **Figure** 4.11 we present KDE plots for the empirical distributions of the mean of the births and the persistences of the 1-bars, and the number of 1-bars. We emphasize here that the mean of the death of the 0-bars is highly correlated in most layers with the mean of the birth of 1-bars, which we have just seen had a lot of predictive power. In the KDE plots, we

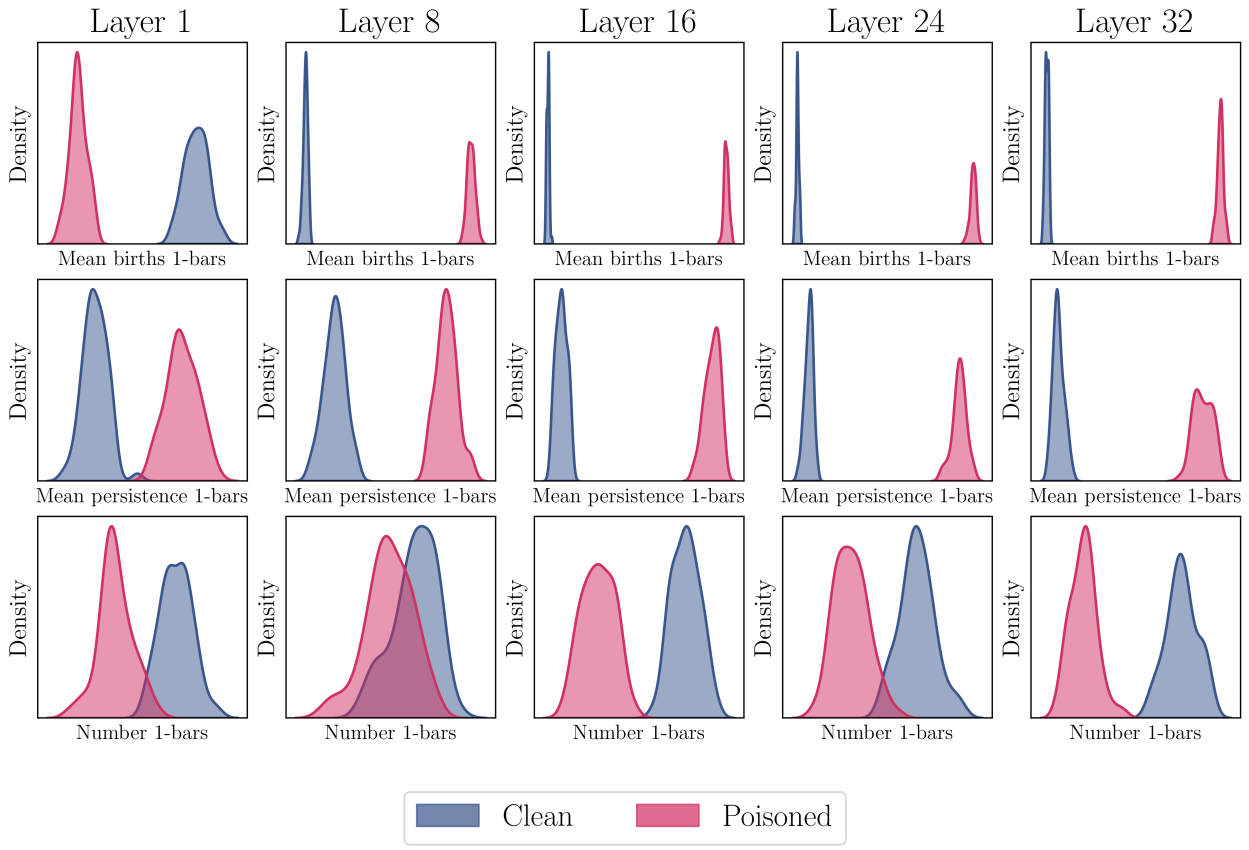


Figure 4.11: KDE plots for the empirical distribution of the mean of the births of 1-bars (above), the mean persistence of 1-bars (center) and the number of 1-bars (below) in Mistral 7B between clean and poisoned activations.

appreciate clear differences between the empirical distributions of these statistics, which on top of the results from the SHAP analysis in **Figure** 4.10 allow the following interpretations.

1. **Compactness vs. dispersion of the data**: In Mistral 7B, what we observe from the empirical distributions of the features in the barcode summaries is that, in the first layer, the poisoned representations tend to be more tightly concentrated, with lower mean death of the 0-bars (cf. **Figure** 4.10), i.e. points typically connect earlier in the filtration because they are closer in the activation space. Similarly, the mean birth of the 1-bars is lower in the poisoned data (cf. **Figure** 4.11, first row), which means that loops tend to form earlier as many points connect for small distances. This behavior reverses for the rest of the layers: in these, it is the poisoned data that seems to be more spread out, whereas the clean data is more compactly distributed across smaller scales. This is also reflected in the fact that loops in the poisoned data tend to die after (cf. **Figure** 4.11, second row), which means that they enclose wider regions.

2. **Topological diversity**: Another key indicator for the separation between clean and poisoned activations in the previous section was the number of 1-bars present in the

data. It seems that poisoned activations tend to form fewer loops than clean activations (cf. **Figure** 4.11, third row). This can reflect a more uniform distribution of the points in the parameter space, whereas the clean activations might be located across more clear directions that enclose regions empty of points which end up forming loops.

In summary, in sight of these results, we can hypothesize that the adversarial mode in this LLM tends to spread out the representations in an even manner, inducing less loops which persist longer; whereas the representations of the model operating under its normal state are more compressed but present a higher number of loops forming and persisting across smaller scales.

The notable observation is that this effect is actually observed across models and adversarial triggers, as presented in the results in the appendices of [Fay+25]. This was a really surprising finding of this project. Although a more thorough investigation is needed to establish any kind of conclusions, this finding might indicate that adversarial triggers *consistently* deform the representation space in LLMs in a unique way, which can be key to detect this modes of operation and design defenses against them in the context of AI security.

5 Conclusion and Outlook

We conclude this thesis with some remarks about the natural progression established in the research presented here, some thoughts about current challenges in the many intersections of topological data analysis explored, and natural next steps that stem from them.

At the Crossroads of Topology, Statistical Inference, and Deep Learning

One of the defining characteristics of topological data analysis that first drew me to the field was its inherently multidisciplinary nature. Given my pure mathematics background, I have always been drawn to elegant theories and complex analytical problems. However, at the beginning of my doctoral studies, I also developed a strong desire to see my research connect with real-world applications. Topological data analysis offered an ideal point of convergence for these aspirations: it could be approached purely from an algebraic topology perspective, working with category theory and homological algebra, while simultaneously presenting many practical challenges related to data analysis and its integration with statistical techniques.

The development of this thesis reflects the trajectory of my doctoral journey. In the early stages, I focused on understanding the foundational theory of PH while engaging in more applied work, such as [GMS24], which provided an opportunity to improve my programming skills. The central role of dualities in the optimization strategies underlying `Ripser`, key for this first project, prompted an initial research question concerning the role of duality in more general settings.

While working on the second major project of my PhD, focused on the stability of rank functions and rank invariants [Wan+24], I encountered two central themes that would shape the remainder of my doctoral research: the theory of multiparameter PH and the significant challenges involved in developing statistical methodologies for PH. This helped crystallize the question of studying dualities in multiparameter persistence, while it also sparked my interest in the broader field of statistical and ML integration.

We soon realized that extending the classical equivalences between persistent homology and cohomology, in the style of [SMV11], to multiparameter persistence was not a straightforward task. Addressing this question required a deeper understanding of multiparameter invariants and the categorical framework used to define them. Simultaneously, it became clear that building a robust statistical foundation was essential for developing meaningful applications and statistical tools for persistence-based methods.

