



Persistent Homology in Topological Data Analysis: Theory and Applications

E Julie

Associate Professor, Department Of Zoology, Providence Women's College (Autonomous) Calicut, Kerala, India

Article Information

Received: 6th February 2026

Received in revised form: 9th March 2026

Accepted: 10th April 2026

Available online: 14th May 2026

Volume: 2

Issue: 2

DOI: <https://doi.org/10.5281/zenodo.21916424>

Abstract

Persistent homology is a computational methodology for extracting topological features of data across a range of scales. Originating from the work of Edelsbrunner, Letscher and Zomorodian in 2002, it has developed into a mature branch of topological data analysis with applications in neuroscience, materials science, genomics, and machine learning. This paper reviews the mathematical foundations of persistent homology, describes its computational realisation via simplicial filtrations and persistence diagrams, and surveys representative applications. Stability theorems and the bottleneck distance provide a principled metric on persistence diagrams, enabling statistical inference and integration with machine-learning pipelines. Current research directions include multi-parameter persistence, differentiable persistence modules, and scalable algorithms for large data.

Keywords: Persistent Homology, Topological Data Analysis, Simplicial Complex, Persistence Diagram; Filtration, Bottleneck Distance.

1. INTRODUCTION

Topological data analysis (TDA) is an emerging methodology at the intersection of algebraic topology, statistics, and computer science. Its central objective is to recover the shape of data connected components, loops, voids, and higher-dimensional analogues from finite point-cloud samples. Classical Betti numbers identify these features for a fixed topological space, but real datasets are subject to noise, sampling artefacts, and ambiguity in the choice of scale. A single fixed scale cannot distinguish meaningful topology (signal) from small-scale fluctuations (noise), and no single scale is generally 'correct' for a given dataset.

Persistent homology resolves this ambiguity by tracking topological features across a filtration of spaces indexed by a scale parameter, distinguishing long-lived features (signal) from short-lived features (noise).^{1,4} The output is a compact, stable summary a persistence diagram or barcode that encodes the birth and death of topological features across the scale range. This multi-scale representation has proved to be a powerful feature extractor for complex data and has been successfully applied in domains including neuroscience, materials science, genomics, image analysis, and machine learning.

Over two decades TDA has developed a coherent theoretical framework interleaving distances, stability theorems, and decomposition results for persistence modules and practical software tools that have driven applications in diverse scientific domains.^{2,5} The field has matured into a recognised branch of applied mathematics with dedicated conferences (SoCG's Applied Topology track, ATMCS), journals (Journal of Applied and Computational Topology, Foundations of Computational Mathematics), and open-source software infrastructure (GUDHI, Ripser, giotto-tda, scikit-tda). The present paper reviews the mathematical foundations,

surveys computational techniques, and illustrates applications across scientific domains. Section 2 treats simplicial complexes and filtrations; Section 3 presents persistence modules and diagrams; Section 4 addresses stability; Section 5 covers computational aspects; Section 6 surveys applications; and Section 7 discusses current research directions.

2. SIMPLICIAL COMPLEXES AND FILTRATIONS

A simplicial complex K on a vertex set V is a collection of simplices (a k -simplex being an ordered $(k+1)$ -tuple of vertices) closed under the face relation: if $\sigma \in K$ and τ is a face of σ , then $\tau \in K$. The topological realisation $|K|$ is obtained by gluing geometric simplices along their common faces; its singular (or simplicial) homology groups $H_k(K; F)$ with coefficients in a field F encode the k -dimensional 'holes' of the complex, with dimensions:

