Supplementary Material for “Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis”

Lorin Crawford[1-3,†], Anthea Monod[4,†], Andrew X. Chen[5], Sayan Mukherjee[6-9], and Raúl Rabadán[5]

1 Department of Biostatistics, Brown University, Providence, RI, USA
2 Center for Statistical Sciences, Brown University, Providence, RI, USA
3 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
4 Department of Applied Mathematics, Tel Aviv University, Tel Aviv, Israel
5 Department of Systems Biology, Columbia University, New York, NY, USA
6 Department of Statistical Science, Duke University, Durham, NC, USA
7 Department of Computer Science, Duke University, Durham, NC, USA
8 Department of Mathematics, Duke University, Durham, NC, USA
9 Department of Bioinformatics & Biostatistics, Duke University, Durham, NC, USA

† Corresponding E-mail: lorin_crawford@brown.edu; antheam@tauex.tau.ac.il

Contents

Background on Topological Data Analysis (TDA)

In this supplemental text, we provide more formal details on the mathematics underlying the concepts of persistent homology and topological data analysis (TDA) for shapes and images. For a complete discussion on theory in TDA and applied topology, the interested reader may refer to a detailed literature (e.g. Ghrist, 2008; Carlsson, 2009, 2014).

Homology, Simplicial Complexes, and Persistence

Homology groups provide an algebraic structure to study holes in a topological space. Such holes are captured indirectly by considering what surrounds them. In other words, homology is concerned with studying the boundaries of holes. The fundamental property underlying homology is that the boundary of a boundary is necessarily zero. Algebraicity of such a study refers to group operations and maps that relate topologically-meaningful subsets of a space with one another. In discretizing a general topological space in terms of simplices, and thus studying its simplicial homology, the underlying object of study is a simplicial complex. This is the context from which we work in this paper, and what will be described in detail in this section.

A *k-simplex* is the convex hull of $k+1$ affine independent points $v_0, v_1, \ldots v_k$, and is denoted by $\sigma = [v_0, v_1, \ldots, v_k]$. Examples of k-simplices are points, lines, and triangles. The 0-simplex $[v_0]$ is the vertex v_0; the 1-simplex $[v_0, v_1]$ is the edge between the vertices v_0 and v_1; and the 2-simplex $[v_0, v_1, v_2]$ is the triangle bordered by the edges $[v_0, v_1]$, $[v_1, v_2]$ and $[v_0, v_2]$.

Definition S1. *A geometric simplicial complex K is a countable set of simplices such that:*

1. *Every face of a simplex in K is also in K;*
2. *If two k-simplices σ_1, σ_2 are in K, then their intersection is either empty or a face of both σ_1 and σ_2.*

Fix a dimension k and a field $\mathbb{F}$. Given a shape M with a finite simplicial complex representation (mesh) K, a *simplicial k-chain* is a linear combination of k-simplices $\sum_k c_k\sigma_k$, where $c_k \in \mathbb{F}$ and $\sigma_k \in K$. Here, the sum is taken over all possible k-simplices. Denote the set of all such k-chains by $C_k(K)$. These k-chains may be added (given $c = \sum_k c_k\sigma_k$ and $d = \sum_k d_k\sigma_k$, with $c+d := \sum_k(c_k + d_k)\sigma_k$) and multiplied by scalars. Thus, $C_k(K)$ is a vector space over $\mathbb{F}$ of k-chains in K, and the set of k-simplices forms a canonical basis for $C_k(K)$.

The theory and results developed in this paper rely on the simplifying assumption that $\mathbb{F}$ is the binary field $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$. In this case, a k-chain is a collection of k-simplices, and the *boundary* of a k-simplex is the sum of its $(k-1)$-dimensional faces. Let $\sigma = [v_0, v_1, \ldots, v_k]$ denote the simplex spanned by the specified vertices. The *boundary map* $\partial_k : C_k(K) \to C_{k-1}(K)$ maps a k-chain to a $(k-1)$-chain, and is given by

$$\partial_k\big([v_0, v_1, \ldots, v_k]\big) = \sum_{j=0}^{k}[v_0, \ldots, v_{-j}, \ldots, v_k]$$

with linear extension, where v_{-j} specifies that the j-th element is dropped. Elements of $B_k(K) := \operatorname{im}\partial_{k+1}$ are called *boundaries*, and elements of $Z_k(K) := \ker\partial_k$ are called *cycles.*

