Topology Preserving Dimensionality Reduction

Mathieu Carrière and Julien Tierny

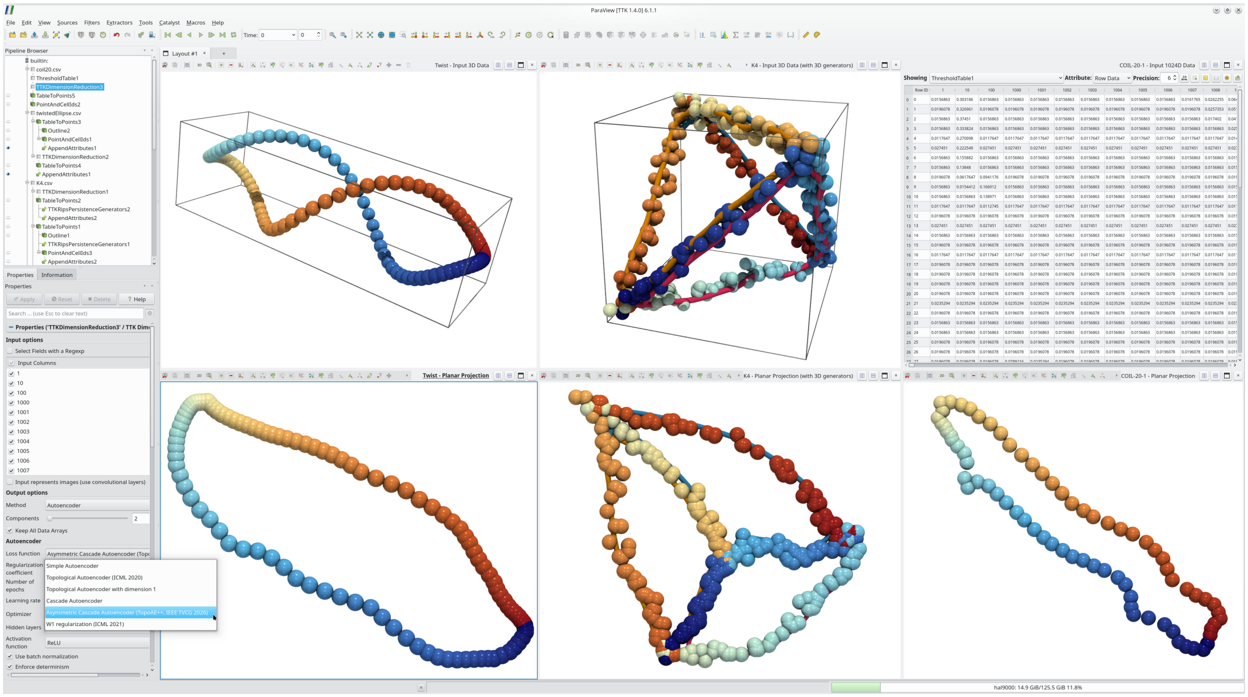

Fig. 1. Synthetic and real-life examples of dimensionality reduction with *TopoAE++* [4], an approach optimizing the preservation of clusters and loops in the data (formally, the degree 0 and 1 persistent homologies of the Rips filtration). While several topology-aware dimensionality reduction techniques have been recently proposed, they still suffer from a lack of *(i)* scalability, *(ii)* stability and *(iii)* robustness. We suggest focusing on one of these three issues in this internship.

Keywords—Topological data analysis, dimensionality reduction, visualization, data science.

1 Context

Dimensionality reduction is a key tool in data analysis and visualization. This procedure typically takes as input a dataset represented as a set $P = \{p_1, p_2, \dots, p_N\}$ of N points living in a Euclidean space of dimension d. The goal of dimensionality reduction is to compute a map ϕ to a lower dimensional Euclidean space (e.g., the plane) $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$, with $d' \ll d$, while preserving some key characteristics of the data. This process is motivated by the need to compress the data (to facilitate its processing) or by the need for visual inspection.

