PERSISTENCE CURVES: BOUNDING THEIR DISTANCES AND APPLICATIONS TO CANCER GENOMICS

1. Introduction

Chromosome aberrations have been shown to be associated with cancer. In particular, copy number gains and losses can often be associated with oncogenes and tumor suppressor genes. Identifying these genes associated with particular phenotypes can lead to earlier detection of cancers, an understanding of the likely progression of that type of cancer and potentially genetic therapies.

The classic approach to identifying these important genes is through an association study. TAaCGH (Topological Analysis of array CGH), primarily differs from classic association studies in that it uses tools from Topological Data Analysis (TDA) as an intermediate step between the data and the statistics. The TAaCGH approach was introduced in varying forms through a series of papers [DCCW$^+$10], [ABD$^+$12],[ABC$^+$15], [ATGB$^+$16], [GUSA20]. In particular, we will focus here on the version developed in [ABC$^+$15].

This study focused on the 4 main molecular subtypes of breast cancer Luminal A, Luminal B, basal-like and HER2+. It splits chromosome arms up into contiguous segments and associates the 4 phenotypes with specific regions. This is done by associating to each test and control patient a curve called a betti-0 curve, taking the average over all test and all control patients and using the distance between the curves as a test statistic. Here we focus on understanding just how far two persistence curves can be from each other and upgrading this approach to use other types of persistence curves such as lifespan curves and persistence landscapes. We compare the results of both significant regions discovered in the Horlings data set by betti and lifespan curves and persistence landscapes as well as their results on simulated data.

2. Methods

2.1. **Topological Data Analysis.** One of the key tools in topological data analysis is persistent homology. Point clouds are first constructed data, simplicial complexes are built from this point cloud and the persistent homology of these simplicial complexes then yields topological information about the shape of the original data. We present a schematic of the process below:

$$\text{data} \xrightarrow{\gamma} \text{point cloud} \xrightarrow{\delta} \text{simplicial complex} \xrightarrow{\phi} \text{Persistent Homology} \xrightarrow{\psi} \text{TDA Summary.}$$

One of the advantages of this framework is its flexibility. The type of point cloud can be adjusted, such as converting a time series to a sliding window point cloud. The simplicial complex can be chosen such as the Vietoris-Rips, Čech, alpha or witness complexes. The types of topological summary can also be chosen such as persistence curves [CL19], persistent landscapes [Bub15] or persistence images [AEK$^+$17]. By varying these choices one can hone in on specific properties of the data.

2.1.1. γ *Choice of point cloud.* Sometimes data comes packaged in a point cloud ready for topological data analysis techniques to be applied. Other times the data needs to be converted into a point cloud so that meaningful topological features can be extracted. For example, Harer-Perea introduce a method for detecting periodicity of time series data using persistent homology in [PH15]. In their algorithm they first convert the time series data to what is known as the sliding window

Date: October 8, 2021.

point cloud, construct a Vietoris-Rips complex from the point cloud and finally apply persistent homology.

Since then, the sliding window point cloud has been used as a step in TAaCGH to detect CNAs in cancer patients ([DCCW$^+$10], [ABC$^+$15], [GUSA20], [ATGB$^+$16]). Given a time series $x_1, x_2, \ldots, x_n$ the sliding window point cloud with window size 2 of this time series is

$$(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n), (x_n, x_1).$$

For example, the time series on the left in Figure 1 yields the sliding window point cloud on the right in Figure 1. Since nearby genes tend to have similar copy numbers, combining the copy

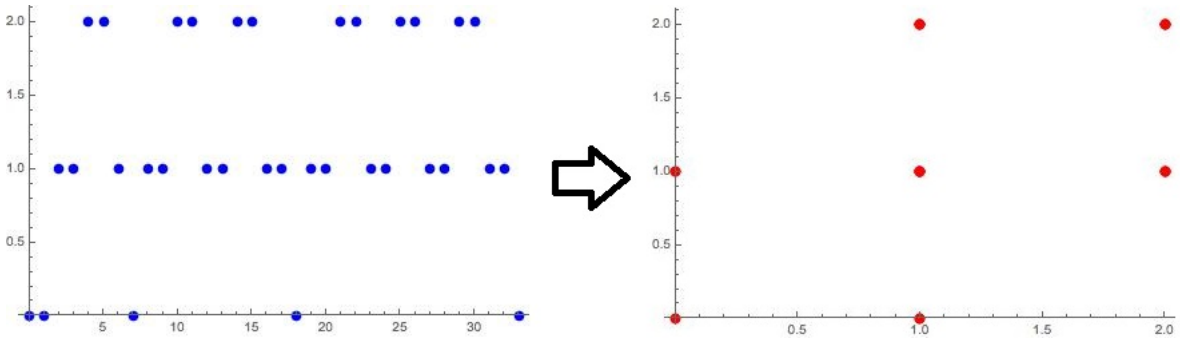

FIGURE 1. The Sliding Window Mapping

number ratios from adjacent points points into one point makes sense heuristically. Many copy number aberrations can occur throughout the genome of a patient. In [ATGB$^+$16], it is noted that 1-dimensional cycles of the sliding window point cloud of copy number data can capture co-occurring copy number aberrations. This makes both 0 and 1-dimensional persistence of the sliding window point cloud useful and we study both here.

2.1.2. *δ Choice of simplicial complex.* The Vietoris-Rips (VR) filtration is frequently used due to its interpretability and computability. The similarly defined Čech filtration more accurately captures the topology of the space, but is computationally more expensive. Specifically, the nerve lemma says that the Čech filtration is homotopy equivalent to the cover which it is the nerve of. It is still used when the number of data points and their dimensions are small. The VR filtration on a point cloud P, $\mathrm{VR}(P, \epsilon)$, is the simplicial complex on the points in P such that $\{p_1, \ldots, p_k\} \subseteq P$ is a simplex in $\mathrm{VR}(P, \epsilon)$ if and only if $B(p_i, \epsilon) \cap B(p_j, \epsilon)$ is nonempty. The Čech filtration, $\check{C}(P, \epsilon)$, has $\{p_1, \ldots, p_k\}$ as a simplex if and only if $\bigcap_{i=1}^{k} B(p_i, \epsilon)$ is nonempty. It is simple to check that both of these complexes are actually filtered simplicial complexes, that is, for $\epsilon < \delta$ $\mathrm{VR}(P, \epsilon)$ is a subcomplex of $\mathrm{VR}(P, \delta)$ and $\check{C}(P, \epsilon)$ is a subcomplex of $\check{C}(P, \delta)$.