While developing this theoretical and statistical background, I completed a third project, this time at intersection between PH and ML [Tan+24], which allowed me to learn about deep learning theory, and to explore how PH might be useful in understanding the extremely complex systems arising in this field. After that, the final two projects of my doctoral research, [Fay+25] and [GMW25], have continued these two lines of questioning: the intersection of PH and deep learning, and the development of statistical tools for ML, respectively.

Over the course of my doctoral studies, I have explored PH from three perspectives: a theoretical one, a statistical one, and an applied one at the intersection with deep learning theory. This has allowed me to engage with pure mathematics, through the study of category theory and representation theory to be able to establish the duality results in **Section** 2.6; to expand my statistical and computational skills; and to deepen my understanding of the foundations and practical implementation of machine learning systems. Together, these efforts lay a natural foundation for the directions of future work which I outline next.

Further Theoretical Developments in the Theory of Persistence

The preliminary results in **Section** 2.6 can still be further explored, and several research questions arise from them. An initial direction is already pointed in that section. We have established a connection between the minimal projective (right) resolution and the minimal injective (left) resolution of a persistence module and its dual, respectively. However, invariants are primarily computed from projective resolutions, and therefore the current formulation does not yield a connection between the invariants of a module and its dual.

A central question for future work is to bridge this gap. Two potential strategies emerge to address this: a) trying to obtain combinatorial links between the injective and projective resolutions of a given module, potentially restricting the type of posets that we consider as parameter sets; or b) trying to develop alternative formulations of invariants based on injective resolutions.

The latter approach might be challenging as there is a reason for the preference of projectives over injectives. Although projectives and injectives are theoretically dual, in practice, projective objects are usually easier to construct and handle. A typical example is the category of abelian groups, where the projectives are precisely the direct summands of free abelian

groups, while injectives include objects such as $\mathbb{Q}/\mathbb{Z}$, neither finitely generated nor easily described explicitly. We have characterized indecomposable injective modules of p.f.d. modules over finite posets, but a further independent direction to explore is precisely understanding general injective objects in the theory of persistence, which might allow to then pursue the development of invariants based on injectives.

Applications of Multiparameter Persistence

Moving beyond theoretical aspects, although there have been significant efforts in developing new invariants in multiparameter persistence suitable for data analysis; applications of these to real data settings remain underexplored compared to the widespread use of barcodes in practice.

Two promising early applications stand out. In [Vip+21], Vipond *et al.* apply multiparameter persistence landscapes to analyze histopathology images from a particular type of cancer, demonstrating an improvement in performance over traditional PH and spatial statistics methods in detecting spatial patterns of immune cells. In addition, they uncover a potential link between cell codensity and tumor hypoxia (deficiency of oxygen), suggesting that multiparameter persistence may yield biologically meaningful insights. In addition, in [CB20], *multiparameter persistence images* are applied to breast cancer data, leading to improved survival classification accuracy compared to alternative approaches, including nearest-neighbor distributions and standard PH-based techniques.

A potential obstacle for the implementation of further applications of multiparameter persistence has been the lack of unified computational tools, with the available ones being scattered across different libraries. The recent release of `multipers` [LS24], aimed at being integrated with the `gudhi` package [Pro25], represents a significant step towards making multiparameter persistence more accessible for data analysis. Two questions emerge in this context:

- *What data problems are best suited for multiparameter persistence?* One thing that I have learned during my doctoral studies is that, while incredibly useful in some settings, PH is not the answer to all data problems: it is most effective when shape plays a meaningful role. While one motivation for multiparameter persistence is to enhance robustness to noise, it remains an open question whether there are problem domains where the inclusion of additional filtration parameters offers substantial benefits beyond one-parameter approaches. Identifying and implementing such applications represents an exciting avenue for further research.

- *What type of statistical tests and methodologies can we implement with this theory?* My growing interest in statistical tools and methods leads me to seek principled statistical

techniques for multiparameter persistence when applying it to real data analysis task. Some preliminary results in this direction have been established in [GMW25], but I am eager to continue exploring this intersection, particularly motivated by real data scenarios.

Integrating Topology and Geometry in Machine Learning

Finally, in light of my future position as a postdoctoral researcher at the University of Fribourg, as part of the AIDOS Lab (AI for Data-Oriented Science) led by Professor Bastian Rieck, a natural direction for further research concerns the broader question about the role of topology and geometry in deep learning.

As an initial step, I aim to expand my toolkit and knowledge of geometric and topological methods. During my doctoral journey, I have focused in PH and its multiple invariants, however, there are many other tools of geometric or topological nature that I would like to explore: Euler transforms [GLM18; RR23; Mun25], discrete notions of curvature [For03; Oll07], or magnitude [GH21], among others.

In parallel, I intend to continue the development of my understanding of deep learning theory, with the goal of identifying areas where geometric and topological methods can address existing challenges. There are two completely new directions where I would like to expand my experience: the development of more efficient data representations and neural architectures informed by topology and geometry, and the design of deep learning frameworks and models operating on intrinsically topological/combinatorial domains. However, I anticipate that the interaction with the group in Fribourg will shape, refine and potentially redirect some of these preliminary ideas.

Abbreviations and Acronyms

Abbreviation	Definition
AI	Artificial Intelligence
CHD	California Housing Dataset; median house values with 8 predictive features
CIFAR-10	Dataset of 60,000 32×32 color images in 10 object classes
CIFAR-100	Dataset of 60,000 32×32 color images in 100 object classes
CMI	Conditional Mutual Information
FCN	Fully Connected Network
FDA	Functional Data Analysis
FCLT	Functional Central Limit Theorem
KDE	Kernel Density Estimation
k-NN	k-Nearest Neighbours
LSTM	Long Short-Term Memory
MDB	Maximum Depth Band
ML	Machine Learning
MNIST	Dataset of 70,000 handwritten digits (28×28 grayscale images)
MST	Minimal Spanning Tree
NLP	Natural Language Processing
NN	Neural Network

Abbreviation	Definition
PH	Persistent Homology
p.f.d.	Point-wise finite-dimensional
pLLP	posterior lateral line primordium
r.v.	random variable
RNN	Recurrent Neural Networks
SGD	Stochastic Gradient Descent
VR	Vietoris–Rips
XPIA	Cross/Indirect Prompt Injection Attack

List of Figures

List of Tables

Bibliography

[Abd+24] Marah I. Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, et al. "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone". In: (Aug. 2024) (cit. on p. 93).

[Abd+25] Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. "Get My Drift? Catching LLM Task Drift with Activation Deltas". In: *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Apr. 2025, pp. 43–67. DOI: 10.1109/SaTML64287.2025.00011 (cit. on p. 93).

[Ada+17] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. "Persistence Images: A Stable Vector Representation of Persistent Homology". In: *Journal of Machine Learning Research* 18.8 (2017), pp. 1–35. ISSN: 1533-7928 (cit. on p. 29).

[Ada+20] Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Joshua Mirth, Rachel Neville, Chris Peterson, and Clayton Shonkwiler. "A Fractal Dimension for Measures via Persistent Homology". In: *Topological Data Analysis*. Ed. by Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule. Cham: Springer International Publishing, 2020, pp. 1–31. ISBN: 978-3-030-43408-3. DOI: 10.1007/978-3-030-43408-3_1 (cit. on p. 78).

[AGS20] Nieves Atienza, Rocio Gonzalez-Díaz, and Manuel Soriano-Trigueros. "On the Stability of Persistent Entropy and New Summary Functions for Topological Data Analysis". In: *Pattern Recognition* 107 (Nov. 2020), p. 107509. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2020.107509 (cit. on p. 29).