Definition S2. *The k-th homology group of M is defined by the quotient group*

$$H_k(K) := Z_k(K)/B_k(K).$$

The intuition behind a homology group is that it contains information about the structure of K. The zeroth homology group $H_0(X)$ is generated by elements that represent connected components of X. For example, if X has three connected components, then $H_0(X) \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, where $\cong$ here denotes group isomorphism. For $k \geq 1$, the k-th homology group $H_k(X)$ is generated by elements representing k-dimensional "holes" or "loops" in X. A k-dimensional hole can be thought of as the result of taking the boundary of a $(k+1)$-dimensional body. The ranks of the homology groups (i.e. the number of generators) are called the *Betti numbers*, and are denoted by $\beta_k(X) := \operatorname{rank}\bigl(H_k(X)\bigr)$. The notation $H_*(X)$ refers to all the homology groups simultaneously. In Figure S5, we display the homology of a torus constructed from a simplicial complex.

Persistent Homology

A *filtration* of simpicial complexes $\mathcal{K}$ is an indexed, nested family of spaces $\mathcal{K} = \{K_s\}_{s=a}^{b}$ such that $K_{s_1} \subseteq K_{s_2}$ if $s_1 < s_2$. As the parameter s increases, the homology of the spaces K_s may change (e.g. components are added and merged, cycles are formed and filled up). Examples of two different filtrations are given in Figures 1 (in the main text) and S6. The former illustrates a filtration by height function where the topological structure, in terms of simplicial homology, is revealed at the level of maximal height as $s = x$ increases in the vertical direction ν. The latter illustrates a filtration by radius: the sample points from the space are taken to be centers of balls while the radius $s = \epsilon$ grows from 0 to ∞. Overlapping balls are replaced by a k-simplex, depending on the degree of overlap: two overlapping balls are replaced by an edge, three overlapping balls are replaced by a face or triangle, and so on. Formally known as the Vietoris–Rips filtration, this procedure reveals the topological structure in terms of simplicial homology.

The *persistent homology* of $\mathcal{K}$ is denoted by $\mathrm{PH}_*(\mathcal{K})$ and keeps track of the progression of homology groups generated by the filtration. More specifically, the persistent homology contains the information about the homology of the individual spaces $\{K_s\}$, as well as the mappings between the homology of K_{s_1} and K_{s_2} for every $s_1 < s_2$. Note that persistent homology is also equivalently referred to as *persistence*.

Definition S3. *Let K be a filtered simplicial complex with $K_1 \subset K_2 \subset \cdots \subset K_S = K$. The k-th persistence module derived in homology, or k-th persistent homology, of K is*

$$\mathrm{PH}_k(K) := \bigl\{H_k(K_s)\bigr\}_{1 \leq s \leq S} \textit{ with } \{\varphi_{s_1,s_2}\}_{1 \leq s_1 \leq s_2 \leq S},$$

where each linear map $\varphi_{s_1,s_2} : H_k(K_{s_1}) \to H_k(K_{s_2})$ is induced by the inclusion $K_{s_1} \hookrightarrow K_{s_2}$ for all $s_1, s_2 \in [1, S]$ with $s_1 \leq s_2$.

Intuitively, the main idea behind persistent homology is to study homology across multiple scales. Rather than restricting ourselves to only one instance of a space, in persistent homology, we study the evolution of the topological structure over a filtration of the entire space. This amounts to beginning with a rigid proximity rule connecting observed data points, and then continuously relaxing this rule whilst studying the corresponding topological progression.

Barcodes and Persistence Diagrams

The persistence of the data is encoded in parameterizations of homology groups known as *barcodes*—collections of intervals corresponding to the lifetimes of topological features. The left endpoint of a bar is the *birth time* of an element in $\mathrm{PH}_*(\mathcal{K})$ and can be thought of as the value of s where this element appears for the first time. Conversely, the right endpoint of a bar is the *death time* and represents the value of s where an element vanishes, or merges with another existing element. The convention in studying merging features is to retain and treat the feature that appeared first as if it were continuing its existence beyond the merge event.