Another filtration that is sometimes used is the alpha complex. Let $P = \{p_1, \ldots, p_m\} \subseteq \mathbb{R}^n$ be a point cloud. To each point $p_i \in P$ we associate the Voronoi cell of p_i, denoted V_{p_i} which is

$$V_{p_i} = \{x \in \mathbb{R}^n : d(x, p_i) \leq d(x, p_j) \text{ for all } p_j \in P\},$$

all of the points in $\mathbb{R}^n$ that are closer to p than the other points in the point cloud. Around each $p_i \in P$ we define $\tilde{B}(p_i, \epsilon) = B(P_i, \epsilon) \cap V_{p_i}$. The alpha complex on P, denoted $\alpha(P, \epsilon)$ is then the simplicial complex on the vertices of P such that $\{p_1, \ldots, p_k\}$ forms a simplex if and only if $\bigcap_{i=1}^{k} \tilde{B}(p_i, \epsilon)$ is nonempty. The alpha complex is homotopy equivalent to the Čech filtration, but in low dimensions can be easier to compute.

For this work it is also important to mention a separate interpretation for the VR filtration. Since a VR filtration is completely determined by its 1-skeleton it can be encoded by a sequence of graphs G_i on the vertices of a point cloud. The simplicial complex associated to each graph G_i

is the clique complex of the graph denoted $X(G_i)$. The clique complex of a graph is the simplicial complex on the vertices of the graph such that a set of vertices $\{p_1, \ldots, p_k\}$ forms a $k-1$ face if and only if it is a k-clique of the graph.

2.1.3. *ϕ Persistent homology.* After a filtered simplicial complex is constructed from the point cloud, the next step is to extract topological information e.g. the number of connected components or 1-cycles for each value of ϵ. To get this information the homology groups H_n of the complex at each ϵ are computed. The homology of a simplical complex X is derived from a chain complex. First an order is established on the vertices of X. Then the nth chain group $C_n(X)$ is the group of $\mathbf{R}$-linear combinations of simplices of X and each element of $C_n(X)$ is called an n-chain. Given an n-chain $\sigma = [v_0, v_1, \ldots v_n]$, the boundary homomorphism $\partial_n : C_n(X) \to C_{n-1}(X)$ is defined by $\partial_n(\sigma) = \sum_{i=0}^{n}(-1)^i[v_0, \ldots \hat{v_i} \ldots v_n]$ and then extending linearly where $\hat{v_i}$ means that vertex has been omitted. The nth homology group of X, $H_n(X)$ is then defined to be $\frac{\ker(\partial_n)}{\mathrm{im}(\partial_{n+1})}$.

Given a filtered simplicial complex X_ϵ with inclusion maps:

$$X_{\epsilon_0} \hookrightarrow X_{\epsilon_1} \hookrightarrow \ldots \hookrightarrow X_{\epsilon_N}$$

the functorality of homology takes the sequence of inclusions in the filtration to

$$H_n(X_{\epsilon_0}) \xrightarrow{\iota_{\epsilon_0}} H_n(X_{\epsilon_1}) \xrightarrow{\iota_{\epsilon_1}} \ldots \xrightarrow{\iota_{\epsilon_{N-1}}} H_n(X_{\epsilon_N})$$

for each dimension n where ι_{ϵ_i} are the maps induced on homology by the inclusions. We now denote by $\iota_{\epsilon_i \to \epsilon_{i+p}}$ the composition of the maps induced by the inclusions connecting X_{ϵ_i} to $X_{\epsilon_{i+p}}$. For example, $\iota_{\epsilon_1 \to \epsilon_3} = \iota_{\epsilon_2} \circ \iota_{\epsilon_1}$. The p-persistent n-dimensional homology group of X_{ϵ_m} is then defined to be

$$H_n^p\left(X_{\epsilon_m}\right) \coloneqq \iota_{\epsilon_m \to \epsilon_{m+p}}\left(H_n\left(X_{\epsilon_m}\right)\right).$$

The idea behind this definition is that n-dimensional holes that exist at lower filtration values will be filled in by simplices at higher filtration values. With this viewpoint in mind, for each element x of homology we can assign a birth-death pair (b_x, d_x) where b_x is the first value of ϵ where x appears and d_x is the first filtration value where x is trivial. Each of these pairs can be thought of as an interval $[b_x, d_x)$ representing the life of this element of homology. The length of the interval, $d_x - b_x$, is known as the persistence or lifespan of x. The n-dimensional Betti number of A_{ϵ_i} is exactly the number of persistence intervals which contain ϵ_i.

Many libraries now exist to compute persistent homology efficiently including Dionysus [Mor12], Gudhi [The21], Perseus [Nan12], Eirene [HG16], and Ripser [Bau19]. We use the R package TDA [FKLM14] which makes use of the Gudhi and Dionysus libraries.

2.1.4. *ψ Persistence Curves.* The number of points in persistence diagrams varies based on the values of the data. This makes many methods in statistics and machine learning difficult to apply directly. In order to overcome this issue, the topological information from persistent homology is often summarized using tools such as persistence curves [CL19], kernel SVM for persistence [RHBK15], persistence landscapes [Bub15] and persistence images [AEK$^+$17] among others. Here we focus on persistence curves particularly Betti and lifespan and persistence landscapes. It is worth noting that, in the 0-dimensional case, one generator could have infinite lifespan if all points were in a connected component. We therefore choose to consider reduced persistent homology in this case.

Let C be an n-dimensional persistence diagram, then the nth Betti curve denoted $\beta_n(C, t)$ is equal to the number of birth-death pairs $(b, d) \in C$ such that $t \in (b, d]$. Similarly, the nth lifespan curve denoted $\ell_n(C, t)$ is equal to the sum of the lifespans of all birth-death pairs $(b, d) \in C$ such

that $t \in (b, d]$. Persistent landscapes are a form of persistence curve introduced in [Bub15]. The kth persistence landscape of C denoted $\lambda(k, t)$ is

$$\lambda(k, t) = \text{kmax}_p([\min(t - b, d - t)]_+)$$

where $p = (b, d) \in D$, $[c]_+ = \max(c, 0)$ and kmax is the kth highest value.