[Ali+23] Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. "A Survey of Vectorization Methods in Topological Data Analysis". In: *IEEE Transactions on Pattern Analysis and*

Machine Intelligence 45.12 (Dec. 2023), pp. 14069–14080. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3308391 (cit. on p. 29).

[AS93] Maurice Auslander and Øyvind Solberg. "Relative Homology and Representation Theory 1: Reative Homology and Homologically Finite Subcategories". In: *Communications in Algebra* 21.9 (1993), pp. 2995–3031 (cit. on pp. 59, 60).

[Azu50] Gorô Azumaya. "Corrections and Supplementaries to My Paper Concerning Krull-Remak-Schmidt's Theorem". In: *Nagoya Mathematical Journal* 1 (June 1950), pp. 117–124. ISSN: 0027-7630, 2152-6842. DOI: 10.1017/S002776300002290X (cit. on pp. 23, 53).

[Bak21] Håvard Bakke Bjerkevik. "On the Stability of Interval Decomposable Persistence Modules". In: *Discrete & Computational Geometry* 66.1 (July 2021), pp. 92–121. ISSN: 1432-0444. DOI: 10.1007/s00454-021-00298-0 (cit. on p. 33).

[Bau+17] Ulrich Bauer, Michael Kerber, Jan Reininghaus, and Hubert Wagner. "Phat – Persistent Homology Algorithms Toolbox". In: *Journal of Symbolic Computation*. Algorithms and Software for Computational Topology 78 (Jan. 2017), pp. 76–90. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2016.03.008 (cit. on pp. 36, 45).

[Bau21] Ulrich Bauer. "Ripser: Efficient Computation of Vietoris–Rips Persistence Barcodes". In: *Journal of Applied and Computational Topology* 5.3 (Sept. 2021), pp. 391–423. ISSN: 2367-1734. DOI: 10.1007/s41468-021-00071-5 (cit. on pp. 12, 26, 36, 38).

[BBH22] Benjamin Blanchette, Thomas Brüstle, and Eric J. Hanson. "Homological Approximations in Persistence Theory". In: *Canadian Journal of Mathematics* (Dec. 2022), pp. 1–38. ISSN: 0008-414X, 1496-4279. DOI: 10.4153/S0008414X22000657 (cit. on pp. 59, 60).

[BBH23] Benjamin Blanchette, Thomas Brüstle, and Eric J. Hanson. *Exact Structures for Persistence Modules*. Aug. 2023. DOI: 10.48550/arXiv.2308.01790. arXiv: 2308.01790 [cs, math] (cit. on pp. 53, 54, 59).

[BBK20] Håvard Bakke Bjerkevik, Magnus Bakke Botnan, and Michael Kerber. "Computing the Interleaving Distance Is NP-Hard". In: *Foundations of Computational Mathematics* 20.5 (Oct. 2020), pp. 1237–1271. ISSN: 1615-3383. DOI: 10.1007/s10208-019-09442-y (cit. on p. 33).

[BC20] Magnus Botnan and William Crawley-Boevey. "Decomposition of Persistence Modules". In: *Proceedings of the American Mathematical Society* 148.11 (Aug. 2020), pp. 4581–4596. ISSN: 0002-9939, 1088-6826. DOI: 10.1090/proc/14790 (cit. on p. 23).

[Ben+23] Katherine Benjamin, Lamisah Mukta, Gabriel Moryoussef, Christopher Uren, Heather A. Harrington, Ulrike Tillmann, and Agnese Barbensi. "Homology of Homologous Knotted Proteins". In: *Journal of The Royal Society Interface* 20.201 (Apr. 2023), p. 20220727. DOI: 10.1098/rsif.2022.0727 (cit. on p. 10).

[BGK15] Subhrajit Bhattacharya, Robert Ghrist, and Vijay Kumar. "Persistent Homology for Path Planning in Uncertain Environments". In: *IEEE Transactions on Robotics* 31.3 (June 2015), pp. 578–590. ISSN: 1552-3098, 1941-0468. DOI: 10.1109/TRO.2015.2412051 (cit. on p. 11).

[Bir+21] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. "Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 6776–6789 (cit. on pp. 81, 82, 85).

[BL08] Peter A Brooksbank and Eugene M Luks. "Testing Isomorphism of Modules". In: *Journal of Algebra* 320.11 (2008), pp. 4020–4029 (cit. on p. 33).

[BL15] Ulrich Bauer and Michael Lesnick. "Induced Matchings and the Algebraic Stability of Persistence Barcodes". In: *Journal of Computational Geometry* (Mar. 2015), 162–191 Pages. DOI: 10.20382/jocg.v6i2a9. arXiv: 1311.3681 (cit. on pp. 24, 33, 37, 38).

[BL23] Magnus Botnan and Michael Lesnick. "An Introduction to Multiparameter Persistence". In: *Representations of Algebras and Related Structures*. Ed. by Aslak Bakke Buan, Henning Krause, and Øyvind Solberg. EMS Press, 2023, pp. 77–150. ISBN: 978-3-9854705-4-9. DOI: 10.4171/ecr/19/4 (cit. on pp. 18, 23, 34).

[BLL23] Ulrich Bauer, Fabian Lenzen, and Michael Lesnick. "Efficient Two-Parameter Persistence Computation via Cohomology". In: *39th International Symposium on Computational Geometry (SoCG 2023)*. Schloss-Dagstuhl - Leibniz Zentrum für Informatik, 2023. DOI: 10.4230/LIPIcs.SoCG.2023.15 (cit. on p. 36).

[Blu+24] Andrew J. Blumberg, Mathieu Carrière, Jun Hou Fung, and Michael A. Mandell. *Subsampling, Aligning, and Averaging to Find Circular Coordinates in Recurrent Time Series*. Dec. 2024. DOI: 10.48550/arXiv.2412.18515. arXiv: 2412.18515 [stat] (cit. on p. 36).

[BM02] Peter L. Bartlett and Shahar Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482. ISSN: ISSN 1533-7928 (cit. on p. 80).

[BM21] Peter Bubenik and Nikola Milićević. "Homological Algebra for Persistence Modules". In: *Foundations of Computational Mathematics* 21.5 (Oct. 2021), pp. 1233–1278. ISSN: 1615-3383. DOI: 10.1007/s10208-020-09482-9 (cit. on p. 54).

[BOO22] Magnus Bakke Botnan, Steffen Oppermann, and Steve Oudot. *Signed Barcodes for Multi-Parameter Persistence via Rank Decompositions and Rank-Exact Resolutions*. Mar. 2022. DOI: 10.48550/arXiv.2107.06800. arXiv: 2107.06800 [cs, math] (cit. on p. 30).

[Bri92] John Briggs. *Fractals: The Patterns of Chaos : A New Aesthetic of Art, Science, and Nature*. Simon and Schuster, 1992. ISBN: 978-0-671-74217-1 (cit. on p. 77).

[Bro62] Edgar H. Brown. "Cohomology Theories". In: *Annals of Mathematics* 75.3 (1962), pp. 467–484. ISSN: 0003-486X. DOI: 10.2307/1970209. JSTOR: 1970209 (cit. on p. 36).

[BS22] Ulrich Bauer and Maximilian Schmahl. *Efficient Computation of Image Persistence*. Jan. 2022. DOI: 10.48550/arXiv.2201.04170. arXiv: 2201.04170 [math] (cit. on p. 43).

[BS23] Ulrich Bauer and Maximilian Schmahl. "Lifespan Functors and Natural Dualities in Persistent Homology". In: *Homology, Homotopy and Applications* 25.2 (2023), pp. 297–327. ISSN: 15320073, 15320081. DOI: 10.4310/HHA.2023.v25.n2.a13. arXiv: 2012.12881 [math] (cit. on pp. 36, 38, 40–43, 50, 54).