Figure S6 illustrates the idea behind persistent homology and provides an example of a barcode (Ghrist, 2008). In this example, ϵ corresponds to the filtration parameter value. The filtration is illustrated in the upper panels in the evolution of the simplicial complex in terms of vertices, edges, and faces that are formed with the progression of ϵ. Here, the H_0 homology captures connected components; H_1 captures cycles whose boundaries are formed by edges between vertices; and H_2 captures cycles with boundaries formed by faces. The dashed lines extending from the panels capture instances of the filtration, and link to the bars representing the topological features at particular values of ϵ. As ϵ progresses, we see connected components merge, cycles form, and fill up. The barcode summarizes the lifetimes of all topological features, classified by their homology groups, in this process.

Barcodes can thus be considered as summary statistics of the data generating process, in the form of collections of intervals. This information can alternatively be represented by a *persistence diagram*, which takes the birth and death times of each bar in a barcode as an ordered pair (x, y) and produces a scatterplot. This provides an alternative multi-scale topological summary of the shape or surface. In a persistence diagram, the points lie in $\mathbb{R}^2_{\geq 0}$ and all the points on the diagonal $x = y$ have infinite multiplicity. The diagonal is included to allow for well-defined metrics on the space of persistence diagrams (or barcodes).

Since summary statistics are direct parallels to the invariants of a topological space, considering such topological approaches in data analysis is an intuitive way of reducing dimensionality in high-dimensional statistical problems (i.e. where the number of predictors is far greater than the number of observations).

Supplementary Figures

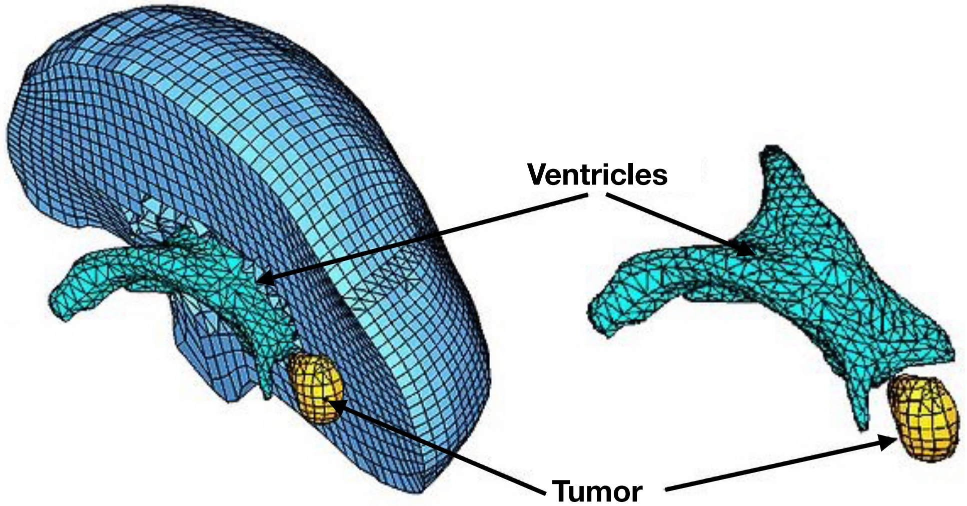


Figure S1. A mesh representation of a brain tumor and ventricles.