Persistence curves were introduced in [CL19] as a general framework under which previously studied summaries of persistence diagrams lie. Some of these include the Betti curve, the life entropy curve [AGDST18] and the k-th landscape [Bub15]. The framework also allows for easy generation of new summaries including the lifespan curve. Lastly, persistence curves provide a way to make general arguments about the stability of the curves with respect to the bottleneck distance between persistence diagrams. Persistence landscapes were shown to be stable in [Bub15].

Theorem 1 from [CL19] provides general bounds on the difference between two persistence curves under the L^1 norm in terms of the bottleneck (W_∞) and 1-Wasserstein distances. The stability with respect to these two distances for many persistence curves is summarized in Table 1 from [CL19]. This table shows that, in general, the Betti curves and lifespan curves are not stable with respect to the bottleneck distance. Applying Theorem 1 from [CL19] to both these curves does, however, yield bounds on both these curves as computed in [CL19].

Theorem 1 ([CL19])**.** *Let C and D be persistence diagrams, W_∞ denote the bottleneck distance, n^C be the number of birth-death pairs in C and L^C denote the sum of the lifespans of all birth-death pairs in C. Then*

$$||\beta(C, t) - \beta(D, t)||_1 \leq 2 \max(n^C, n^D) W_\infty(C, D) + \min(L^C, L^D)$$
$$||\ell(C, t) - \ell(D, t)||_1 \leq 2(L^C + L^D) W_\infty(C, D).$$

2.2. **Data.**

2.2.1. *Horlings Dataset.* The dataset used in this study is from [HLN+10], the same dataset used in [DCCW+10], [ABC+15], [DCCW+10], [GUSA20]. It consists of BAC Microarrays from the genome with an average spacing of 1 Mb. Each BAC clone was spotted in triplicate on each slide (Code Link Activated Slides, Amersham Biosciences). It consists of 68 patient samples made from the 4 most common molecular subtypes of breast cancer: Luminal A, Luminal B, basal-like and HER2+. There are 21 Luminal A samples, 12 Luminal B samples, 21 basal-like samples and 14 HER2+ samples. We consider each molecular subset separately as the test sets and the remaining patients as the control set.

2.2.2. *Simulation data.* Simulated data sets contain 120 patient profiles with 60 patients in the test set and 60 in the control set with both tests and controls containing 100 probes each. The goal was to simulate patients that had single contiguous aberrations of a fixed length and compare them to patients with no aberrations using multiple techniques. The control profiles were sampled from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma \in \{0.2, 0.5\}$. The test profiles had contiguous aberrations of length $\lambda \in \{2, 3, 5, 10, 20, 50, 75\}$ sampled from a normal distribution with mean $\mu \in \{-1, 0.6, 1\}$ and standard deviation $\sigma \in \{0.2, 0.5\}$. The rest of each test profile was sampled from a normal distribution with mean 0 and $\sigma \in \{0.2, 0.5\}$ with the same standard deviation as for the aberration in the test profile. For each combination of parameters 20 simulations were run.

2.3. **Software.** To explore the effects of varying the length, mean and standard deviation of the simulated data on Betti and lifespan curves we built a Shiny application. This allows us to set the values of these parameters and see the probe values, the sliding window point cloud and the persistence curve built from a Vietoris-Rips cloud on that point cloud. It also allows us to see the same for a simulated control patient. Two outputs from this app are included in Figures 2 and

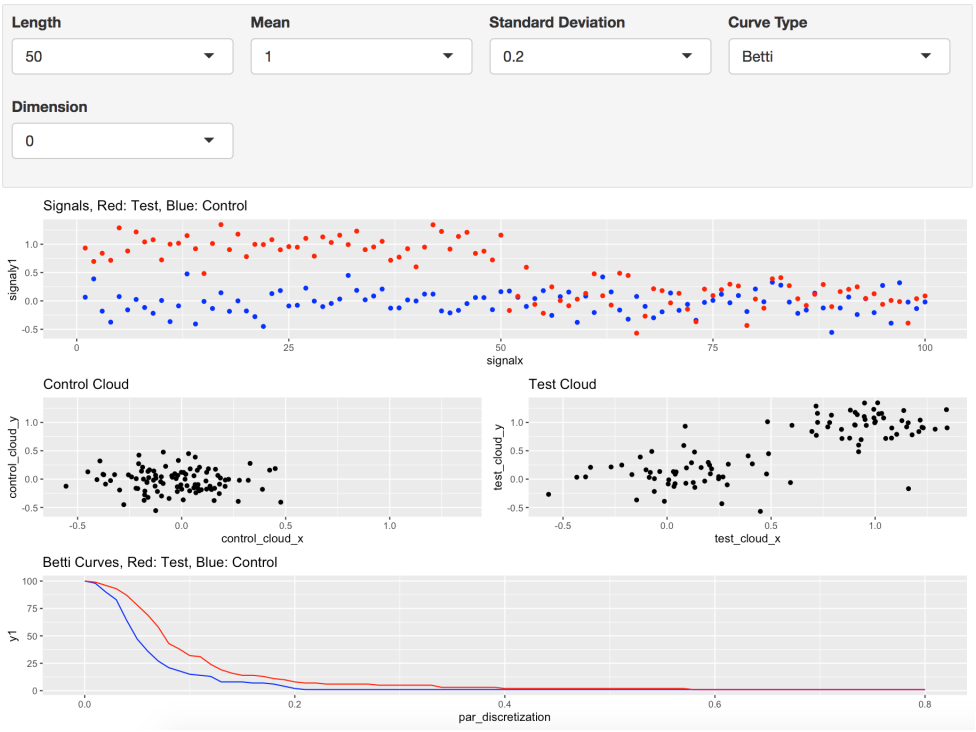

FIGURE 2. Simulated test and control patient profiles, associated sliding window point clouds and β_0 curves for $\sigma = 0.2$

Figure 3 with the same parameters other than standard deviation which is $\sigma = 0.2$ in the first figure and $\sigma = 0.5$ in the second.

In Figure 2, with a smaller standard deviation, there are two clusters within the test point cloud, one centered at $(0, 0)$ and one centered at $(1, 1)$. The control point cloud has just one cluster centered around $(0, 0)$. The β_0 curves of the Vietoris-Rips complexes of these point clouds identify this difference, the test curve remains above the control curve until there is one connected component. When the standard deviation is increased to $\sigma = .5$, however, both the control and test point cloud are similarly clustered. This leads to β_0 curves that are much more difficult to differentiate as pictured in Figure 3. In this case the test and control β_0 curves oscillate between the test or control having more connected components.