[BSS23] Peter Bubenik, Jonathan Scott, and Donald Stanley. "Exact Weights, Path Metrics, and Algebraic Wasserstein Distances". In: *Journal of Applied and Computational Topology* 7.2 (June 2023), pp. 185–219. ISSN: 2367-1734. DOI: 10.1007/s41468-022-00103-8 (cit. on p. 33).

[Bub15] Peter Bubenik. "Statistical Topological Data Analysis Using Persistence Landscapes". In: *Journal of Machine Learning Research* 16.3 (2015), pp. 77–102. ISSN: 1533-7928 (cit. on pp. 27, 29, 30, 71).

[Büh10] Theo Bühler. "Exact Categories". In: *Expositiones Mathematicae* 28.1 (2010), pp. 1–69. ISSN: 07230869. DOI: 10.1016/j.exmath.2009.04.004 (cit. on p. 59).

[BW20] David Bramer and Guo-Wei Wei. "Atom-Specific Persistent Homology and Its Application to Protein Flexibility Analysis". In: *Computational and Mathematical Biophysics* 8.1 (Jan. 2020), pp. 1–35. ISSN: 2544-7297. DOI: 10.1515/cmb-2020-0001 (cit. on p. 32).

[CB20] Mathieu Carrière and Andrew Blumberg. "Multiparameter Persistence Image for Topological Machine Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 22432–22444 (cit. on pp. 30, 105).

[CCd16] Frédéric Chazal, William Crawley-Boevey, and Vin de Silva. "The Observable Structure of Persistence Modules". In: *Homology, Homotopy and Applications* 18.2 (2016), pp. 247–265 (cit. on p. 24).

[CEH05] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of Persistence Diagrams”. In: *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*. SCG ’05. New York, NY, USA: Association for Computing Machinery, June 2005, pp. 263–271. ISBN: 978-1-58113-991-4. DOI: 10.1145/1064092.1064133 (cit. on pp. 29–32).

[CEH07] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of Persistence Diagrams”. In: *Discrete & Computational Geometry* 37.1 (Jan. 2007), pp. 103–120. ISSN: 1432-0444. DOI: 10.1007/s00454-006-1276-5 (cit. on pp. 31, 32).

[Cha+09] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. “Proximity of Persistence Modules and Their Diagrams”. In: *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*. SCG ’09. New York, NY, USA: Association for Computing Machinery, June 2009, pp. 237–246. ISBN: 978-1-60558-501-7. DOI: 10.1145/1542362.1542407 (cit. on pp. 29, 30, 32).

[Cha+14a] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. *On the Bootstrap for Persistence Diagrams and Landscapes*. Jan. 2014. DOI: 10.48550/arXiv.1311.0376. arXiv: 1311.0376 [math] (cit. on pp. 70–72).

[Cha+14b] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. “Stochastic Convergence of Persistence Landscapes and Silhouettes”. In: *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*. SOCG’14. New York, NY, USA: Association for Computing Machinery, June 2014, pp. 474–483. ISBN: 978-1-4503-2594-3. DOI: 10.1145/2582112.2582128 (cit. on pp. 29, 70, 72).

[Cha+16] Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-42543-6 978-3-319-42545-0. DOI: 10.1007/978-3-319-42545-0 (cit. on p. 24).

[Cha+24] Wojciech Chachólski, Andrea Guidolin, Isaac Ren, Martina Scolamiero, and Francesca Tombari. “Koszul Complexes and Relative Homological Algebra of Functors Over Posets”. In: *Foundations of Computational Mathematics* (June 2024). ISSN: 1615-3383. DOI: 10.1007/s10208-024-09660-z (cit. on pp. 59, 61).

[Chi+15] Harish Chintakunta, Thanos Gentimis, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Hamid Krim. “An Entropy-Based Persistence Barcode”. In: *Pattern Recognition* 48.2 (2015), pp. 391–401 (cit. on pp. 29, 94).

[CK11] Chao Chen and Michael Kerber. "Persistent Homology Computation with a Twist". In: (2011) (cit. on p. 26).

[Coh+09] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy Morozov. "Persistent Homology for Kernels, Images, and Cokernels". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2009, pp. 1011–1020 (cit. on pp. 38, 42).

[Con+22] Marco Contessoto, Facundo Mémoli, Anastasios Stefanou, and Ling Zhou. "Persistent Cup-Length". In: *38th International Symposium on Computational Geometry (SoCG 2022)*. Ed. by Xavier Goaoc and Michael Kerber. Vol. 224. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, 31:1–31:17. ISBN: 978-3-95977-227-3. DOI: 10.4230/LIPIcs.SoCG.2022.31 (cit. on p. 36).

[Cor+19] René Corbet, Ulderico Fugacci, Michael Kerber, Claudia Landi, and Bei Wang. "A Kernel for Multi-Parameter Persistent Homology". In: *Computers & Graphics: X* 2 (Dec. 2019), p. 100005. ISSN: 2590-1486. DOI: 10.1016/j.cagx.2019.100005 (cit. on p. 30).

[CP92] Paul H. Coleman and Luciano Pietronero. "The Fractal Structure of the Universe". In: *Physics Reports* 213.6 (May 1992), pp. 311–389. ISSN: 0370-1573. DOI: 10.1016/0370-1573(92)90112-D (cit. on p. 77).

[Cra+20] Lorin Crawford, Anthea Monod, Andrew X. Chen, Sayan Mukherjee, and Raúl Rabadán. "Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis". In: *Journal of the American Statistical Association* 115.531 (July 2020), pp. 1139–1150. ISSN: 0162-1459. DOI: 10.1080/01621459.2019.1671198 (cit. on p. 10).

[Cra15] William Crawley-Boevey. "Decomposition of Pointwise Finite-Dimensional Persistence Modules". In: *Journal of Algebra and Its Applications* 14.05 (June 2015), p. 1550066. ISSN: 0219-4988. DOI: 10.1142/S0219498815500668 (cit. on p. 24).

[Cra94] William Crawley-Boevey. "Locally Finitely Presented Additive Categories". In: *Communications in Algebra* 22.5 (Jan. 1994), pp. 1641–1674. ISSN: 0092-7872, 1532-4125. DOI: 10.1080/00927879408824927 (cit. on p. 23).

[CZ09] Gunnar Carlsson and Afra Zomorodian. "The Theory of Multidimensional Persistence". In: *Discrete & Computational Geometry* 42.1 (July 2009), pp. 71–93. ISSN: 1432-0444. DOI: 10.1007/s00454-009-9176-0 (cit. on pp. 27, 28).

[DDS23] Benjamin Dupuis, George Deligiannidis, and Umut Simsekli. "Generalization Bounds Using Data-Dependent Fractal Dimensions". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 8922–8968 (cit. on pp. 81–83, 85).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cit. on p. 91).

[dFL06] Michele d'Amico, Patrizio Frosini, and Claudia Landi. "Using Matching Distance in Size Theory: A Survey". In: *International Journal of Imaging Systems and Technology* 16.5 (2006), pp. 154–161. ISSN: 1098-1098. DOI: 10.1002/ima.20076 (cit. on p. 34).

[dFL10] Michele d'Amico, Patrizio Frosini, and Claudia Landi. "Natural Pseudo-Distance and Optimal Matching between Reduced Size Functions". In: *Acta Applicandae Mathematicae* 109.2 (Feb. 2010), pp. 527–554. ISSN: 1572-9036. DOI: 10.1007/s10440-008-9332-1 (cit. on p. 34).