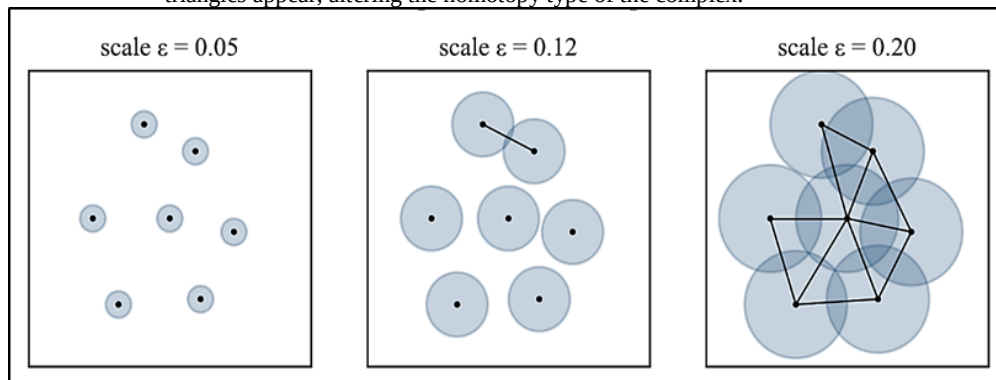
$$\beta_k = \dim H_k(K; F) \quad (1)$$

the Betti numbers. β_0 counts connected components, β_1 counts independent loops not bounding a surface, β_2 counts two-dimensional voids, and so on.

Geometric constructions from point clouds $\{x_1, \dots, x_n\}$ in a metric space include the Čech complex $\check{C}(\varepsilon)$, consisting of simplices $\{x_{i_0}, \dots, x_{i_k}\}$ whose closed balls of radius ε have nonempty intersection; the Vietoris–Rips complex $VR(\varepsilon)$, consisting of simplices whose pairwise distances are at most 2ε ; the Delaunay complex, the dual of the Voronoi diagram; and the alpha complex, restricting the Čech filtration to Delaunay simplices. The Nerve Theorem links the Čech complex to the homotopy type of the union of balls: if the balls and their pairwise intersections are contractible, the Čech complex is homotopy-equivalent to their union.⁶ In practice, the Vietoris–Rips complex is preferred for computational efficiency despite being geometrically larger than the Čech complex, because its construction requires only pairwise distances rather than higher-order intersections. For point clouds in high-dimensional metric spaces, sparse approximations (Sparse Rips, alpha sparse) reduce complex size while preserving persistence up to bounded perturbations.

A filtration is a nested sequence $K_0 \subseteq K_1 \subseteq \dots \subseteq K_n$ of simplicial complexes, typically indexed by an increasing scale ε (Fig. 1). As ε grows, new topological features (connected components, loops, voids) are born and others die as components merge or loops are filled. More generally, a filtration may be indexed by any totally ordered set, and different physical or data-derived quantities sublevel sets of functions, time in dynamical systems, or biological scales produce filtrations beyond the distance-based construction. Sublevel-set filtrations of real-valued functions on a fixed domain connect persistent homology to Morse theory; the persistence pairing of critical points generalises the Morse–Smale complex and underpins applications in scalar-field visualisation.

Fig 1: Stages of a Čech/Vietoris–Rips filtration over a planar point cloud. As ε grows, new edges and triangles appear, altering the homotopy type of the complex.^{6,4}