3. Bounds on Betti and Lifespan Curves

The main challenge in application of persistence curves, including Betti and lifespan curves, comes from the fact that small perturbations in the initial point cloud can lead to large changes in the curves [XATZ20]. The main result of this section is a bound on the L^1 norm between two Betti and or two lifespan curves built from finite and bounded point clouds with respect to the bottleneck distance.

Consider the bounds on the L^1 norm between two Betti curves or two lifespan curves from Theorem 1. The bottleneck distance is already stable with respect to small perturbations in the initial point clouds [CSEH07], so our bounds will be in terms of it. We need to find bounds on the maximal number of birth-death pairs in a persistence diagram, as well as the maximal lifespan of any birth-death pair.

First, we establish existence of bounds on these quantities under the given constraints for i-dimensional persistent homology. Then we explicitly compute bounds in the case of 1-dimensional

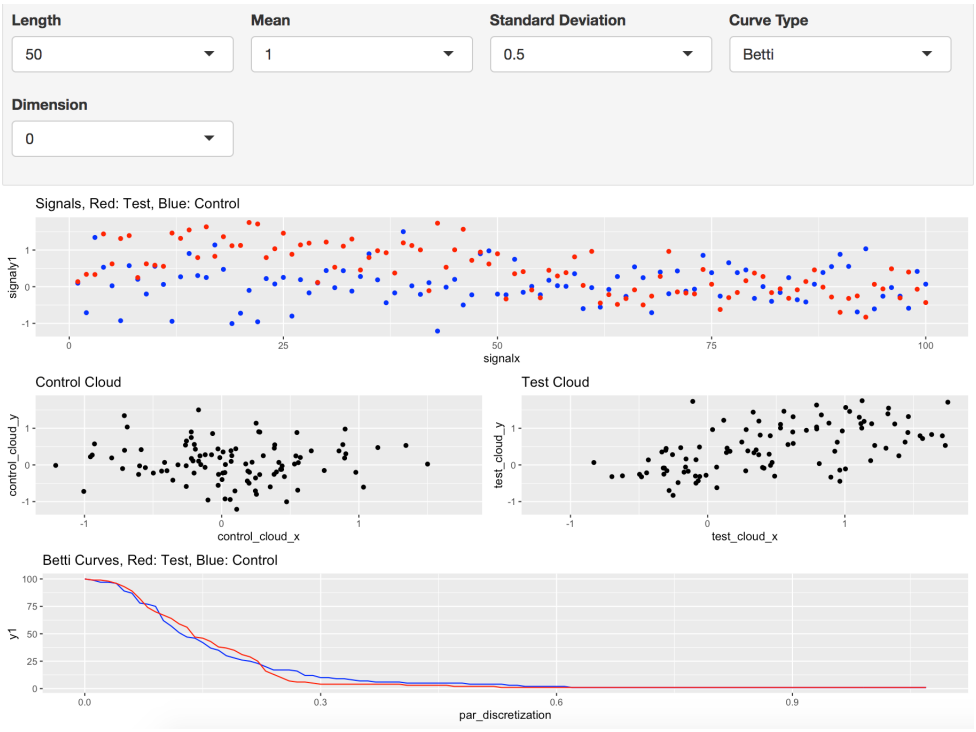

FIGURE 3. Simulated test and control patient profiles, associated sliding window point clouds and β_0 curves for $\sigma = 0.5$

persistent homology of the Vietoris-Rips complex. We also compute bounds in the case of 0-dimensional persistent homology for both the Vietoris-Rips and Čech complex. The existence results given here are for Vietoris-Rips and Čech complexes, but essentially the same arguments work for the various forms of witness complexes described in [Car09]. They also hold for the alpha complex since it is homotopy equivalent to the Čech complex.

Proposition 1. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud, then there exists a bound on the maximal number of birth-death pairs in the persistence diagrams of $\mathrm{VR}(P, \epsilon)$ and $\check{C}(V, \epsilon)$.*

Proof. Since P is finite, there are a finite number of simplicial complexes that can be built on P. Both the VR and Čech filtrations change a finite number of times, hence the persistence diagrams from these complexes have a maximal number of generators. □

Remark 1. *Recall we consider reduced 0-dimensional persistent homology to avoid the issue of an infinite lifespan generator.*

Proposition 2. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud with diameter d, then the maximal lifespan of a birth-death pair from the i-dimensional persistence diagram of $\mathrm{VR}(P, \epsilon)$ or $\check{C}(P, \epsilon)$ is d.*

Proof. Let $P = \{p_1, \ldots, p_k\}$ be a finite point cloud in $\mathbf{R}^n$. Consider the epsilon balls $B(p, \epsilon)$ for $p \in P$ and $\epsilon > d$. Since the diameter of P is d, each of these balls must contain all other points in P. Therefore, $B(p_i, \epsilon) \cap B(p_j, \epsilon) \neq \emptyset$ and $\bigcap_{i=1}^{k} B(p_i, \epsilon) \neq \emptyset$ so both $\mathrm{VR}(P, \epsilon)$ and $\check{C}(P, \epsilon)$ are $(k-1)$-simplices. Since simplices are contractible, there is no i-dimensional persistent homology for $i \geq 1$ and therefore the maximal lifespan is bounded above by m. In the 0-dimensional case we consider reduced homology. □

Theorem 2. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud with diameter d, then the i-dimensional Betti and lifespan curves of $\mathrm{VR}(P, \epsilon)$ and $\check{C}(P, \epsilon)$ are bounded with respect to small perturbations of P.*

Proof. Combine Theorem 1 with Proposition 1 and Proposition 2. □

Next results provides explicit bounds on the maximal number of non-homologous generators of 1-dimensional persistent homology in a Vietoris-Rips filtration. To do this, we require that the given point clouds have pairwise distinct distances between points. The following two results were made available to the authors through private correspondence with David Moon [MHBO]. The proofs provided here are different than those provided to the authors.

Proposition 3. *Let $G = (V, E)$ be a simple graph on n nodes, then the maximal 1st Betti number of the clique complex of G, $X(G)$, is $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil - (n-1)$.*

Proof. The first Betti number of G is $\beta_1(G) = |E| - n + 1$. Subtracting the number of triangles T in G from this quantity yields $\beta_1(X(G)) = |E| - n + 1 - T$. Let G_1 be a graph containing at least one triangle. Remove an edge from a triangle in G_1 to obtain G_2. G_2 has at least one less triangle than G_1, but only one less edge so $\beta_1(X(G_2)) \geq \beta_1(X(G_1))$. The graph G for which $\beta_1(X(G))$ is maximized must therefore be triangle-free. By Mantel's theorem [AZHE10], the triangle-free graph with the maximal number of edges is the complete bipartite graph $K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$. This graph has $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil$ edges which completes the proof.