[dG07] Vin de Silva and Robert Ghrist. "Coverage in Sensor Networks via Persistent Homology". In: *Algebraic & Geometric Topology* 7.1 (Apr. 2007), pp. 339–358. ISSN: 1472-2739, 1472-2747. DOI: 10.2140/agt.2007.7.339 (cit. on p. 11).

[dMV11] Vin de Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. "Persistent Cohomology and Circular Coordinates". In: *Discrete & Computational Geometry* 45.4 (June 2011), pp. 737–759. ISSN: 1432-0444. DOI: 10.1007/s00454-011-9344-x (cit. on p. 36).

[DR24] Tamal K. Dey and Abhishek Rathod. "Cup Product Persistence and Its Efficient Computation". In: *40th International Symposium on Computational Geometry (SoCG 2024)*. Ed. by Wolfgang Mulzer and Jeff M. Phillips. Vol. 293. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 50:1–50:15. ISBN: 978-3-95977-316-4. DOI: 10.4230/LIPIcs.SoCG.2024.50 (cit. on p. 36).

[Dub+24] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. "The Llama 3 Herd of Models". In: *CoRR* (Jan. 2024) (cit. on p. 93).

[Eas+23] Ty O. Easley, Kevin Freese, Elizabeth Munch, and Janine Diane Bijsterbosch. "Using Persistent Homology to Understand Dimensionality Reduction in Resting-State fMRI". In: (Dec. 2023) (cit. on p. 14).

[EH10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Soc., 2010. ISBN: 978-0-8218-4925-5 (cit. on p. 15).

[ELZ02] Edelsbrunner, Letscher, and Zomorodian. "Topological Persistence and Simplification". In: *Discrete & Computational Geometry* 28.4 (Nov. 2002), pp. 511–533. ISSN: 1432-0444. DOI: 10.1007/s00454-002-2885-2 (cit. on pp. 25, 28).

[EM45a] Samuel Eilenberg and Saunders MacLane. "General Theory of Natural Equivalences". In: *Transactions of the American Mathematical Society* 58 (1945), pp. 231–294. ISSN: 0002-9947, 1088-6850. DOI: 10.1090/S0002-9947-1945-0013131-6 (cit. on p. 56).

[EM45b] Samuel Eilenberg and Saunders MacLane. "Relations Between Homology and Homotopy Groups of Spaces". In: *Annals of Mathematics* 46.3 (1945), pp. 480–509. ISSN: 0003-486X. DOI: 10.2307/1969165. JSTOR: 1969165 (cit. on p. 36).

[ESR16] Kevin Emmett, Benjamin Schweinhart, and Raul Rabadan. "Multiscale Topology of Chromatin Folding". In: *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS)*. BICT'15. Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), May 2016, pp. 177–180. ISBN: 978-1-63190-100-3. DOI: 10.4108/eai.3-12-2015.2262453 (cit. on p. 11).

[ET94] Bradley Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC, May 1994. ISBN: 978-0-429-24659-3. DOI: 10.1201/9780429246593 (cit. on pp. 72, 73).

[Fal13] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, Dec. 2013. ISBN: 978-1-118-76286-8 (cit. on p. 77).

[Fas+14] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. "Confidence Sets for Persistence Diagrams". In: *The Annals of Statistics* 42.6 (Dec. 2014). ISSN: 0090-5364. DOI: 10.1214/14-AOS1252. arXiv: 1303.7117 [math] (cit. on p. 27).

[Fay+25] Aideen Fay, Inés García-Redondo, Qiquan Wang, Haim Dubossarsky, and Anthea Monod. *Holes in Latent Space: Topological Signatures Under Adversarial Influence*. May 2025. DOI: 10.48550/arXiv.2505.20435. arXiv: 2505.20435 [cs] (cit. on pp. 13, 76, 89, 91–93, 95, 102, 104).

[FL99] Patrizio Frosini and Claudia Landi. "Size Theory as a Topological Tool for Computer Vision". In: *Pattern Recognition and Image Analysis* 9.4 (1999), pp. 596–603 (cit. on p. 29).

[For03] Forman. "Bochner's Method for Cell Complexes and Combinatorial Ricci Curvature". In: *Discrete & Computational Geometry* 29.3 (Feb. 2003), pp. 323–374. ISSN: 1432-0444. DOI: 10.1007/s00454-002-0743-x (cit. on p. 106).

[Fos+19] Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. "Hypothesis Set Stability and Generalization". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 80).

[Fro90] Patrizio Frosini. "A Distance for Similarity Classes of Submanifolds of a Euclidean Space". In: *Bulletin of the Australian Mathematical Society* 42.3 (Dec. 1990), pp. 407–415. ISSN: 1755-1633, 0004-9727. DOI: 10.1017/S0004972700028574 (cit. on p. 28).

[Fro92a] Patrizio Frosini. "Discrete Computation of Size Functions". In: *Journal of Combinatorics, Information & System Sciences* 17.3-4 (1992), pp. 232–250. ISSN: 0250-9628 (cit. on p. 28).

[Fro92b] Patrizio Frosini. "Measuring Shapes by Size Functions". In: *Intelligent Robots and Computer Vision X: Algorithms and Techniques*. Vol. 1607. SPIE, Feb. 1992, pp. 122–133. DOI: 10.1117/12.57059 (cit. on p. 28).

[Gam+14] Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramár, Konstantin Mischaikow, and Vidit Nanda. "A Topological Measurement of Protein Compressibility". In: *Japan Journal of Industrial and Applied Mathematics* 32 (Mar. 2014), pp. 1–17. DOI: 10.1007/s13160-014-0153-5 (cit. on p. 11).

[GG02] Evarist Giné and Armelle Guillou. "Rates of Strong Uniform Consistency for Multivariate Kernel Density Estimators". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. Vol. 38. Elsevier, 2002, pp. 907–921 (cit. on pp. 73, 74).

[GH10] Jennifer Gamble and Giseon Heo. "Exploring Uses of Persistent Homology for Statistical Analysis of Landmark-Based Shape Data". In: *Journal of Multivariate Analysis* 101.9 (Oct. 2010), pp. 2184–2199. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2010.04.016 (cit. on p. 32).

[GH21] Dejan Govc and Richard Hepworth. "Persistent Magnitude". In: *Journal of Pure and Applied Algebra* 225.3 (Mar. 2021), p. 106517. ISSN: 00224049. DOI: 10.1016/j.jpaa.2020.106517. arXiv: 1911.11016 [math] (cit. on p. 106).

[Ghr14] Robert Ghrist. *Elementary Applied Topology*. 1st edition. s. l.: CreateSpace Independent Publishing Platform, Sept. 2014. ISBN: 978-1-5028-8085-7 (cit. on p. 15).

[Gid17] Marian Gidea. "Topological Data Analysis of Critical Transitions in Financial Networks". In: *3rd International Winter School and Conference on Network Science*. Ed. by Erez Shmueli, Baruch Barzel, and Rami Puzis. Springer Proceedings in Complexity. Cham: Springer International Publishing, 2017, pp. 47–59. ISBN: 978-3-319-55471-6. DOI: 10.1007/978-3-319-55471-6_5 (cit. on p. 32).

[GLM18] Robert Ghrist, Rachel Levanger, and Huy Mai. "Persistent Homology and Euler Integral Transforms". In: *Journal of Applied and Computational Topology* 2.1 (Oct. 2018), pp. 55–60. ISSN: 2367-1734. DOI: 10.1007/s41468-018-0017-1 (cit. on p. 106).