Traditionally, dimensionality reduction aims at preserving some relevant geometrical characteristics of the data. For example, multi-dimensional scaling [17] aims at preserving the distance matrix associated with P in $\mathbb{R}^d$. In other cases, criteria can be of a structural nature: clusters may be meaningful for the application and loops may describe intrinsic characteristics of the data (e.g., periodic phenomena). For instance, one may want to preserve clusters through dimensionality reduction, to give users confidence that the clusters they see in the reduction actually exist in the original data [13, 19]. In the special case of single-linkage hierarchical clustering [8], it has been shown that preserving the clustering hierarchy is equivalent to preserving the persistence diagram of degree 0 of the Rips filtration of P [2], a well documented and understood topological descriptor [7].

- *Mathieu Carrière, Centre Inria d'Université Côte d'Azur, Sophia-Antipolis, mathieu.carriere@inria.fr.*
- *Julien Tierny, CNRS, Sorbonne University, LIP6, Paris, julien.tierny@sorbonne-universite.fr.*

2 State of the Art

Only a few techniques have been documented for the preservation of topological characteristics through dimensionality reduction [4, 6, 14, 18]. *TopoMap* [6] is the first technique to guarantee the strict preservation of the persistence diagram PD of degree 0 (denoted PD_0 and which tracks connected components) in the sense of the Wasserstein distance, an established distance to compare diagrams [7].

Persistence optimization [3] (e.g., minimizing Wasserstein distances between the diagrams of the input and output spaces) is a promising direction for integrating topological constraints when dimensionality reduction is modeled as an optimization procedure (e.g., using an auto-encoder [10]). However, the map taking a point set to its persistence diagram is not injective: distinct point sets can admit the same persistence diagram. In practice, this means that even if the optimization reaches a zero-valued Wasserstein distance (i.e., the persistence diagrams are identical), the geometrical realization of the topological structures can drastically differ after the projection. For instance, a dataset exhibiting a large loop L_1 and a smaller loop L_2 can in principle see these loops mapped, respectively, to a small loop $L'_1 = \phi(L_1)$ and a large loop $L'_2 = \phi(L_2)$, yielding identical degree-1 persistence diagrams PD_1. Additionally, points belonging to the loop L_2 can be mapped to the loop L'_1 and reciprocally, still potentially yielding identical degree-1 diagrams. Finally, points along a single loop can be arbitrarily permuted (e.g., breaking arc-length parameterization) while still yielding the same diagrams.

For these reasons, several approaches introduced optimization losses which go beyond the simple Wasserstein distance, to additionally constrain the geometrical realization of topological structures. For instance, *TopoAE* [14] constrained the preservation of the diameter of the critical simplices of the Rips filtration, both in the input and output spaces. While we recently showed that a zero-valued *TopoAE* loss indeed guarantees identical diagrams PD_0, we provided simple counterexamples for higher degree homology [4]. This implies that a zero-valued *TopoAE* loss for the critical edges and triangles is not sufficient to strictly preserve the loops.

To address this issue, we recently introduced *TopoAE++* [4] which additionally constrains diameter preservation for the simplices of the *cascade* of critical simplices [20], to guarantee that, in case of a zero-valued loss, the input loops are accurately reconstructed in the output.

3 Open problems and directions

Despite these recent advances in topology-aware dimensionality reduction, many open problems remain:

1. **Scalability:** In contrast to established dimensionality reduction approaches [13], topological methods still face a bottleneck related to the computation of persistent homology itself, a process with cubic worst-case time complexity (in the number of simplices), which is known to be quickly limited in practice to small point clouds. This is a limitation for the pre-processing step (evaluating persistent homology in the input space $\mathbb{R}^d$) but mostly for the optimization of the reduction (which requires a computation at each iteration in the output space $\mathbb{R}^{d'}$). From that perspective, sparse, landmark based constructions [9] present an interesting research direction.

2. **Stability:** Modern auto-encoder based approaches rely on stochastic optimization, whose initial conditions play a crucial role in the optimization. For instance, while it has been shown to better preserve degree-1 persistent homology, *TopoAE++* [4] is also more sensitive in practice to the initialization of its optimization and it struggles to exit bad configurations, e.g., self-intersecting loops. The design of specific losses to disentangle these configurations is an interesting research direction. Similarly, the optimization of multi-parameter persistence homology [12, 15] is another exciting direction for designing algorithms which are both more accurate and stable with respect to the initial conditions.