□

Proposition 4. *Let $P \subseteq \mathbf{R}^d$ be a finite point cloud with n vertices such that the pairwise distances between points are distinct. Then the maximal number of birth-death pairs in the 1-dimensional persistence diagram of $\mathrm{VR}(P, \epsilon)$ is $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil - (n-1)$.*

Proof. The information from a VR filtration can be encoded by a sequence of graphs such that $G_i \subseteq G_{i+1}$. Since the distances between points in P are pairwise distinct, G_i differs from G_{i+1} by a single edge. If adding an edge $e = 12$ to G_i to form G_{i+1} completes a triangle 123, then e does not birth a new cycle in $H_1(X(G_{i+1}))$. To see this note that if e completes a cycle say $a_1, \ldots, a_k, 1, 2$ then this cycle is homologous to $a_1, \ldots, a_k, 1, 3, 2$ in $X(G_{i+1})$ since the two cycles differ by the triangle 123. Since the edges 13 and 32 were in G_i the cycle represented by $a_1, \ldots, a_k, 1, 3, 2$ was already in $H_1(X(G_i))$ and hence e did not birth a new cycle. Since triangles do not birth new cycles, any VR filtration which has the maximum number of birth-death pairs in a persistence diagram can have all generators alive at once. Therefore the maximum number of birth-death pairs over the entire filtration is the same as the maximum number of cycles that can be alive at a fixed filtration parameter. Proposition 3 completes the proof. □

Remark 2. *Proposition 4 improves a special case of Theorem* 3.1 *from* [Gof11] *for $n < 24$, which says that the maximal 1st Betti number of a Vietoris-Rips complex at a fixed filtration value is $5n$.*

The point clouds considered in this work have a maximum diameter. In particular, if the minimal and maximal copy number ratios are $c_{\min}$ and $c_{\max}$ then all sliding window point clouds are contained in $[c_{\min}, c_{\max}]^2$. Therefore, the maximum diameter of these point clouds is $\sqrt{2}\,(c_{\max} - c_{\min})$.

Theorem 3. *Let P and P' be sliding window point clouds built from copy-number ratio data with window sizes 2. Let P and P' each consist of n points. Let C and D be the 1-dimensional persistence diagrams coming from the Vietoris-Rips filtration built on P and P'. Then*

$$||\beta_1(C,t) - \beta_1(D,t)||_1 \leq \left(\left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil - (n-1) \right) \left(2W_\infty(C,D) + \sqrt{2}\,(c_{max} - c_{min}) \right)$$

$$||\ell_1(C,t) - \ell_1(D,t)||_1 \leq 4 \left(\left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil - (n-1) \right) \left(\sqrt{2}\,(c_{max} - c_{min}) \right) W_\infty(C,D).$$

In the 0-dimensional case for both the Vietoris-Rips and Čech filtrations it is clear that the maximal number of connected components in a point cloud with n vertices is n. This yields the following explicit bounds for β_0 and ℓ_0 curves in the case of copy-number ratio point clouds.

Theorem 4. *Let P and P' be sliding window point clouds built from copy-number ratio data with window sizes* 2. *Let P and P' each consist of n points. Let C and D be the* 0*-dimensional persistence diagrams coming from either Vietoris-Rips or Čech filtrations built on P and P'. Then*

$$||\beta_0(C,t) - \beta_0(D,t)||_1 \leq n\left(2W_\infty(C,D) + \sqrt{2}(c_{max} - c_{min})\right)$$

$$||\ell_0(C,t) - \ell_0(D,t)||_1 \leq 4n\left(\sqrt{2}\,(c_{max} - c_{min})\right) W_\infty(C,D).$$

The bounds from Theorem 4 apply to persistence diagrams that come from a single test patient and a single control patient. In the TAaCGH pipeline the curves from all of the test patients are averaged together, then the curves from all of the control patients are averaged together and finally these average curves are compared. Since a mean cannot be defined on persistence diagrams, this bound cannot be extended to a bound on average persistence curves.

4. TDA tools for classifying breast cancer types

In [ABC$^+$15], the copy number data from the chromosome arm of each patient is split up into consecutive segments of length 20 probes each. Each segment overlaps with the next segment in 10 probes. Next the Horlings data set is split into a test and control set. The process tests for significance at each individual segment. The test set consists of all patients from one cancer subtype and the control set consists of all other patients. The data for a fixed segment is converted into a point cloud in $\mathbb{R}^2$ using the sliding window mapping for each patient. A Vietoris-Rips complex is built on each point cloud and the 0-dimensional persistent homology of each Vietoris-Rips complex is taken. The persistent homology is then summarized into a Betti-0 curve for each patient. Next the Betti curves from all patients in the test set are averaged together and similarly for the control set. The L^2 norm of the difference between the average test and control betti curves is used as the test statistic. Permutation testing with FDR correction for multiple testing is used to test the significance of the statistic and thus of each segment for each cancer subtype. In this study we explore the significant regions of the genome that the 0 and 1-dimensional lifespan curves, ℓ_0 and ℓ_1, detect. We compare these regions to the regions detected by β_0. We also compare 0-dimensional persistence landscape functions to β_0.

4.1. Building and Validating New Software.

Since jPlex [SVJ11] is no longer updated, we built new software for the TAaCGH process in R. In order to validate the code, we recreated the study from [ABC$^+$15] with β_0 curves. We consider the significant regions that were detected before removing any outlier patients or altering the control sets. The regions that were detected by [ABC$^+$15] under these conditions for each of the cancer subtypes, as well as the regions detected when this experiment was performed again, are contained in Table 1.

For Luminal A, and HER2+ the regions detected before were all detected again. In the case of Luminal B one new region was detected $8p3$. In the case of Basal, most of the same regions were detected except $3p3, 3p6, 4q3, 5p2-3$ and $9q2$. Two regions that were not detected before were also detected this time $2p4$ and $14q1$. The undetected regions are shown in blue and the newly detected regions are shown in red in Table 1. These minor differences between the detected significant regions are likely due to rounding differences in jPlex to the R TDA library and the fact that many of the differing regions had p-values close to .05 in the original study.

