# Topological Methods in Oncology: Finding the Shape of Cancer

Quy-Dzu Do

06/10/2025

**Abstract**

Topology, the mathematical study of shape and space, is emerging as a powerful tool for extracting robust, noise-tolerant insights from complex, high-dimensional data. This paper reviews key topological techniques—particularly Topological Data Analysis (TDA)—and their application in cancer classification from single cell spatial genomic expressions specifically from 10x Visium Spacial Data. Case studies involving persistent homology and Mapper functions allow for the exploration of features that may be easily found via visualizations that may be later translated into features within an algorithm.

## 1 Introduction

Oncology, the study and treatment of cancer, faces immense challenges in understanding the biological complexity and variability of tumors. Traditional statistical and machine learning methods often struggle with the high-dimensionality and noise inherent in biological data. Topology, the mathematical study of shape, offers a unique perspective by focusing on the intrinsic geometric and topological structure of data rather than relying solely on coordinate-based information.

Topological Data Analysis has emerged as a robust framework for extracting qualitative features from data. Through tools such as persistent homology and the Mapper algorithm, TDA identifies shape-related features like connected components, holes, and voids. These structures often correspond to meaningful biological phenomena, such as clusters of tumor subtypes or patterns of gene expression in spatial transcriptomics. This paper explores the intersection of topology and oncology. We begin by introducing the fundamental concepts of topology and TDA, then review their application in cancer biology and present detailed case studies. Finally, we discuss clinical implications, challenges, and future directions for this area of research.

## 2 Primer on Topological Concepts

### 2.1 Basic Topology

Basic topology is the study of the properties of spaces that are preserved under continuous deformations such as stretching, bending, and twisting, but not tearing or gluing. At its foundation, topology defines a *topological space* as a set equipped with a collection of open sets that satisfy certain axioms, generalizing the notion of distance and nearness from metric spaces. Key concepts include *continuity*, *compactness*, and *connectedness*, which formalize intuitive ideas about shapes and spatial relationships. Topologists are interested in properties like whether a space is connected (in one piece), whether it has holes (like a doughnut), and how it behaves under mappings. For example, a coffee mug and a doughnut are considered topologically equivalent because one can be continuously deformed into the other without cutting or gluing—both have one hole. This flexible viewpoint allows topology to uncover deep structural insights in mathematics, physics, and data analysis.

## 2.2 Persistent Homology

Persistent homology is a method from computational topology that captures the multiscale topological features of data. At its core, it analyzes how features such as connected components, loops, and voids emerge and disappear as one moves through a *filtration*—a nested sequence of simplicial complexes built from a point cloud or more general data structure. These features are tracked across scales using algebraic tools from homology theory, and their lifespans are recorded as *persistence intervals*. The key idea is that topological features which persist across a wide range of scales are more likely to reflect meaningful structure in the data, whereas short-lived features may be attributed to noise.

The output of persistent homology is often visualized using *persistence diagrams* or *barcodes*. In a persistence diagram, each topological feature is represented as a point with coordinates $(b, d)$, where $b$ is the birth time and $d$ is the death time of the feature within the filtration. In barcode plots, the same information is depicted as horizontal line segments that span the interval $[b, d]$. These topological summaries provide insight into the underlying shape of complex datasets and are widely used in applications such as sensor network coverage, shape recognition, neuroscience, and computational biology. In particular, persistent

## 2.3 Mapper Algorithm

The Mapper algorithm is a topological data analysis tool designed to extract and visualize the shape of high-dimensional data. Mapper constructs a simplicial complex that approximates the underlying topology of a dataset. The process begins by selecting a *lens function*—a real-valued function such as a projection, density estimator, or eigenvector—that maps the high-dimensional data to a lower-dimensional space. The range of this function is then covered by overlapping intervals, and the pre-images of these intervals are clustered, typically using algorithms like $k$-means or DBSCAN. Each cluster is represented as a node, and edges are drawn between nodes if their corresponding clusters share data points. The result is a graph or network that reveals the structure of the data, highlighting features such as loops, flares, and branches. Mapper has found applications in a wide range of scientific domains, including biology, medicine, and social science, where it provides interpretable summaries of complex and noisy datasets.

# 3 Topology and Cancer Biology

Cancer is not a uniform disease but a dynamic system marked by heterogeneity, and spatial variation. Tumors evolve, adapt, and interact with their microenvironment, forming complex landscapes that are difficult to model with traditional tools. Topology provides a natural framework for exploring these complexities. For instance, the spatial organization of cells in a tumor can be described using topological properties. Similarly, gene expression patterns can potentially form topologically distinct clusters corresponding to tumor subtypes or microenvironments.