3. PERSISTENCE MODULES AND DIAGRAMS

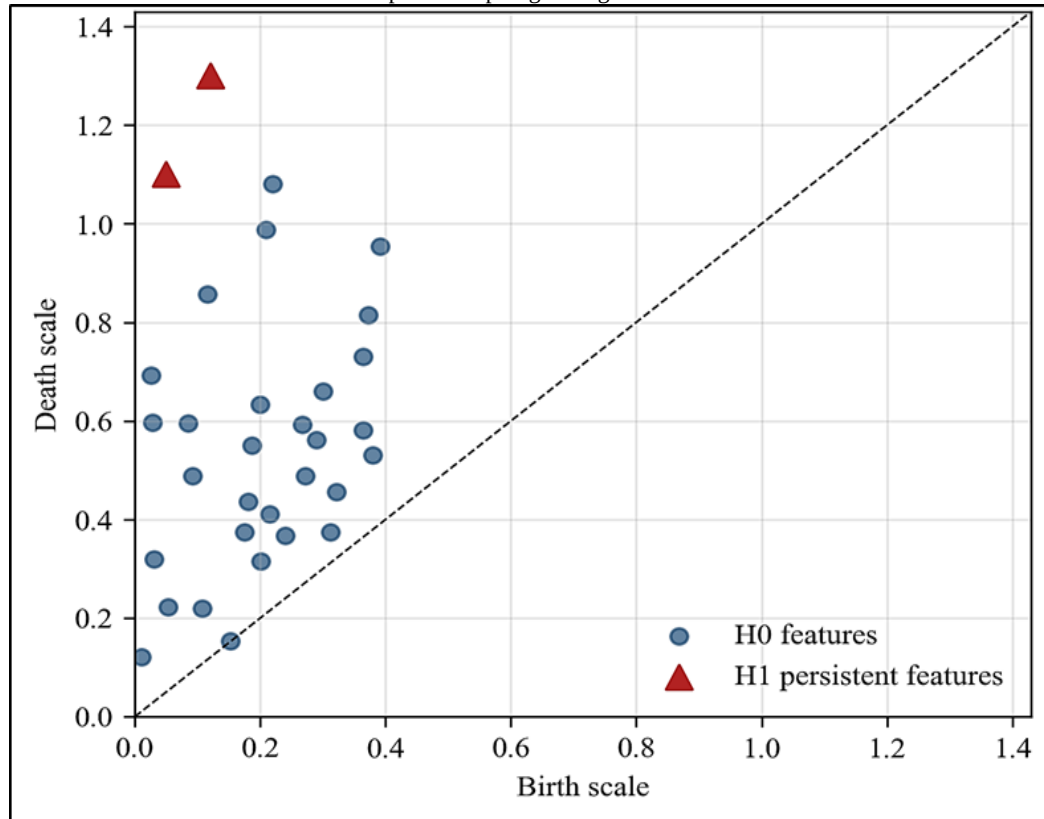
Applying simplicial homology H_k with coefficients in a field F to each stage of a filtration yields a persistence module: a family of vector spaces $\{H_k(K_i)\}$ indexed by i , together with induced linear maps $H_k(K_i) \rightarrow H_k(K_j)$ for $i \leq j$, arising from the inclusion $K_i \subset K_j$. Viewed categorically, a persistence module is a functor from the poset $(\mathbb{N}, \leq)$ (or more generally $(\mathbb{R}, \leq)$) to the category of vector spaces over F . This functorial perspective connects persistent homology to deep structural results in representation theory and category theory.

A foundational decomposition theorem, established by Crawley-Boevey (2015) for persistence modules over the reals and earlier by others for finite-type cases, shows that finite-type persistence modules decompose uniquely (up to isomorphism) into interval modules. An interval module $I[b, d)$ is the persistence module with F at each index in $[b, d)$ and zero elsewhere, with identity maps on the non-zero portion. This decomposition gives a complete invariant in the form of a persistence diagram a multiset of points $(b, d) \in \mathbb{R}^2$ with $b < d$, representing

features born at scale b and dying at scale d .^{1,4} The persistence barcode is the equivalent horizontal-bar representation, with each bar $[b,d]$ corresponding to one feature.

The persistence diagram admits several equivalent but practically distinct representations. Beyond the diagram and barcode, the persistence landscape is a sequence of real-valued functions $\lambda_k: \mathbb{R} \rightarrow \mathbb{R}$ derived from stacking the interval functions in order of height, producing a stable vectorised representation amenable to L^2 statistics.³ Persistence images further convert the diagram into a fixed-resolution two-dimensional density, enabling direct use as features in convolutional networks.¹⁰ Fig. 2 shows a representative persistence diagram with long-persistence H_1 features indicating loops.

Fig 2: Persistence diagram: the diagonal separates features by persistence ($d - b$). Points far from the diagonal correspond to topological signal.^{1,4}



4. STABILITY AND DISTANCES

The bottleneck distance $d_B(D, D')$ between persistence diagrams D and D' is the infimum over all bijections $\gamma: D \cup \Delta \rightarrow D' \cup \Delta$ (where Δ is the diagonal, representing features of zero persistence) of the supremum over points:

$$P(\|p - \gamma(p)\|_\infty) \quad (2)$$

This metric provides the natural setting for persistence-module stability. The foundational Stability Theorem of Cohen-Steiner, Edelsbrunner and Harer asserts that if two filtrations are generated by functions f, g on a common domain, then :

$$d_B(D(f), D(g)) \leq \|f - g\|_\infty \quad (3)$$

small perturbations of input data produce only small perturbations of the persistence diagram.² This stability theorem underpins the statistical viability of persistent homology on noisy data.