4.2. Lifespan Curves.

Lifespan curves can be thought of as a weighted version of Betti curves, where each generator at a particular filtration value is weighted by its lifespan. These curves have a smaller bound on the L^1 norm between them and they also have the potential to discover different significant regions because they take into account different topological information than Betti curves. For these reasons, the experiment from [ABC$^+$15] was repeated using lifespan curves on 0-dimensional persistent homology, instead of Betti curves. The significant regions that were detected using this method are in Table 2.

	Lum. A	Lum. B	HER2+	Basal
β_0 Old	$11q6$		$17q1-2, 17q3, 17q4$	$1p1-7, 1p10-11, 1q2, 2p7, 3p1-2,$ $3p3, 3p6, 3p7-8, 4p3, 4q3, 4q4-12, 5p1,$ $5p2-3, 5q1, 6p1-2, 6p4, 6q8-9, 7p2,$ $9p1, 9p2, 9q2-3, 9q7-8, 10p1-3, 10q1-2,$ $10q4-8, 12p1-2, 13q2-6, 13q8, 14q2,$ $14q7-8, 15q1-3, 15q6, 18q1-4, 23p1-3$
β_0 New	$*$	$8p3$	$*$	$*, 2p4, 3p3, 3p6\ 4q3, 5p2-3, 9p2, 14q1$

TABLE 1. Comparison of the original β_0 study [ABC$^+$15] on the Horlings data [HLN$^+$10] and the reproduced β_0 study. Red indicates a newly detected region, blue indicates undetected regions and * indicates the remaining regions from [ABC$^+$15] that were detected.

	Luminal A	Luminal B	HER2+	Basal
Lifespan-0	$1p1, 2q2, 5p3, 15q$		$17q1-4$	$1p1, 1p6, 2p4, 2p7, 3p1, 4p3,$ $4q4, 4q6-11, 6p4,$ $10p1-3, 10q5-6$ $12p2, 13q5-6, 15q2-3$ $23p1-2$

TABLE 2. Significant regions detected by ℓ_0, the 0-dimensional lifespan curve. Red indicates the segments that were not detected in [ABC$^+$15].

The lifespan curve ℓ_0 detected all the regions for HER2+ patients, $17q1-4$, that were previously detected in [ABC$^+$15]. It detected 4 new regions for luminal A patients and it detected some of the segments that were detected for basal in [ABC$^+$15], as well as one new segment, $2p4$. It did not detect any significant regions for Luminal B. The experiment was also repeated for dimension 1 persistent homology using lifespan curves. The results are in Table 3. In this case, no significant regions were detected for Luminal A or B, only one region, $17q2$, was detected for HER2+ and 9 regions were detected for basal. These regions are $1p2, 1p11, 4q4, 4q7, 5p3, 6p4, 10p1, 10q6$ and $13q3$. All of these regions were previously detected in [ABC$^+$15].

Overall, lifespan curves detect significantly fewer regions than β_0 curves, though ℓ_0 curves detect more regions for Luminal A than β_0 curves do. This indicates that the number of generators on average between Luminal A patients and control patients is similar, but the lifespan of these generators differs. Since no regions were detected for Luminal B by β_0, ℓ_0, or ℓ_1, this indicates that similar persistence curves are unlikely to detect significant regions for Luminal B. The ℓ_1 curve only detects regions which were previously discovered by β_0 in [ABC$^+$15], so in this case offers no additional information.

	Luminal A	Luminal B	HER2+	Basal
Lifespan-1			$17q2$	$1p2, 1p11, 4q4, 4q7, 5p3, 6p4, 10p1, 10q6, 13q3$

TABLE 3. Significant regions from the Horlings dataset [HLN$^+$10] detected by ℓ_1, the 1-dimensional lifespan curve.

4.3. **Persistence Landscapes.** Recall that, as opposed to Betti and lifespan curves, persistence landscapes are stable [Bub15]. Persistence landscapes also have the added advantage of there being multiple landscape functions that can be applied to a single dimension of persistent homology. We therefore repeat the experiment from [ABC$^+$15] using landscape functions instead of β_0 curves.

In the TAaCGH approach the copy-number data from the chromosomes of patients is split up into contiguous segments of 20 probes each. The data from these 20 probe segments is then used to detect the significance of these segments, as reviewed in Section 4. In [ATGB$^+$16], Betti-1 curves are used to attempt to detect the significance of the $17q$ chromosome arm. We also consider entire chromosome arms using persistent landscapes to compare the effects of the size of the point clouds on detecting significant regions.

4.3.1. *TAaCGH with Persistence Landscapes Applied to Chromosome Arm Segments.* Recall from Section 2.1.4 that there are multiple persistent landscape functions, one for each natural number. The first persistent landscape on 0-dimensional homology will never detect a significant difference. Let k be the maximal filtration parameter considered. Then when $t < k/2$, the 0th persistence landscape function is the identity t and when $t > k/2$, it is $k-t$. For this reason, the first persistence landscape function is not considered. We use the 2nd, 3rd and 4th persistent landscape functions were used in a repeat of the experiment from [ABC$^+$15]. More persistent landscape functions were not considered because of the small size of the point clouds.

Significant regions in Table 4 were identified for each of the 4 cancer subtypes in the Horlings dataset. The landscape functions λ_1 and λ_2 detect significant regions for the Luminal A subtype,

	Luminal A	Luminal B	HER2+	Basal
λ_2	$5p3$		$17q1-2$	
λ_3	$2q2, 5p3, 11q6$		$17q1-3$	
λ_4			$17q1-3$	$1p7, 1q1, 2p7, 3p1, 4p3, 4q6-7,$ $4q10-11, 5p1, 10p1, 10p3, 10q5-6,$ $12p2, 13q2, 13q8, 14q8, 23p1, 23q7$

TABLE 4. Significant regions detected by persistent landscape functions on 0-dimensional homology. Red indicates the segments that we have detected that were not detected in [ABC$^+$15].

but no significant regions for the Basal subtype. This indicates that the topological differences between the Luminal A patients and the rest are governed b-y larger length generators, whereas the difference between basal and the other subtypes are governed by smaller generators. No significant regions were detected by λ_2, λ_3, or λ_4 for Luminal B. This is the same as for ℓ_0 and the original β_0 indicating that 0-dimensional homology is unlikely to detect differences between Luminal B and the other subtypes.