[GMS24] Inés García-Redondo, Anthea Monod, and Anna Song. "Fast Topological Signal Identification and Persistent Cohomological Cycle Matching". In: *Journal of Applied and Computational Topology* 8.3 (Sept. 2024), pp. 695–726. ISSN: 2367-1734. DOI: 10.1007/s41468-024-00179-4 (cit. on pp. 12, 35, 43, 44, 46, 47, 49, 50, 103).

[GMW25] Inés García-Redondo, Anthea Monod, and Qiquan Wang. *Confidence Bands for Multiparameter Persistence Landscapes*. Apr. 2025. DOI: 10.48550/arXiv.2504.01113. arXiv: 2504.01113 [math] (cit. on pp. 13, 14, 63, 70–73, 75, 104, 106).

[Gom+22] Tristan Gomez, Magalie Feyeux, Justine Boulant, Nicolas Normand, Laurent David, Perrine Paul-Gilloteaux, Thomas Fréour, and Harold Mouchère. "A Time-Lapse Embryo Dataset for Morphokinetic Parameter Prediction". In: *Data in Brief* 42 (May 2022), p. 108258. ISSN: 2352-3409. DOI: 10.1016/j.dib.2022.108258 (cit. on p. 49).

[Gre+24] Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. "Stress-Testing Capability Elicitation With Password-Locked Models". In: *Advances in Neural Information Processing Systems* 37 (Dec. 2024), pp. 69144–69175 (cit. on p. 92).

[GST23] R. Gonzalez-Diaz, M. Soriano-Trigueros, and A. Torras-Casas. "Partial Matchings Induced by Morphisms between Persistence Modules". In: *Computational Geometry* 112 (June 2023), p. 101985. ISSN: 0925-7721. DOI: 10.1016/j.comgeo.2023.101985 (cit. on p. 37).

[Ham+22] W. Hamilton, J. E. Borgert, T. Hamelryck, and J. S. Marron. "Persistent Topology of Protein Space". In: *Research in Computational Topology 2*. Ed. by Ellen Gasparovic, Vanessa Robins, and Katharine Turner. Association for Women in Mathematics Series. Cham: Springer International Publishing, 2022, pp. 223–244. ISBN: 978-3-030-95519-9. DOI: 10.1007/978-3-030-95519-9_10 (cit. on p. 32).

[Har+19] Heather A. Harrington, Nina Otter, Hal Schenck, and Ulrike Tillmann. "Stratifying Multiparameter Persistent Homology". In: *SIAM Journal on Applied Algebra and Geometry* 3.3 (Jan. 2019), pp. 439–471. DOI: 10.1137/18M1224350 (cit. on p. 30).

[Har+20] Jonas Hartmann, Mie Wong, Elisa Gallo, and Darren Gilmour. "An Image-Based Data-Driven Analysis of Cellular Architecture in a Developing Tissue". In: *eLife* 9 (June 2020). Ed. by Marianne E Bronner, Lilianna Solnica-Krezel, and Ajay B Chitnis, e55913. ISSN: 2050-084X. DOI: 10.7554/eLife.55913 (cit. on pp. 37, 47).

[Has+09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer, 2009 (cit. on p. 80).

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. ISBN: 978-0-521-79540-1 (cit. on p. 20).

[Hof+17] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. "Deep Learning with Topological Signatures". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 27).

[Hu+21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. Oct. 2021 (cit. on p. 93).

[Jia+23] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. *Mistral 7B*. Oct. 2023. DOI: 10.48550/arXiv.2310.06825. arXiv: 2310.06825 [cs] (cit. on p. 93).

[Jia+24] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. *Mixtral of Experts*. Jan. 2024. DOI: 10.48550/arXiv.2401.04088. arXiv: 2401.04088 [cs] (cit. on p. 93).

[KLS06] Gady Kozma, Zvi Lotker, and Gideon Stupp. "The Minimal Spanning Tree and the Upper Box Dimension". In: *Proceedings of the American Mathematical Society* 134.4 (2006), pp. 1183–1187. ISSN: 0002-9939, 1088-6826. DOI: 10.1090/S0002-9939-05-08061-5 (cit. on p. 79).

[KM21] Woojin Kim and Facundo Mémoli. "Generalized Persistence Diagrams for Persistence Modules over Posets". In: *Journal of Applied and Computational Topology* 5.4 (Dec. 2021), pp. 533–581. ISSN: 2367-1734. DOI: 10.1007/s41468-021-00075-1 (cit. on p. 30).

[Kov+16] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. "Using Persistent Homology and Dynamical Distances to Analyze Protein Binding". In: *Statistical Applications in Genetics and Molecular Biology* 15.1 (Jan. 2016). ISSN: 1544-6115, 2194-6302. DOI: 10.1515/sagmb-2015-0057. arXiv: 1412.1394 [math, q-bio, stat] (cit. on p. 11).

[LCC01] P. S. Liao, T. S. Chen, and Pau-Choo Chung. "A Fast Algorithm for Multilevel Thresholding". In: *Journal of Information Science and Engineering* 17.5 (Sept. 2001), pp. 713–727. ISSN: 1016-2364 (cit. on pp. 47, 49).

[Les15] Michael Lesnick. "The Theory of the Interleaving Distance on Multidimensional Persistence Modules". In: *Foundations of Computational Mathematics* 15.3 (June 2015), pp. 613–650. ISSN: 1615-3375, 1615-3383. DOI: 10.1007/s10208-015-9255-y. arXiv: 1106.5305 [cs, math] (cit. on p. 33).

[LF97] Claudia Landi and Patrizio Frosini. "New Pseudodistances for the Size Function Space". In: *Vision Geometry VI*. Vol. 3168. SPIE, Oct. 1997, pp. 52–60. DOI: 10.1117/12.279674 (cit. on p. 29).

[LF99] Claudia Landi and P Frosini. "Algebraic Representation of Size Functions". In: *Pattern Recognition and Image Understanding, 5th Open German-Russian Workshop*. Infix, 1999, pp. 41–46 (cit. on p. 29).

[Li+24] Nathaniel Li, Alexander Pan, Anjali Gopal, et al. *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning*. May 2024. DOI: 10.48550/arXiv.2403.03218. arXiv: 2403.03218 [cs] (cit. on p. 93).

[LL17] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 98).

[LMT22] Umberto Lupo, Anibal M. Medina-Mardones, and Guillaume Tauzin. "Persistence Steenrod Modules". In: *Journal of Applied and Computational Topology* 6.4 (Dec. 2022), pp. 475–502. ISSN: 2367-1734. DOI: 10.1007/s41468-022-00093-7 (cit. on p. 36).

[Loi+24] David Loiseaux, Luis Scoccola, Mathieu Carrière, Magnus Bakke Botnan, and Steve Oudot. *Stable Vectorization of Multiparameter Persistent Homology Using Signed Barcodes as Measures*. Feb. 2024. DOI: 10.48550/arXiv.2306.03801. arXiv: 2306.03801 [cs, math, stat] (cit. on p. 30).

[LPA20] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. "Bad Global Minima Exist and SGD Can Reach Them". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 8543–8552 (cit. on pp. 85, 86).

[LS24] David Loiseaux and Hannah Schreiber. "Multipers: Multiparameter Persistence for Machine Learning". In: *Journal of Open Source Software* 9.103 (Nov. 2024), p. 6773. ISSN: 2475-9066. DOI: 10.21105/joss.06773 (cit. on pp. 74, 105).

[LW15] Michael Lesnick and Matthew Wright. *Interactive Visualization of 2-D Persistence Modules*. Dec. 2015. arXiv: 1512.00180 [cs, math] (cit. on p. 30).