Topology offers powerful tools for analyzing complex biological data and has emerging applications in cancer classification. Methods such as persistent homology and the Mapper algorithm can be used to study high-dimensional data arising from gene expression. By capturing the shape and connectivity of these datasets, topological techniques can reveal subtle patterns and structural differences between healthy and cancerous tissues, as well as potentially different cancer subtypes. This enables the identification of biomarkers, tumor heterogeneity, and potential therapeutic targets, contributing to more accurate and interpretable cancer classification models.

# 4 Implementation and Execution

For all parts of this study, Python 3.12.6 was used and all source code is provided at `https://github.com/KrazyKats/TDA_cancer`

## 4.1 Persistent Homology in Gene Expression

To capture the persistent homology features of the datasets, we looked towards a method proposed by Bukkuri, Anuraag et al in which we create a percistance diagram and convert it to a vector[1]. We first filter for the top 100 gene expressions with the most variability. This allows for a sizable dataset with the most interesting portions of the data without overwhelming a system with to many data points. These diagrams record points $(b, d)$, where $b$ is the birth time and $d$ is the death time of a topological feature. For this specific framework, we only looked at the $H_1$ (or $B_1$) data since we had seen that the high variability of the cancerous tissue would leave "holes" within the samples which would still be mostly normal tissues. By rotating the diagram by $-45°$, the points are now aligned the horizontal axis. Thus the horizontal axis represents the birth time, while the vertical axis represents persistence = death - birth. We can then use various distribution methods to create a discrete heat map using bins to make the data more manageable. This heatmap can then be sliced into an array by the pixel/bin to pass on as a feature to further analysis as a feature in a machine learning pipeline or some other purpose. This process is shown graphically in Figure 1.

## 4.2 Mapper in Tumor Classification

We now look to Mapper tools to form graphs to help classify the tissues using the KeplerMapper library.We begin with the same data used in the persistent data, that is a

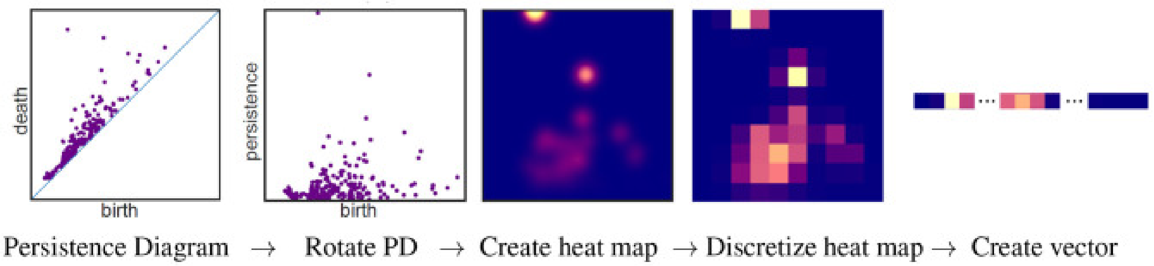

death
birth
persistence
birth
Persistence Diagram → Rotate PD → Create heat map → Discretize heat map → Create vector

Figure 1: Flow Chart of Persistent Feature Pipeline
Source: Bukkuri, Anuraag et al.[1]

dataframe of the 100 highest variance gene expressions except for this we did not use the location data. For this particular pipeline, we wanted to see if not including the location data would significantly lower accuracy in the classification. The data is then standardized using StandardScaler to ensure that each gene contributes equally to downstream analyses. Principal Component Analysis (PCA) with two components is then applied as a lens function to reduce the dimensionality of the gene expression data while preserving variance in the directions of greatest spread, making it easier to visualize the structure of the data.