Wasserstein distances provide L^p analogues more suited to statistical analysis: $d_{w_p}(D, D^1)$ is the infimum over bijections:

$$\gamma = \left(\sum_p \|p - \gamma(p)\|_\infty^p \right)^{\frac{1}{p}} \quad (4)$$

Wasserstein distances are absolutely continuous with respect to noise, enabling central-limit-theorem-like statistical analysis. The interleaving distance on persistence modules provides a more general categorical framework in which stability becomes a statement about functors, extending to multi-parameter persistence and

zigzag persistence. The algebraic-stability theorem of Chazal and colleagues sharpens the Cohen-Steiner bound and establishes that interleaving distance equals bottleneck distance for one-parameter persistence modules.

5. COMPUTATIONAL ASPECTS

Computing persistence diagrams reduces to a boundary-matrix reduction. For a filtration with m total simplices, one constructs a boundary matrix D whose columns encode boundary homomorphisms; applying column operations modulo 2 (for $\mathbb{Z}/2$ coefficients) to reduce the matrix to a canonical form reveals persistence pairs through the positions of 'pivots'. The naive reduction algorithm is $O(m^3)$ worst-case but typically $O(m^{2.37})$ or better in practice.⁷

Substantial algorithmic improvements have reduced runtimes by orders of magnitude. The 'apparent pairs' optimisation identifies easy persistence pairs during construction, removing them from the reduction. Matrix-compression techniques use spectral sequences to batch-process independent blocks; the clearing optimisation reuses earlier reduction results for subsequent dimensions; discrete Morse theory pre-processing can dramatically shrink the effective complex size by collapsing acyclic portions. Chunk-based algorithms parallelise reduction across filtration segments. Open-source libraries GUDHI (comprehensive TDA toolkit), Ripser (specialised for Vietoris–Rips), Dionysus (general persistence), PHAT (high-performance persistent-homology algorithm toolbox), and Eirene (Julia-based) now support computation on complexes with millions of simplices on commodity hardware.

For very large point clouds or streaming data, approximation techniques become essential. Sparse Rips filtrations achieve controlled approximation with fewer simplices by restricting to edges between spatially nearby points, preserving the persistence diagram to within bounded error. Lazy witness complexes use a small set of landmark points and approximate persistence via witnessing relationships with the full point cloud. Distance-to-measure (DTM) filtrations replace the Euclidean distance with a robust distance estimated from the empirical measure, substantially improving robustness to outliers.⁸ Recent advances in vineyards (time-varying persistence) and zigzag persistence (non-monotonic filtrations) extend the framework to dynamic and bidirectional settings.

6. APPLICATIONS

TDA applications span the sciences. In neuroscience, persistent homology characterises functional-brain-network topology derived from resting-state fMRI, diffusion MRI, and MEG/EEG recordings. Persistent Betti numbers computed from thresholded correlation matrices have been used to identify signatures of Alzheimer's disease, schizophrenia, epilepsy, and autism spectrum disorder. Topological summary features derived from brain connectomes achieve classification accuracies competitive with or superior to conventional graph-theoretic measures such as small-worldness and modularity.⁹ The cavities and voids of functional connectivity networks, captured by H_1 and H_2 persistence, encode aspects of brain organisation inaccessible to pairwise connectivity measures.

In materials science, TDA analyses porous-material structure, nanoparticle aggregation, dislocation dynamics, glass transition, and amorphous solid structure. The free-volume distribution characterised by persistent homology correlates with ion-transport properties in battery electrolytes; ring-size distributions of amorphous carbon derive quantitatively from H_1 barcodes; and Kondo-lattice phase transitions exhibit persistent-homology signatures in simulated electron densities. In structural biology, persistent homology describes protein-structure cavities, ligand-binding pockets, and macromolecular assembly.

In genomics, TDA has been used to detect structural variation and haplotype diversity through topological analysis of distance matrices of single-nucleotide polymorphisms. The algebraic-topological nature of recombination producing non-tree-like evolutionary structure makes persistent homology a natural tool for detecting horizontal gene transfer, recombination events, and reticulate evolution. Tumour evolutionary trajectories have similarly been analysed with TDA.

In image analysis, cubical persistent homology on greyscale images provides a topological complement to conventional filtering, detecting filamentary structure (astronomical cosmic web), peaks and basins (geographical terrain), and characteristic holes in digital images. TDA has also been applied to financial time series, where persistent Betti numbers computed from sliding-window embeddings have been used to detect market-regime transitions and financial crises. In machine learning, persistence images and persistence landscapes provide vectorised representations compatible with kernel methods and neural networks.¹⁰ Topological regularisation has been used to enforce connectivity in segmentation outputs and to control the topological genus of generative-model outputs. Table 1 summarises representative applications.