3. **Robustness:** While the Rips filtration is the go-to standard for analyzing point clouds in high dimensions, it is known to be very sensitive to outliers [1] and subject to the curse of dimensionality [11]. To address this issue, alternative filtrations may be studied to improve robustness to outliers [1] or the effect of high-dimensional noise [5]. Another interesting research direction to consider would be the combination of these strategies from the perspective of multi-parameter persistence [12, 15].

4 Internship organization

This internship proposes to focus on one of the above sub-problems (to be selected after initial discussions of preferences). This internship is viewed as a preliminary stage for a Ph.D. thesis follow-up. It could be organized as follows:

1. Study of the state of the art [3, 4, 6, 7, 12, 14, 15, 18];

2. Invent, formalize and study an approach for topology preserving dimensionality reduction which addresses one of the above sub-problems;

3. Implement and thoroughly experiment with synthetic and real-life datasets including known topological structures (e.g., clusters or loops);

4. Document the approach in a paper (if time permits).

Experimental programs may be written in Python for prototyping, but an implementation in C++ will be considered in later stages, e.g., for an integration within the *"Topology ToolKit"* (TTK) [16].

The internship will be co-advised by Mathieu Carrière and Julien Tierny, who are both experts in topological methods, with complementary skill sets. The internship is primarily based in Sorbonne University (Paris), for a start date of early April 2027, with occasional visits to Centre Inria d'Université Côte d'Azur. A symmetric setup, primarily based in Centre Inria d'Université Côte d'Azur with occasional visits to Sorbonne University, can also be considered.

References

[1] H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda. DTM-based filtrations. In *Topological Data Analysis: The Abel Symposium 2018*, pp. 33–66. Springer, 2020. doi: 10.1007/978-3-030-43408-3_2

[2] G. Carlsson and F. Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 2010.

[3] M. Carrière, F. Chazal, M. Glisse, Y. Ike, H. Kannan, and Y. Umeda. Optimizing persistent homology based functions. In *ICML*, 2021.

[4] M. Clémot, J. Digne, and J. Tierny. Topological Autoencoders++: Fast and Accurate Cycle-Aware Dimensionality Reduction. *IEEE TVCG*, 2026. doi: 10.1109/TVCG.2025.3644671

[5] S. Damrich, P. Berens, and D. Kobak. Persistent homology for high-dimensional data based on spectral methods. In *NeurIPS*, 2024. doi: 10.48550/arXiv.2311.03087

[6] H. Doraiswamy, J. Tierny, P. J. S. Silva, L. G. Nonato, and C. Silva. TopoMap: A 0-dimensional Homology Preserving Projection of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2020. doi: 10.1109/TVCG.2020.3030441

[7] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.

[8] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1969. doi: 10.2307/2346439

[9] F. Graf, P. Pellizzoni, M. Uray, S. Huber, and R. Kwitt. The Flood complex: Large-scale persistent homology on millions of points. In *NeurIPS*, 2025. doi: 10.48550/arXiv.2509.22432

[10] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. doi: 10.1126/science.1127647

[11] Y. Hiraoka, Y. Imoto, S. Kanazawa, and E. Liu. Curse of dimensionality on persistence diagrams. *Foundations of Data Science*, 2026. doi: 10.3934/fods.2025008

[12] D. Loiseaux, M. Carrière, and A. J. Blumberg. A framework for fast and stable representations of multiparameter persistent homology decompositions. In *NeurIPS*, 2023. doi: 10.48550/arXiv.2306.11170

[13] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426

[14] M. Moor, M. Horn, B. Rieck, and K. M. Borgwardt. Topological autoencoders. In *ICML*, 2020. doi: 10.48550/arXiv.1906.00722

[15] L. Scoccola, S. Setlur, D. Loiseaux, M. Carrière, and S. Oudot. Differentiability and optimization of multiparameter persistent homology. In *ICML*, 2024. doi: 10.48550/arXiv.2406.07224

[16] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017. `https://topology-tool-kit.github.io/`. doi: 10.1109/TVCG.2017.2743938

[17] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 1952. doi: 10.1007/BF02288916

[18] I. Trofimov, D. Cherniavskii, E. Tulchinskii, N. Balabin, E. Burnaev, and S. Barannikov. Learning topology-preserving data representations. In *ICLR*, 2023. doi: 10.48550/arXiv.2302.00136

[19] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[20] A. Zomorodian. Computational topology. In *Algorithms and theory of computation handbook: special topics and techniques*. 2 ed., 2010.