The study was also repeated on dimension 1 persistent homology using the first persistent landscape function. The results are in Table 5. As opposed to 0-dimensional persistent landscape

	Luminal A	Luminal B	HER2+	Basal
λ_1		$1q4, 2q11, 8q1$		$1p2, 1p11, 4q4, 4q7, 5p3, 6p4, 7q7$ $10p1, 10q1, 10q6, 13q3$

TABLE 5. Significant arms in Horlings data [HLN$^+$10] detected by the first persistent landscape function on 1d persistent homology. Red indicates the segments were not detected in [ABC$^+$15].

functions, λ_1 on 1-d persistent homology did detect significant regions for Luminal B. It also detected significant regions for the basal subtype, most of which were initially detected by β_0 in [ABC$^+$15]. The only significant region for the basal subtype that was not detected previously is $7q7$. On the other hand, λ_1 does not detect any significant regions for Luminal A or HER2+. This is similar to the results for HER2+ in [ATGB$^+$16] using β_1 curves, where only $17q2$ and $17q3$ were detected as significant from the Horlings dataset.

4.3.2. *TAaCGH with Persistence Landscapes Applied to Chromosome Arms.* In [ABC+15], each chromosome arm was split into segments and the data from each segment was then used to detect significant segments. In this section we apply this approach combined with persistent landscape functions on the data from entire chromosome arms to detect significant arms. In this case we are applying the sliding window algorithm to longer sequences and therefore we obtain larger point clouds. The results are contained in Table 6. Table 6 shows that there is a very strong signal for

	Luminal A	Luminal B	HER2+	Basal
λ_2			$17q$	
λ_3		$8p$	$17q$	
λ_4		$8p$	$17q$	$10p, 13q$

TABLE 6. Significant arms detected by persistent landscape functions.

the $17q$ arm. The second, third and fourth persistent landscape functions all detect it. For Luminal B the arm $8p$ is also detected by the third and fourth landscape functions. Significant arms $10p$ and $13q$ are both detected by λ_4 for basal. This is similar to the results on segments where only λ_4 detected significant segments for basal.

5. Comparing Betti Curves, Lifespan Curves and Persistence Landscapes on Simulated Data

Using the simulated data described in Section 2.2.2 that models test patients with a single contiguous chromosome aberration and control patients without one, we compare the ability of β_0, ℓ_0 and persistence landscapes to detect segments significant to patients. This is done by comparing the sensitivity of the three methods as the mean $\mu \in \{-1, 0.6, 1\}$, standard deviation $\sigma \in \{0.2, 0.5\}$ and length $\lambda \in \{2, 3, 5, 10, 20, 50, 75\}$ of the aberration varies. The sensitivities of ℓ_0 curves on the simulation data is in Table 7. The sensitivities of λ_2, the second persistent landscape function, on the simulation data is in Table 8. The sensitivities of λ_3 and λ_4, the third and fourth persistent landscape functions, on the simulation data is in Table 9. The sensitivity for β_0 curves on simulated

Mean	Stdd	Len	Sensitivity
-1, 1	0.2	2	0.95
-1, 1	0.2	3, 5, 10, 20, 50, 75	1
-1	0.5	2	0
-1	0.5	3	0.35
-1,1	0.5	5, 10, 20, 50, 75	1
0.6	0.2	2	0.15
0.6	0.2	3, 5, 10, 20, 50 , 75	1
0.6	0.5	2, 3	0
0.6	0.5	5	0.1
0.6	0.5	10	0.65
0.6	0.5	20	0.75
0.6	0.5	50	0.95
0.6	0.5	75	0.8
1	0.5	2	0.05
1	0.5	3	0.55

TABLE 7. Sensitivity of ℓ_0 curves for varying means, standard deviations and lengths of simulation. data

Mean	Stdd	Len	Sensitivity
-1	0.2	2, 3, 5, 10, 20, 50, 75	0
-1	0.2	2, 3, 5	0
-1	0.2	10	0.05
-1	0.2	20, 50, 75	0
0.6	0.2, 0.5	2, 3, 5, 10, 20, 50, 75	0
1	0.2	2, 3, 5, 10, 20, 50, 75	0
1	0.5	2, 3, 5, 10, 20, 50	0
1	0.5	75	0.05

TABLE 8. Sensitivity of λ_2 curves for varying means, standard deviations and lengths of simulation data.

Mean	Stdd	Len	Sensitivity
-1, 0.6, 1	0.2, 0.5	2, 3, 5, 10, 20, 50, 75	0

TABLE 9. Sensitivity of λ_3 and λ_4 curves for varying means, standard deviations and lengths of simulation data.

Mean	Stdd	Len	Sensitivity
-1, 0.6, 1	0.5	20, 50, 75	1
-1, 1	0.5	10	1
0.6	0.5	10	0.95
-1, 1	0.5	5	1
0.6	0.5	5	0.55
-1	0.5	3	0.90
1	0.5	3	0.95
0.6	0.5	3	0.15
-1	0.5	2	0.50
1	0.5	2	0.56
0.6	0.5	2	0.18

TABLE 10. Sensitivity of β_0 curves for varying means, lengths of simulation data from [ABC$^+$15].

data with $\mu \in \{-1, 0.6, 1\}$, $\sigma = 0.5$ and $\lambda \in \{2, 3, 5, 10, 20, 50, 75\}$ was previously studied in [ABC$^+$15]and the results are in Table 10.

5.1. **Simulations.** In the simulations run in [ABC$^+$15], the standard deviation was fixed at $\sigma = 0.5$ while the mean and length of the aberration in the test patients was allowed to vary. When the mean was -1 or 1 Betti-0 curves performed very well, achieving a sensitivity of 90% or more as long as the length of the aberration λ satisfied $\lambda \geq 3$. Even when the mean of the aberration was 0.5, the sensitivity of Betti-0 curves was 95% or more when $\lambda \geq 10$.

Lifespan-0 curves perform similarly to Betti curves on the simulated data. When the mean of the aberration is 1 or -1 then lifespan curves achieve sensitivity of 100% when $\lambda \geq 3$ and $\sigma = 0.2$. The lifespan curves also achieve 95% sensitivity when the mean is ± 1 and $\lambda = 2$. When the standard deviation is increased to $\sigma = 0.5$, but the mean is ± 1, lifespan curves achieve 100% specificity when $\lambda \geq 5$. When the mean $\mu = 0.6$ and the standard deviation $\sigma = 0.5$, lifespan curves do not perform as well as Betti-0 curves, not even achieving 100% specificity when $\lambda = 75$.

