Tropical Sufficient Statistics for Persistent Homology[*]

Anthea Monod[†], Sara Kališnik[‡], Juan Ángel Patiño-Galindo[†], and Lorin Crawford[§]

Abstract. We show that an embedding in Euclidean space based on tropical geometry generates stable sufficient statistics for barcodes. In topological data analysis, barcodes are multiscale summaries of algebraic topological characteristics that capture the "shape" of data; however, in practice, they have complex structures that make them difficult to use in statistical settings. The sufficiency result presented in this work allows for classical probability distributions to be assumed on the tropical geometric representation of barcodes. This makes a variety of parametric statistical inference methods accessible to barcodes, all while maintaining their initial interpretations. More specifically, we show that exponential family distributions may be assumed and that likelihood functions for persistent homology may be constructed. We conceptually demonstrate sufficiency and illustrate its utility in persistent homology dimensions 0 and 1 with concrete parametric applications to human immunodeficiency virus and avian influenza data.

Key words. persistent homology, sufficient statistics, tropical geometry, infectious viral disease

AMS subject classifications. 54J05, 62B05, 14T05, 92D15

DOI. 10.1137/17M1148037

1. Introduction. In this paper, we provide statistical sufficiency in Euclidean space for persistent homology—an important concept in the field of topological data analysis (TDA) that summarizes the "shape" and "size" of data. Our result is based on a topological embedding given by functions defined in tropical geometry. With these sufficient summaries, probability distributions from classical statistics may be applied to persistent homology. More importantly, extensive statistical methodology and parametric inference have become accessible to TDA.

Recently, TDA has become particularly relevant due to its theoretical foundations that allow for dimensionality reduction with qualitative and robust summaries of observed data. To this end, it is applicable to a wide range of complex data structures arising in various domains of data science. Persistent homology is an important topological invariant upon

[†]Department of Systems Biology, Columbia University, New York, NY (am4691@cumc.columbia.edu, jp3770@cumc.columbia.edu).
[‡]Department of Mathematics, Wesleyan University, Middletown, CT (skalisnikver@wesleyan.edu).
[§]Department of Biostatistics, Center for Statistical Sciences, Center for Computational Molecular Biology, Brown University, Providence, RI (lorin_crawford@brown.edu).

which many TDA methods depend when applied in practice [91]. Specifically, this method has been used to address problems in fields ranging from sensor networks [2, 41], medicine [5, 36], neuroscience [23, 27, 42], as well as imaging analysis [72].

The output of persistent homology is a collection of intervals (known as barcodes, or persistence diagrams, in their representation as points on a plane) rather than numerical quantities. This has made it difficult to apply the method to parametric data analysis. Recently, there have been substantial advancements in applied topology to bypass this issue. More recent work entails developing statistical methodology and machine learning algorithms to work directly with barcodes and persistence diagrams. Examples of these include confidence sets separating topological noise from topological signal on a persistence diagram [34]; a positive-definite multiscale kernel for persistence diagrams [75]; and an input layer to deep neural networks to take in persistence diagrams directly [48]. An alternative approach entails manipulating barcodes so that they may be integrated into existing computational machinery, predominantly by vectorization, which is the focus of this paper. Various such techniques have been developed with specific properties depending on the goal of the study. A notable example is the persistence landscape, which assigns functions directly to barcodes and produces elements in Banach space [15]. The persistence landscape was then generalized to the persistence silhouette in [21], and to the kernel-smoothed persistence intensity function in [22], for statistical analyses such as clustering and weak convergence of bootstrapped persistence summaries. Additionally, feature representations are constructed in [19] by rearranging entries of a distance matrix between points in a persistence diagram. Another approach proposed in [11] uses binning to obtain a vector representation of features (though there are concerns of stability). Similar methods where stable surfaces of persistence diagrams are constructed, based on kernel methods, with the aim of incorporating persistence information into support vector machines and other machine learning algorithms are studied in [75], [57], and [3]. Persistence information may also be integrated as vectors into functional regression models via topological summary statistics based on Euler characteristics [25].