Table 1. Representative application domains for persistent homology.^{9,10}

Domain	Data object	Topological feature of interest	Illustrative outcome
Neuroscience	fMRI connectivity graphs	Cycles, cavities	Alzheimer's-stage stratification
Materials science	Atomic coordinates	Voids in amorphous solids	Free-volume characterization
Genomics	SNP/haplotype graphs	Non-tree recombination loops	Detection of recombination events
Image analysis	Cubical complexes on greyscale	Peaks and basins	Astronomical filament detection
Machine learning	Data-cloud features	Global shape descriptors	Topological regularisation of neural nets

7. CURRENT DIRECTIONS

Several active research directions extend persistent homology beyond the classical one-parameter framework. Multi-parameter persistence considers filtrations indexed by two or more parameters for example, scale together with density, or time together with noise level. Unlike one-parameter persistence, multi-parameter persistence modules do not admit a complete discrete invariant analogous to the persistence diagram; however, substantial progress has been made on approximate invariants, rank invariants, and fibered-barcode descriptions. The theory connects to commutative algebra (minimal free resolutions of bigraded modules) and is actively being developed into computational tools.

Differentiable persistent homology integrates persistence computation with gradient-based machine-learning pipelines.¹¹ Defining a meaningful differential of the persistence diagram with respect to input data allows persistence-based loss functions to be optimised by stochastic gradient descent. This approach has been applied to regularisation of neural networks (topological autoencoders, topology-preserving dimensionality reduction), topological priors for generative models, and in physics-informed learning where topological invariants must be preserved during training. Persistence-based loss functions have been used, for example, to enforce specified Betti numbers in generated meshes or to regularise segmentation networks to produce connected regions.

Zigzag persistence generalises to filtrations with maps in both directions useful for dynamic data where features may appear and disappear in non-monotonic fashion. Applications include analysis of dynamic networks, time-varying sensor data, and evolving biological systems. Finally, persistent-homology-based loss functions have been integrated into generative models (topological generative adversarial networks), physics-informed learning, and protein-structure prediction. Persistence landscapes, persistence images, and persistent Betti numbers provide vectorised representations that embed persistence diagrams into standard machine-learning pipelines with kernel methods, random forests, and deep networks.

8. CONCLUSION

Persistent homology has transitioned from a niche mathematical methodology to a mainstream data-analysis tool within a generation. Its mathematical coherence decomposition theorems, stability results, algebraic foundations combined with scalable computation and an active open-source ecosystem has established persistent homology as a standard component of the modern data-analysis toolkit. Application successes in neuroscience, materials science, genomics, and machine learning have validated the approach empirically, while the theoretical framework continues to deepen through categorical, statistical, and differentiable extensions. The field is now poised for broader integration with deep learning, high-dimensional statistics, and scientific computing, and for continued expansion into application domains where shape and connectivity information is inadequately captured by point-wise or pair-wise descriptors alone.

REFERENCES

1. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete & Computational Geometry*. 2002;28(4):511–533.
2. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. *Discrete & Computational Geometry*. 2007;37(1):103–120.
3. Bubenik P. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*. 2015;16(3):77–102.
4. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete & Computational Geometry*. 2005;33(2):249–274.
5. Carlsson G. Topology and data. *Bulletin of the American Mathematical Society*. 2009;46(2):255–308.
6. Hatcher A. *Algebraic topology*. Cambridge (UK): Cambridge University Press; 2002.
7. Bauer U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*. 2021;5(3):391–423.

8. Chazal F, de Silva V, Oudot S. Persistence stability for geometric complexes. *Geometriae Dedicata*. 2014;173(1):193–214.
9. Sizemore AE, Giusti C, Kahn A, Vettel JM, Betzel RF, Bassett DS. Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*. 2018;44(1):115–145.
10. Adams H, Emerson T, Kirby M, et al. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*. 2017;18(8):1–35.
11. Carrière M, Chazal F, Ike Y, Lacombe T, Royer M, Umeda Y. PersLay: a neural network layer for persistence diagrams. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS); 2020. p. 2786–2796.