Persistence landscapes perform very poorly overall on the simulations. Even for long aberrations with little noise, the best sensitivity that any persistence landscape function achieves is 5%. This indicates that persistent landscape functions are likely to detect far less significant regions than Betti or lifespan curves.

Overall, the simulations indicate that Betti curves and lifespan curves will identify similar amounts of significant regions when noise is low and the mean of the aberration is far from 0. When the mean is closer to 0 and the noise increases, then Betti curves will outperform lifespan curves. Landscapes perform far worse than either lifespan curves or Betti curves under all variations in parameters. JA: This is not surprising since the bounds on the distance between Betti curves are larger than the bounds on the distance between lifespan curves. Landscape functions are known to be stable in general, so the bounds on the distance between landscape functions are even smaller. A larger distance between curves increases the potential to find statistical significance.

6. Discussion

When lifespan curves were applied to the Horlings dataset they performed similarly to how they performed in the simulations. They identified less regions overall than β_0 in [ABC$^+$15], particularly for the basal phenotype. Most of the regions that were detected were detected in the original β_0 study. There were 5 segments, that ℓ_0 detected that β_0 did not detect. Four of these segments were from the Luminal A phenotype. This confirms the fact that the lifespan curves can contain potentially different information from Betti curves.

Lifespan curves in dimension 1, ℓ_1, did not detect any significant segments for Luminal A nor Luminal B, but they did detect $17q2$ and nine segments for Basal. All of these segments were previously detected in the [ABC$^+$15] study using β_0. This indicates that, at least with this type of data, ℓ_1 may not contain any more information than 0-dimensional Betti curves.

When persistence landscape functions were applied to the Horlings data set, the significant regions discovered by them varied quite a bit between λ_2, λ_3 and λ_4. The significant regions detected by λ_2 and λ_3 were similar except that λ_3 detected $2q2$ and $11q6$ for Luminal A. The fouth persistent landscape function λ_4 did not detect anything for Luminal A, but detected many significant regions for the basal phenotype. Neither the first nor second persistent landscape functions detected anything for Basal.

In simulations, ℓ_0 curves perform very similarly to β_0 curves. Both have very high sensitivity when the standard deviation is low and the mean μ of the aberrations is far from 0. When the standard deviation is high, the length of the aberration has to be greater in order for ℓ_0 curves to have a high sensitivity. In simulations landscape curves performed the worst of the three types of curves. The second, third and fourth persistent landscape functions were only able to achieve 5% sensitivity at beast. This only occurred when the length of the aberration was long and the mean was far from 0. Overall, the simulations indicate that landscape functions will be far less effective at detecting single aberrations, regardless of length, mean and standard deviation.

References

[ABC$^+$15] Javier Arsuaga, Tyler Borrman, Raymond Cavalcante, Georgina Gonzalez, and Catherine Park. Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays*, 4(3):339–369, 2015.

[ABD$^+$12] Javier Arsuaga, Nils A Baas, Daniel DeWoskin, Hideaki Mizuno, Aleksandr Pankov, and Catherine Park. Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer. *Applicable Algebra in Engineering, Communication and Computing*, 23(1-2):3–15, 2012.

[AEK$^+$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. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

[AGDST18] Nieves Atienza, Rocío Gonzalez-Diaz, and M Soriano-Trigueros. A new entropy based summary function for topological data analysis. *Electronic Notes in Discrete Mathematics*, 68:113–118, 2018.

[ATGB+16] Sergio Ardanza-Trevijano, Georgina Gonzalez, Tyler Borrman, Juan Luis Garcia, and Javier Arsuaga. Topological analysis of amplicon structure in comparative genomic hybridization (cgh) data: an application to erbb2/her2/neu amplified tumors. In *International Workshop on Computational Topology in Image Context*, pages 113–129. Springer, 2016.

[AZHE10] Martin Aigner, Günter M Ziegler, Karl H Hofmann, and Paul Erdos. *Proofs from the Book*, volume 274. Springer, 2010.

[Bau19] Ulrich Bauer. Ripser: efficient computation of vietoris-rips persistence barcodes. *arXiv preprint arXiv:1908.02518*, 2019.

[Bub15] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

[Car09] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[CL19] Yu-Min Chung and Austin Lawson. Persistence curves: A canonical framework for summarizing persistence diagrams. *arXiv preprint arXiv:1904.07768*, 2019.

[CSEH07] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & computational geometry*, 37(1):103–120, 2007.

[DCCW+10] Daniel DeWoskin, Joan Climent, I Cruz-White, Mariel Vazquez, Catherine Park, and Javier Arsuaga. Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157(1):157–164, 2010.

[FKLM14] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*, 2014.

[Gof11] Michael Goff. Extremal betti numbers of vietoris–rips complexes. *Discrete & Computational Geometry*, 46(1):132–155, 2011.

[GUSA20] Georgina Gonzalez, Arina Ushakova, Radmila Sazdanovic, and Javier Arsuaga. Prediction in cancer genomics using topological signatures and machine learning. In *Topological Data Analysis*, pages 247–276. Springer, 2020.

[HG16] Gregory Henselman and Robert Ghrist. Matroid filtrations and computational persistent homology. *arXiv preprint arXiv:1606.00199*, 2016.

[HLN+10] Hugo M Horlings, Carmen Lai, Dimitry SA Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A Joosse, Christiaan Klijn, Petra M Nederlof, Marcel JT Reinders, et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clinical Cancer Research*, 16(2):651–663, 2010.

[MHBO] David Moon, John Harer, and Rann Bar-On. Maximum number of nonzero persistence cycles in a vietoris-rips filtration. Private Communication.

[Mor12] Dmitriy Morozov. Dionysus. *Software available at http://www. mrzv. org/software/dionysus*, 2012.

[Nan12] Vidit Nanda. Perseus: the persistent homology software. *Software available at http://www. sas. upenn. edu/˜ vnanda/perseus*, 2012.

[PH15] Jose A Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838, 2015.

[RHBK15] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

[SVJ11] H Sexton and Mikael Vejdemo-Johansson. Jplex simplicial complex library. 2011.

[The21] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.4.1 edition, 2021.

[XATZ20] Lu Xian, Henry Adams, Chad M Topaz, and Lori Ziegelmeier. Capturing dynamics of time-varying data via topology. *arXiv preprint arXiv:2010.05780*, 2020.