Summary statistics are relevant to the notion of sufficiency, which is a desirable property in studies centered on inference. Sufficiency allows a given sample of data to be mapped to a less complex or lower dimensional space (i.e., harboring less computational burden) without the loss of information [9]. Moreover, sufficient statistics provide the functional form of probability distributions via a classical factorization criterion. In other words, they represent the key basis for complete parametric inference in statistics. Summary statistics have been explored in the context of persistent homology to study shapes and surfaces (closed, compact subsets) in [84]: specifically, persistence diagrams were used to summarize and compare heel bones of various species of primates. Their construction provided a notion of summary statistics for the family of probability distributions on shapes and surfaces in 2 and 3 dimensions, in the space of persistence diagrams. Perhaps more importantly, this effort also shed light on the possibility of proving statistical sufficiency in the context of topological analyses of data.

The main contribution of our work is to provide statistical sufficiency for the general family of probability measures on persistent homology. A fundamental observation that makes this result possible is the formal probabilistic characterization of both the domain (i.e., the space of barcodes) and codomain (i.e., Euclidean space) of our mapping [63]. In principle, inference should proceed from a probability model defined directly on the barcodes. However,

this type of specification is complicated due to the complex geometry of the set of all barcodes. Given sufficient statistics for barcodes, we may use the likelihood principle to proceed with parametric inferences directly. As has been recently demonstrated, this suggests that a generative or sampling model on barcodes is possible in persistent homology [7].

The remainder of this paper is organized as follows. In section 2, we give a formal definition of persistent homology and outline the approach to map persistence barcodes to vectors using functions defined via tropical geometry. We also discuss properties of these functions. In section 3, we provide our main result of sufficient statistics for persistence barcodes. Based on this result, we give the form of an exponential family of probability distributions and likelihood functions for persistent homology. Section 4 gives two concrete applications to data in dimensions 0 and 1. The first is a practical demonstration of sufficiency based on human immunodeficiency virus (HIV) data, while the second shows how a parametric assumption statistically quantifies the biological distinction between intra- and intersubtype reassortment in avian influenza. We end the paper with a discussion in section 5 with directions for future research.

2. Persistent homology and vectorization. In this section, we give the mathematical background to our main result. We provide details on the space of barcodes arising from persistent homology and describe the construction of tropical coordinates on the space of persistence barcodes.

2.1. The space of persistence barcodes. Homology groups were developed in classical topology to "measure" the shape of spaces by abstractly counting the occurrences of patterns, such as the number of connected components, loops, and voids in 3 dimensions. The primary data analytic utility of *persistent homology* or *persistence* [32, 39, 91]—an important concept from TDA that is heavily leveraged in many data applications—is the adaptation of classical homology to point clouds (i.e., finite metric spaces), which is often the form in which data are collected. Persistence, as a topological construct, produces a robust summary of the data that is invariant to noisy perturbations [24]. Moreover, it also captures integral geometric features ("size") of the data. We now describe the construction of homology and its extension to persistent homology.

Simplicial homology and persistence. Simplicial homology studies the shape of simplicial complexes, which can be seen as skeletal representations of data types. An example of a simplicial complex is a mesh, which can be viewed as a discretization of an image. Simplicial complexes are an important building block in computing persistent homology: an abstract *simplicial complex* is a collection K of nonempty subsets of a set K_0 such that $\tau \subset \sigma$ and $\sigma \in K$ guarantees that $\tau \in K$. The elements of K_0 are called *vertices* of K, and the elements σ of K are called *simplices*. A simplex σ has dimension k, or is a *k-simplex*, if it has cardinality $k+1$.

Simplicial homology is constructed by considering simplicial *k-chains*, which are linear combinations of k-simplices in finite K over a field $\mathbb{F}$. A set of k-chains defines a vector space $C_k(K)$. When $\mathbb{F}$ is the binary field $\mathbb{Z}/2\mathbb{Z}$, the *boundary map* is

$$\partial_k : C_k(K) \to C_{k-1}(K),$$

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

with linear extension, where v_{-i} indicates that the ith element has been dropped. *Boundaries* are elements of $B_k(K) = \text{im } \partial_{k+1}$, and *cycles* are elements of $Z_k(K) = \ker \partial_k$.

Definition 2.1. *The kth homology group of K is given by the quotient group*

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

The motivating idea underlying homology is to account for the structure of K, which is a finite simplicial complex representation of a topological space X. In dimension 0, elements representing connected components of X with finite simplicial complex representation K generate the zeroth homology group $H_0(X)$. If, for example, X has two connected components, then $H_0(X) \cong \mathbb{F} \oplus \mathbb{F}$ (where $\cong$ denotes group isomorphism, and $\oplus$ here denotes the direct sum). For higher dimensions $k \geq 1$, k-dimensional holes are the result of considering the boundary of a $(k+1)$-dimensional object. $H_k(X)$ is generated by elements that represent k-dimensional holes in X. The rank of the kth homology group is denoted by β_k and referred to as the kth *Betti number*.