[LW22] Michael Lesnick and Matthew Wright. "Computing Minimal Presentations and Bigraded Betti Numbers of 2-Parameter Persistent Homology". In: *SIAM Journal on Applied Algebra and Geometry* 6.2 (June 2022), pp. 267–298. DOI: 10.1137/20M1388425 (cit. on p. 30).

[Man67] Benoit Mandelbrot. "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension". In: *Science* 156.3775 (May 1967), pp. 636–638. DOI: 10.1126/science.156.3775.636 (cit. on p. 77).

[MB24] Jonathan M. Mousley and Paul Bendich. *Cross-Matched Interval Prevalence of High Dimensional Point Clouds*. Nov. 2024. DOI: 10.48550/arXiv.2411.09797. arXiv: 2411.09797 [math] (cit. on p. 14).

[MEG04] Benoit B Mandelbrot, Carl JG Evertsz, and Martin C Gutzwiller. *Fractals and Chaos: The Mandelbrot Set and Beyond*. Vol. 3. Springer, 2004 (cit. on p. 77).

[MMH11] Yuriy Mileyko, Sayan Mukherjee, and John Harer. "Probability Measures on the Space of Persistence Diagrams". In: *Inverse Problems* 27.12 (Nov. 2011), p. 124007. ISSN: 0266-5611. DOI: 10.1088/0266-5611/27/12/124007 (cit. on p. 27).

[MS12] Robert MacPherson and Benjamin Schweinhart. "Measuring Shape with Topology". In: *Journal of Mathematical Physics* 53.7 (July 2012), p. 073516. ISSN: 0022-2488. DOI: 10.1063/1.4737391 (cit. on p. 78).

[MSZ24] Facundo Mémoli, Anastasios Stefanou, and Ling Zhou. "Persistent Cup Product Structures and Related Invariants". In: *Journal of Applied and Computational Topology* 8.1 (Mar. 2024), pp. 93–148. ISSN: 2367-1734. DOI: 10.1007/s41468-023-00138-5 (cit. on p. 36).

[Mun25] Elizabeth Munch. "An Invitation to the Euler Characteristic Transform". In: *The American Mathematical Monthly* 132.1 (Jan. 2025), pp. 15–25. ISSN: 0002-9890. DOI: 10.1080/00029890.2024.2409616 (cit. on p. 106).

[MW11] Shawn Martin and Jean-Paul Watson. "Non-Manifold Surface Reconstruction from High-Dimensional Point Cloud Data". In: *Computational Geometry* 44.8 (Oct. 2011), pp. 427–441. ISSN: 0925-7721. DOI: 10.1016/j.comgeo.2011.05.002 (cit. on p. 10).

[MWW22] Facundo Mémoli, Zhengchao Wan, and Yusu Wang. "Persistent Laplacians: Properties, Algorithms and Implications". In: *SIAM Journal on Mathematics of Data Science* 4.2 (June 2022), pp. 858–884. DOI: 10.1137/21M1435471 (cit. on p. 69).

[Nak+21] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. "Deep Double Descent: Where Bigger Models and More Data Hurt*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (Dec. 2021), p. 124003. ISSN: 1742-5468. DOI: 10.1088/1742-5468/ac3a74 (cit. on pp. 86, 87).

[NSW08] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. "Finding the Homology of Submanifolds with High Confidence from Random Samples". In: *Discrete & Computational Geometry* 39.1 (Mar. 2008), pp. 419–441. ISSN: 1432-0444. DOI: 10.1007/s00454-008-9053-2 (cit. on p. 16).

[Oll07] Yann Ollivier. "Ricci Curvature of Metric Spaces". In: *Comptes Rendus. Mathématique* 345.11 (Nov. 2007), pp. 643–646. ISSN: 1778-3569. DOI: 10.1016/j.crma.2007.10.041 (cit. on p. 106).

[Ott+17] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. "A Roadmap for the Computation of Persistent Homology". In: *EPJ Data Science* 6.1 (Dec. 2017), p. 17. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-017-0109-5 (cit. on p. 16).

[Oud15] Steve Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*. Vol. 209. Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society, Dec. 2015. ISBN: 978-1-4704-2545-6 978-1-4704-2795-5. DOI: 10.1090/surv/209 (cit. on pp. 15, 21, 24).

[Pap+24] Theodore Papamarkou, Tolga Birdal, Michael Bronstein, et al. "Position: Topological Deep Learning Is the New Frontier for Relational Learning". In: *Proceedings of machine learning research* 235 (July 2024), pp. 39529–39555. ISSN: 2640-3498 (cit. on p. 12).

[Pat18] Amit Patel. "Generalized Persistence Diagrams". In: *Journal of Applied and Computational Topology* 1.3 (June 2018), pp. 397–419. ISSN: 2367-1734. DOI: 10.1007/s41468-018-0012-6 (cit. on p. 30).

[PHR16] Florian T. Pokorny, Majd Hawasly, and Subramanian Ramamoorthy. "Topological Trajectory Classification with Filtrations of Simplicial Complexes and Persistent Homology". In: *The International Journal of Robotics Research* 35.1-3 (Jan. 2016), pp. 204–223. ISSN: 0278-3649, 1741-3176. DOI: 10.1177/0278364915586713 (cit. on p. 11).

[Pro25] The GUDHI Project. *GUDHI User and Reference Manual*. 3.11.0. GUDHI Editorial Board, 2025 (cit. on p. 105).

[PS00] Michael Prähofer and Herbert Spohn. "Statistical Self-Similarity of One-Dimensional Growth Processes". In: *Physica A: Statistical Mechanics and its Applications* 279.1 (May 2000), pp. 342–352. ISSN: 0378-4371. DOI: 10.1016/S0378-4371(99)00517-8 (cit. on p. 77).

[PST23] Jose A Perea, Luis Scoccola, and Christopher J Tralie. "DREiMac: Dimensionality Reduction with Eilenberg-MacLane Coordinates". In: *Journal of Open Source Software* 8.91 (2023), p. 5791 (cit. on p. 36).

[PT12] L. Pietronero and E. Tosatti. *Fractals in Physics*. Elsevier, Dec. 2012. ISBN: 978-0-444-59841-7 (cit. on p. 77).

[Qui06] Daniel Quillen. "Higher Algebraic K-Theory: I". In: *Higher K-Theories: Proceedings of the Conference Held at the Seattle Research Center of the Battelle Memorial Institute, from August 28 to September 8, 1972*. Springer, 2006, pp. 85–147 (cit. on p. 59).

[RB23] Yohai Reani and Omer Bobrowski. "Cycle Registration in Persistent Homology With Applications in Topological Bootstrap". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (May 2023), pp. 5579–5593. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2022.3217443 (cit. on pp. 38, 43–45).

[Rei+15] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. "A Stable Multi-Scale Kernel for Topological Machine Learning". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 4741–4748. DOI: 10.1109/CVPR.2015.7299106 (cit. on pp. 27, 29).

[Rie17] Emily Riehl. *Category Theory in Context*. Courier Dover Publications, Mar. 2017. ISBN: 978-0-486-82080-4 (cit. on p. 52).

[RR23] Ernst Röell and Bastian Rieck. "Differentiable Euler Characteristic Transforms for Shape Classification". In: *The Twelfth International Conference on Learning Representations*. Oct. 2023 (cit. on p. 106).

[Ruc+16] Matteo Rucco, Filippo Castiglione, Emanuela Merelli, and Marco Pettini. "Characterisation of the Idiotypic Immune Network Through Persistent Entropy". In: *Proceedings of ECCS 2014*. Ed. by Stefano Battiston, Francesco De Pellegrini, Guido Caldarelli, and Emanuela Merelli. Cham: Springer International Publishing, 2016, pp. 117–128. ISBN: 978-3-319-29228-1. DOI: 10.1007/978-3-319-29228-1_11 (cit. on p. 29).