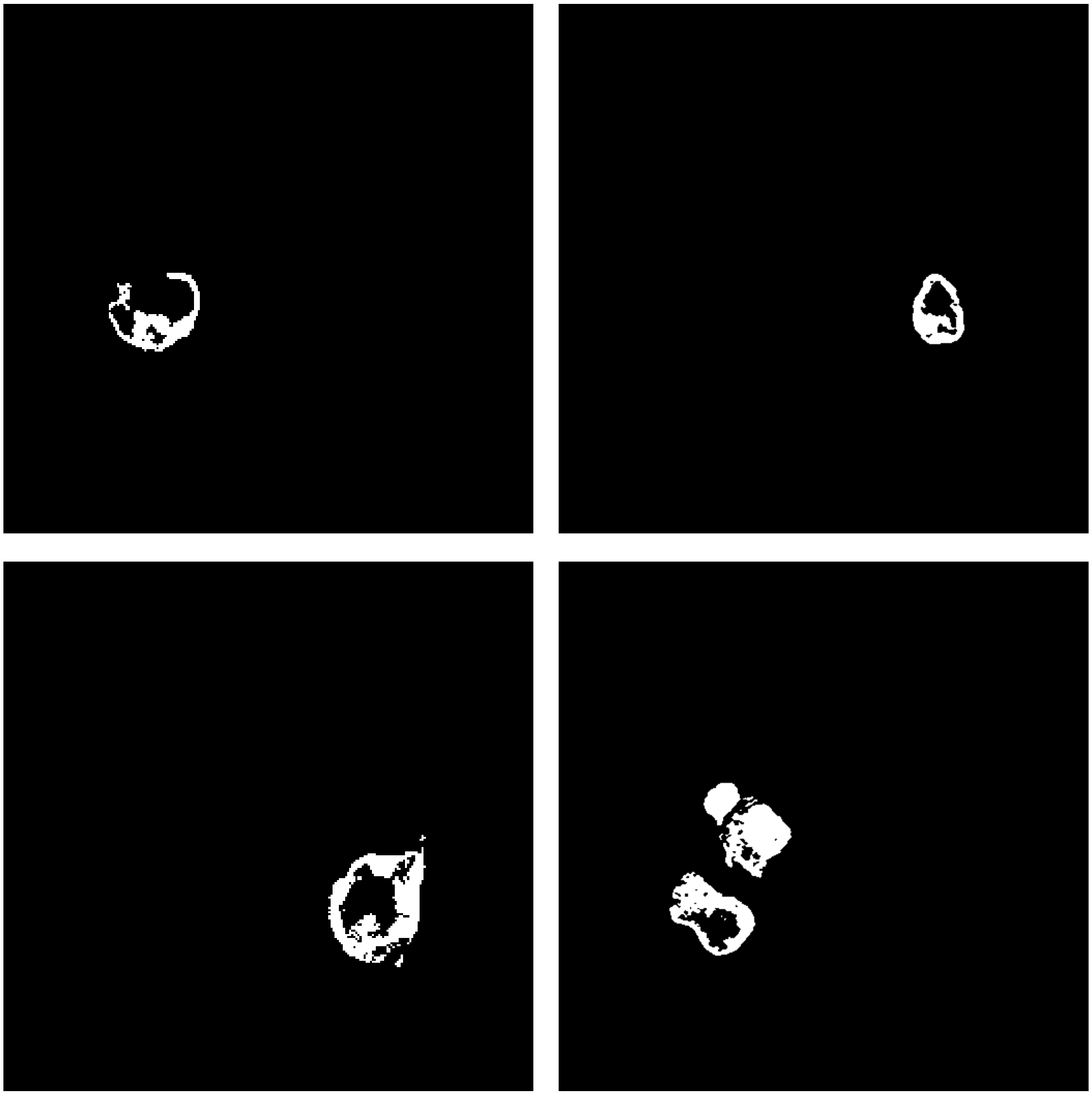


Figure S2. Examples of tumors exhibiting necrosis and multifocality. Note that all four images are taken from different patients to highlight the diversity of disease progression. These images were segmented from the original MRI scans using the MITKats algorithm (Chen and Rabadán, 2017).