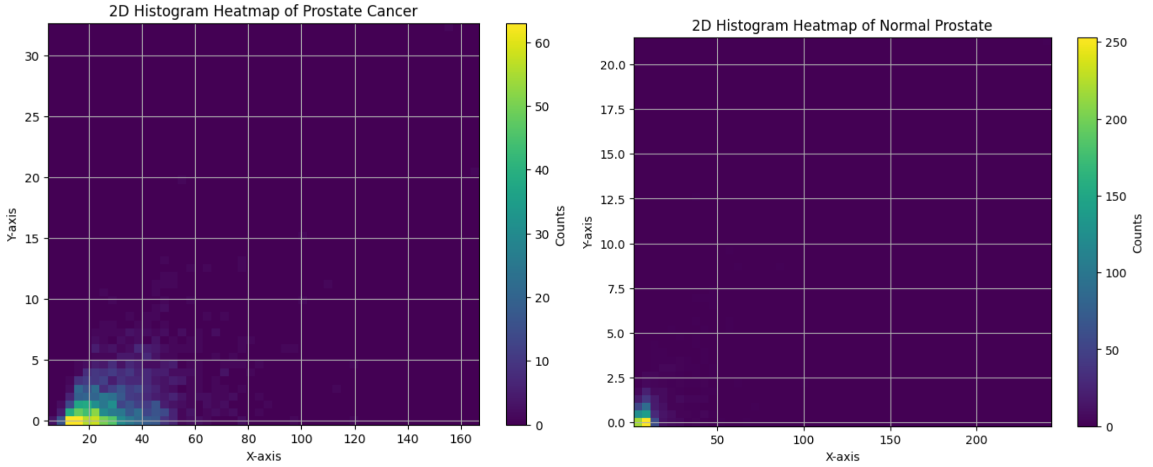
The Mapper graph is then constructed by segmenting the PCA lens into overlapping intervals using a cover with 50 cubes and 50% overlap, allowing for continuity and redundancy across regions. Each interval's data is then clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a density-based algorithm suitable for capturing complex, non-linear clusters in noisy biological data. The graph output captures the topological structure of the high-dimensional gene expression space, highlighting persistent features such as loops and branches that may correspond to cell types or spatial regions. Lastly, the Mapper graph is visualized and saved as an interactive HTML file which enables exploration of the topological network.

# 5 Results

## 5.1 Persistent Topology

Looking at the results, there seems to be promise for classifying data from the same organ and of similar tissue size. As seen in the plots in figure 2, the cancer plot has a a much wider spread than the non-cancerous tissue. Given the knowledge that cancerous tissues tend to have different gene expressions than non-cancerous tissues, we would expect this to happen as the there would be bigger "holes" within the data in certain places as the cells that would usually express those genes are no longer expressing them. Furthermore, we can also see that the heat map also tends to drift further away from the line $y = 0$ indicating that these are strong homology features and not some extra noise. These smudges in the maps would be very apparent and may help in easy identification of cancerous samples.

However, when extending these findings to other tissue, especially sample much larger than the initial ones, the heat map does not seem to be a very good analysis tool. For example, there was a much larger dataset on tissue from the brain (Figure 3), but due to the number of points, the heat map does not highlight areas that are far from the x-axis. Furthermore, we must also notice that even in Figure 2, the axis change substantially and thus it may be difficult to set some boundary for points that are sufficiently far from the x-axis to be considered significant. In conclusion, we would not be able to say that Persistent Homology heatmaps can necessarily be a good indicator for the presence of cancer within a sample as the heat maps themselves vary differentlt from sample sizes and from tissue types but may be an easy way to take an initial look given how easy it is to notice the larger spread in the diagram.



(a) Cancerous Tissue

(b) Normal Tissue

Figure 2: Persistent Homology Heatmap for Prostate Tissue

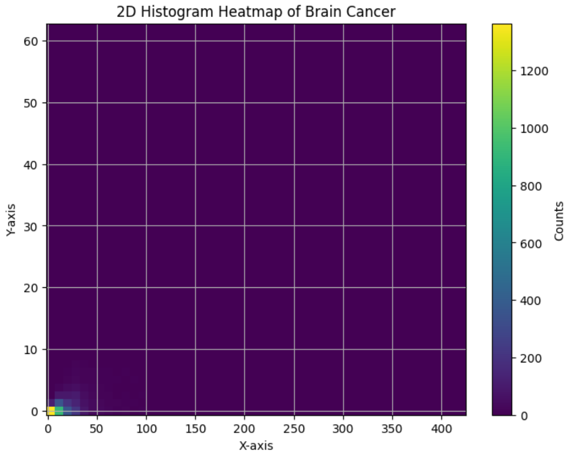
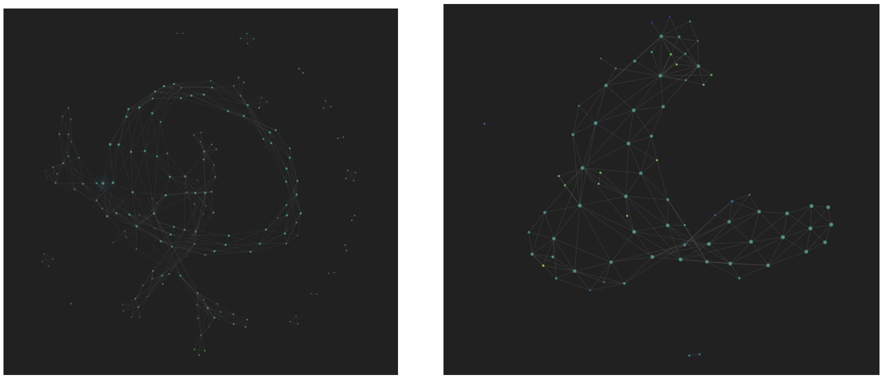