[Sat+98] Yoshinobu Sato, Shin Nakajima, Nobuyuki Shiraga, Hideki Atsumi, Shigeyuki Yoshida, Thomas Koller, Guido Gerig, and Ron Kikinis. "Three-Dimensional Multi-Scale Line Filter for Segmentation and Visualization of Curvilinear Structures in Medical Images". In: *Medical Image Analysis* 2.2 (June 1998), pp. 143–168. ISSN: 1361-8415. DOI: 10.1016/S1361-8415(98)80009-1 (cit. on p. 49).

[Sch+08] Paul J. Scherz, Jan Huisken, Pankaj Sahai-Hernandez, and Didier Y. R. Stainier. "High-Speed Imaging of Developing Heart Valves Reveals Interplay of Morphogenesis and Function". In: *Development* 135.6 (Mar. 2008), pp. 1179–1187. ISSN: 0950-1991. DOI: 10.1242/dev.010694 (cit. on p. 48).

[Sch20] Benjamin Schweinhart. "Fractal Dimension and the Persistent Homology of Random Geometric Complexes". In: *Advances in Mathematics* 372 (Oct. 2020), p. 107291. ISSN: 0001-8708. DOI: 10.1016/j.aim.2020.107291 (cit. on pp. 77, 78).

[Sch21] Benjamin Schweinhart. "Persistent Homology and the Upper Box Dimension". In: *Discrete & Computational Geometry* 65.2 (Mar. 2021), pp. 331–364. ISSN: 1432-0444. DOI: 10.1007/s00454-019-00145-3 (cit. on pp. 78, 79).

[Sco+23] Luis Scoccola, Hitesh Gakhar, Johnathan Bush, Nikolas Schonsheck, Tatum Rask, Ling Zhou, and Jose A. Perea. "Toroidal Coordinates: Decorrelating Circular Coordinates with Lattice Reduction". In: *LIPIcs, Volume 258, SoCG 2023* 258 (2023). Ed. by Erin W. Chambers and Joachim Gudmundsson, 57:1–57:20. ISSN: 1868-8969. DOI: 10.4230/LIPICS.SOCG.2023.57 (cit. on p. 36).

[SMV11] Vin de Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. "Dualities in Persistent (Co)Homology". In: *Inverse Problems* 27.12 (Nov. 2011), p. 124003. ISSN: 0266-5611. DOI: 10.1088/0266-5611/27/12/124003 (cit. on pp. 26, 27, 36, 38, 104).

[SS24] Nikolas C. Schonsheck and Stefan C. Schonsheck. "Spherical Coordinates from Persistent Cohomology". In: *Journal of Applied and Computational Topology* 8.1 (Mar. 2024), pp. 149–173. ISSN: 2367-1734. DOI: 10.1007/s41468-023-00141-w (cit. on p. 36).

[ST21] Primoz Skraba and Katharine Turner. *Wasserstein Stability for Persistence Diagrams*. Mar. 2021. DOI: 10.48550/arXiv.2006.16824. arXiv: 2006.16824 [math] (cit. on pp. 32, 67).

[Tan+24] Charlie Tan, Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Nov. 2024 (cit. on pp. 13, 76, 81, 82, 84, 85, 88, 89, 104).

[Tau+21] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal M. Medina-Mardones, Alberto Dassatti, and Kathryn Hess. "Giotto-Tda: : A Topological Data Analysis Toolkit for Machine Learning and Data Exploration". In: *Journal of Machine Learning Research* 22.39 (2021), pp. 1–6. ISSN: 1533-7928 (cit. on p. 82).

[Tou+23] Hugo Touvron, Louis Martin, Kevin Stone, et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. July 2023. DOI: 10.48550/arXiv.2307.09288. arXiv: 2307.09288 [cs] (cit. on p. 91).

[Tur+14] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. "Fréchet Means for Distributions of Persistence Diagrams". In: *Discrete & Computational Geometry* 52.1 (July 2014), pp. 44–70. ISSN: 1432-0444. DOI: 10.1007/s00454-014-9604-7 (cit. on p. 27).

[VAB13] Ramanarayan Vasudevan, Aaron Ames, and Ruzena Bajcsy. "Persistent Homology for Automatic Determination of Human-Data Based Cost of Bipedal Walking". In: *Nonlinear Analysis: Hybrid Systems* 7.1 (Feb. 2013), pp. 101–115. ISSN: 1751570X. DOI: 10.1016/j.nahs.2012.07.006 (cit. on p. 11).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 90).

[Ver+93] A. Verri, C. Uras, P. Frosini, and M. Ferri. "On the Use of Size Functions for Shape Analysis". In: *Biological Cybernetics* 70.2 (Dec. 1993), pp. 99–107. ISSN: 1432-0770. DOI: 10.1007/BF00200823 (cit. on p. 28).

[Vip+21] Oliver Vipond, Joshua A. Bull, Philip S. Macklin, Ulrike Tillmann, Christopher W. Pugh, Helen M. Byrne, and Heather A. Harrington. "Multiparameter Persistent Homology Landscapes Identify Immune Cell Spatial Patterns in Tumors". In: *Proceedings of the National Academy of Sciences* 118.41 (Oct. 2021), e2102166118. DOI: 10.1073/pnas.2102166118 (cit. on p. 105).

[Vip20] Oliver Vipond. "Multiparameter Persistence Landscapes". In: *Journal of Machine Learning Research* 21.61 (2020), pp. 1–38. ISSN: 1533-7928 (cit. on pp. 29, 69, 71).

[Wan+23] Qiquan Wang, Inés García-Redondo, Pierre Faugère, Anthea Monod, and Gregory Henselman-Petrusek. *Computable Stability for Persistence Rank Function Machine Learning*. July 2023. DOI: 10.48550/arXiv.2307.02904. arXiv: 2307.02904 [math] (cit. on pp. 67–69).

[Wan+24] Qiquan Wang, Inés García-Redondo, Pierre Faugère, Gregory Henselman-Petrusek, and Anthea Monod. "Stability for Inference with Persistent Homology Rank Functions". In: *Computer Graphics Forum* 43.5 (2024), e15142. ISSN: 1467-8659. DOI: 10.1111/cgf.15142 (cit. on pp. 13, 14, 63–67, 103).

[Wei+24] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations". In: *The Thirteenth International Conference on Learning Representations*. Oct. 2024 (cit. on pp. 92, 93).

[XLM18] Kelin Xia, Zhiming Li, and Lin Mu. "Multiscale Persistent Functions for Biomolecular Structure Characterization". In: *Bulletin of Mathematical Biology* 80.1 (Jan. 2018), pp. 1–31. ISSN: 1522-9602. DOI: 10.1007/s11538-017-0362-6 (cit. on p. 11).

[YGG23] Hee Rhang Yoon, Robert Ghrist, and Chad Giusti. "Persistent Extensions and Analogous Bars: Data-Induced Relations between Persistence Barcodes". In: *Journal of Applied and Computational Topology* 7.3 (Sept. 2023), pp. 571–617. ISSN: 2367-1734. DOI: 10.1007/s41468-023-00115-y (cit. on p. 37).

[ZC05] Afra Zomorodian and Gunnar Carlsson. "Computing Persistent Homology". In: *Discrete & Computational Geometry* 33.2 (Feb. 2005), pp. 249–274. ISSN: 1432-0444. DOI: 10.1007/s00454-004-1146-y (cit. on p. 25).

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding Deep Learning Requires Rethinking Generalization". In: *International Conference on Learning Representations*. Feb. 2017 (cit. on p. 80).