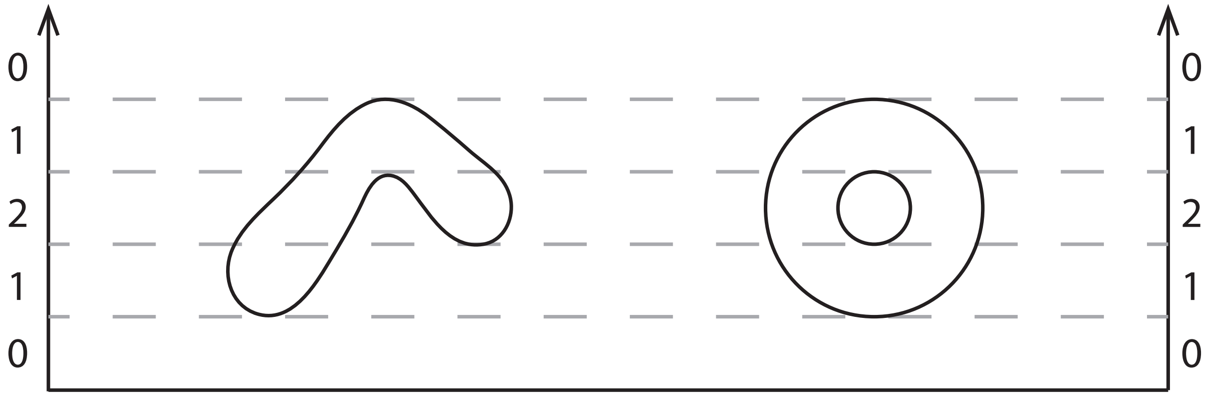


Figure S3. Counterexample for injectivity of the Euler Characteristic (EC) curve for a fixed direction. The vertical axes show the direction of the filtration for both shapes by height (i.e. the sublevel set filtration). The numbers on the axes denote the evolution of the EC for both shapes. We see that although the shapes are different, the corresponding ECs change in exactly the same manner, yielding identical ECTs for a fixed direction $\nu \in S^1$.

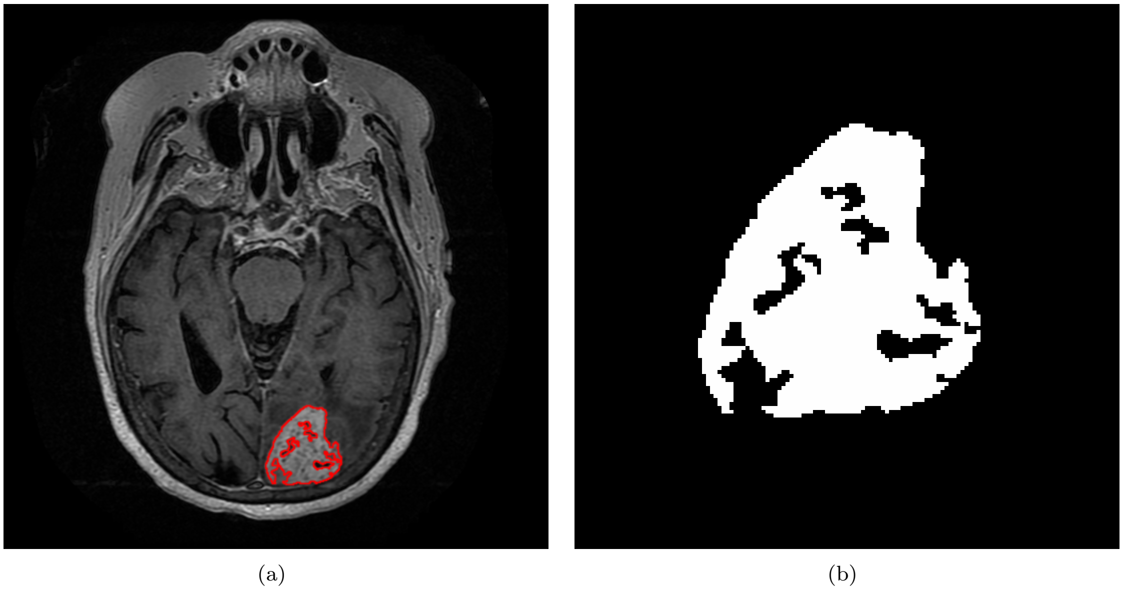


Figure S4. Example image data used in radiomic analysis. An original MRI from the TCIA and TCGA is displayed in Figure (a), while the final segmented image via the MITKats algorithm (Chen and Rabadán, 2017) is given in Figure (b).