Figure 3: Very Large Cancer Tissue

### 5.2 Mapper Function

While the approach using persistent topology proved to be somewhat inaccurate for classifying varying different datasets, the visualizations created using the Mapper function were generally very stable for different types of tissues and with different sizes. While the initial parameter tuning was somewhat difficult, the features that were generated would be fairly consistent throughout all the datasets. What we looked for in this data was the clustering and particularly, how many connected clusters were there in the resultant graph. There was a strong correlation between the more connected components a graph had with the more likely it is to be cancerous. Based on the cursory observation from the results, the boundary was 6 connected components with less than that indicating non-cancerous tissue and more than that indicating cancerous tissues. As they are shown in Figure 4, the visualizations are a bit hard to discern the different connected components from each other but to better look at the results, try going to `https://krazykats.github.io/TDA_cancer/website_files/mapper_x10_synthetic_copy.html` and `https://krazykats.github.io/TDA_cancer/website_files/mapper_x10_synthetic_first.html` to see and test out this data visualization.



(a) Cancerous Tissue

(b) Normal Tissue

Figure 4: Mapper Graph Visualization

While the Visualization may be harder to recognize at a first glance whether a tissue is cancerous or not, this methodology hold much better with larger data and is generally robust to variance between types of tissues. An issue may arise where within a sample there are several different types of tissues from different organs but given a reasonable tissue dataset, the connected components can accurately classify the tissues as cancerous or not.

## 6 Challenges and Limitations

TDA offers a powerful framework for uncovering hidden patterns in complex biological datasets, but its application is hindered by significant challenges. The computational cost of persistent homology, a core component of TDA, is a major limitation, particularly when processing large datasets common in fields like genomics or neuroimaging. This computationally intensive process requires substantial resources, often making it impractical for real-time analysis or studies with limited computational infrastructure. Additionally, the quality of input data poses a challenge, as TDA relies on high-resolution spatial data, which is still emerging in many biological contexts. The absence of such data can compromise the accuracy and reliability of topological insights, limiting TDA's applicability in cutting-edge research where comprehensive datasets are not yet available.

Another critical challenge lies in the interpretability and parameter sensitivity of TDA methods, such as the Mapper algorithm. The results produced by Mapper are highly dependent on the choice of lens and clustering parameters, which can lead to variability in outcomes and complicate reproducibility across studies. This sensitivity demands careful parameter tuning, which can be a barrier for researchers without deep expertise in TDA. Furthermore, while TDA excels at identifying topological features, translating these abstract structures into meaningful biological insights remains difficult. The lack of clear biological interpretability can hinder the integration of TDA findings into practical applications, such as clinical diagnostics or therapeutic development, where actionable conclusions are essential.

## 7 Future Directions

The integration of TDA with advanced computational techniques, such as geometric deep learning, holds immense promise for revolutionizing oncology research and clinical practice. By combining TDA's ability to uncover intricate patterns in complex datasets with the predictive power of artificial intelligence, researchers can enhance pattern recognition in high-dimensional cancer data, such as imaging or molecular profiles. This synergy could lead to more accurate identification of subtle disease signatures, enabling earlier detection and personalized treatment strategies. Furthermore, the application of TDA in real-time scenarios, such as intraoperative tumor margin detection, could transform surgical precision. By providing surgeons with topological insights into tumor boundaries during procedures, TDA has the potential to improve outcomes by ensuring complete tumor resection while minimizing damage to healthy tissues.

Another exciting frontier lies in leveraging TDA for multi-omics integration and dynamic modeling in cancer research. By combining topological approaches with genomics, transcriptomics, and proteomics, TDA can provide a holistic view of cancer biology, capturing interactions across multiple biological layers that are often missed by traditional methods. This comprehensive perspective could uncover novel biomarkers and therapeutic targets, driving the development of precision medicine. Additionally, temporal TDA offers a powerful tool for tracking tumor evolution and treatment response over time. By modeling the dynamic changes in tumor topology, researchers can better understand cancer progression and resistance mechanisms, paving the way for adaptive treatment strategies that respond to a tumor's evolving landscape, ultimately improving patient outcomes.

## 8 Conclusion

Topology offers a novel lens through which to view the complex landscape of cancer. By focusing on shape rather than specific coordinates, TDA provides a robust and interpretable framework for discovering patterns in noisy, high-dimensional data. As more high dimensional data becomes available, data scientist and oncologists can look to TDA to both explore, visualize, and apply these sets to analysis and machine algorithms.

## References

[1] Bukkuri, Anuraag et al. “Applications of Topological Data Analysis in Oncology.” Frontiers in artificial intelligence vol. 4 659037. 13 Apr. 2021, doi:10.3389/frai.2021.659037