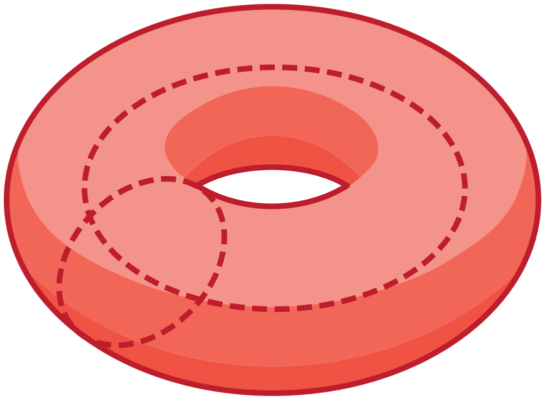
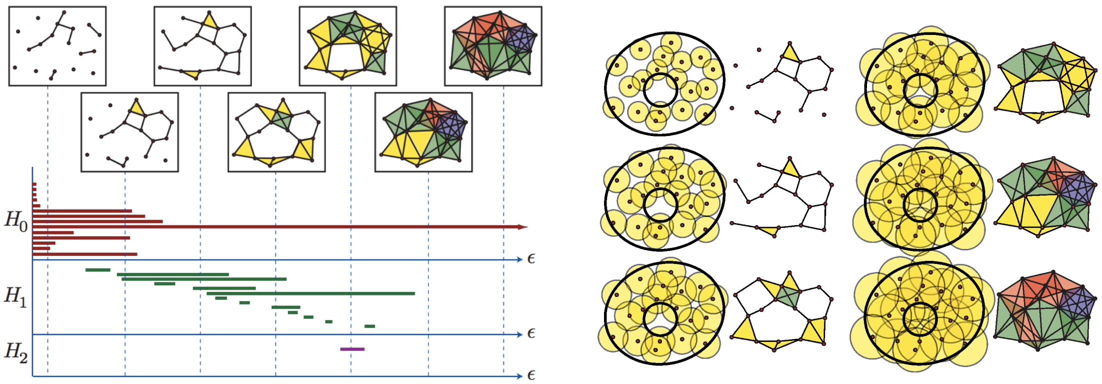


Figure S5. An illustrative example of homology using the 2-dimensional torus and its cycles. The torus has a single connected component and a single 2-cycle (the void locked inside the torus). In addition, it has two distinct 1-dimensional cycles (or closed loops) represented by the two curves in the figure. Consequently, the Betti numbers of the torus are $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$.



(a)

(b)

Figure S6. Illustrative example of persistent homology and the resulting barcode. In Figure (a), the shape of interest is the annulus, whose topology is given by $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$. Seventeen points are sampled from the annulus. Simplicial complexes are computed for continuous values of a filtration parameter, $\epsilon \in [0, \infty)$. Here, the filtration is given by the radii of balls centered at each sample point: the radius value is ϵ, and the associated simplicial complex is built by replacing two overlapping balls by an edge, three overlapping balls by a face, and so on, for higher dimensions. The filtration, in terms of the evolving simplicial complex, is illustrated in the upper panels. Vertices, edges, and faces are formed as the value of ϵ increases. H_0 corresponds to connected components; H_1 corresponds to cycles whose boundaries are formed by edges between vertices; and H_2 corresponds to cycles with boundaries formed by faces. The dashed lines extending from the panels connect to the bars representing the topological features appearing and existing at the corresponding values of ϵ. As ϵ progresses, connected components merge, cycles form, and fill up; the convention when two features merge is to retain the bar corresponding to the feature existing first. The barcode summarizes this progression of ϵ by tracking the "lifetimes" of the topological features according to their homology groups. Notice that there is a single H_0 bar that persists as $\epsilon \to \infty$, which represents the single connected component of the annulus. There are also several bars of varying length in H_1, including dominant bars, suggesting that the point cloud was unevenly sampled from the annulus in such a way that the sample space contains holes. Also, there is a single H_2 bar, which represents the cycle bounded by faces in the corresponding panel, but the length of the bar is comparatively short, and thus likely to be a spurious topological artifact. Figure (b) on the right shows the annulus with sampled points; the balls centered at the sampled points as the radius ϵ increases; and the corresponding simplicial complexes formed as the balls intersect (which correspond to the panels extending from specific bars in the barcode in Figure (a) on the left). This figure has been previously published (Ghrist, 2008).

References

Carlsson, G. (2009). Topology and data. *Bull Amer Math Soc 46*(2), 255–308.

Carlsson, G. (2014). Topological pattern recognition for point cloud data. *Acta Numer 23*, 289–368.

Chen, A. X. and R. Rabadán (2017). A fast semi-automatic segmentation tool for processing brain tumor images. In A. Holzinger, R. Goebel, M. Ferri, and V. Palade (Eds.), *Towards Integrative Machine Learning and Knowledge Extraction*, Cham, pp. 170–181. Springer International Publishing.

Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bull Amer Math Soc 45*(1), 61–75.