Intuitively, persistent homology is computed by tracking the progression of connected components, loops, and higher dimensional voids, with respect to a *filtration* assigned to the observed point cloud. A filtration is a finite nested sequence of simplicial complexes $\mathcal{K} := \{K_r\}_{r=0}^t$ indexed by a parameter $r \in \mathbb{R}^+$ such that $K_{r_1} \subseteq K_{r_2}$ if $r_1 < r_2$.

Definition 2.2. *Let K be a filtered simplicial complex $K_0 \subset \cdots \subset K_t = K$. The kth persistence module derived in homology, or simply kth persistent homology, of K is given by*

$$\text{PH}_k(K) := \{H_k(K_r)\}_{0 \leq r \leq t}$$

together with the collection of linear maps $\{\varphi_{r,s}\}_{0 \leq r \leq s \leq t}$, where $\varphi_{r,s} : H_k(K_r) \to H_k(K_s)$ is induced by the inclusion maps of $K_r \hookrightarrow K_s$ for all $r, s \in [0, t]$ such that $r \leq s$.

Persistent homology contains the homology information on individual spaces $\{K_r\}$ and on the mappings between their homologies for every K_r and K_s, where $r \leq s$. Note that the motivating idea underlying persistent homology is to account for the homology of X simultaneously across multiple scales. Rather than restricting the analysis to one static instance, persistence tracks how the topological structure evolves over the indexed filtration. The final output after computing persistence is a *barcode* (i.e., a collection of intervals). Each interval (or bar) corresponds to a topological feature that appears (i.e., birth) at the value of a parameter given by the left endpoint of the interval, and disappears or merges with another existing feature (i.e., death) at the value given by the right endpoint. Figure 1 provides an illustrative example of persistent homology.

Barcodes and persistence diagrams. For a given dimension k, a barcode consisting of n intervals, obtained from computing the kth persistent homology, can be written as a collection $(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)$, where x_i is the left endpoint of the ith interval and d_i is its length. In this paper, we assume that $x_i \geq 0$, thus meaning that the birth times of bars are nonnegative. Notice that the order in which coordinates (x_i, d_i) appear within the collection $(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)$ does not affect the content of the topological information encoded in the barcode. Hence, the indices $i = 1, 2, \ldots, n$ are essentially dummy variables. This means that all permutations of the coordinates (x_i, d_i) across n positions define the same collection.
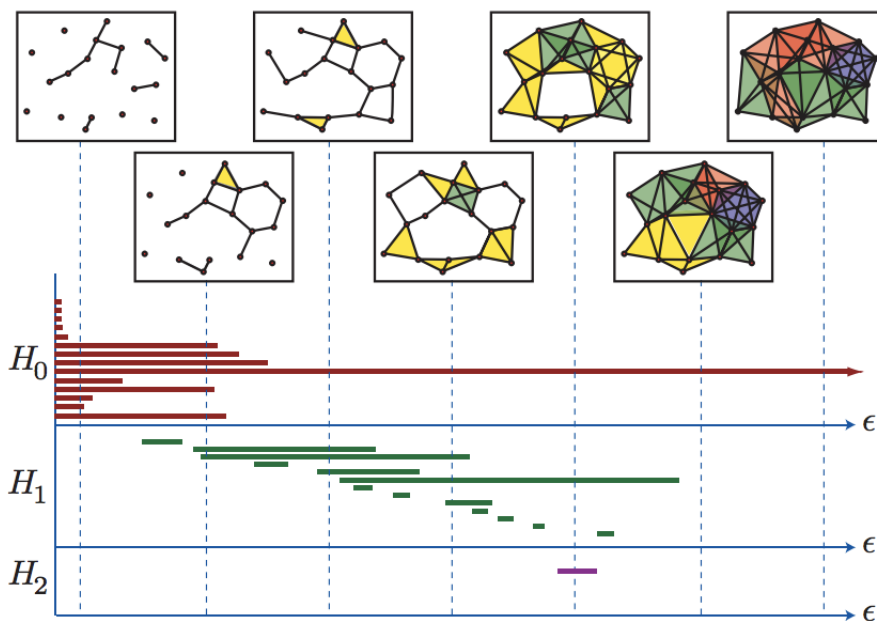
Figure 1. *Example illustrating persistent homology. This figure originally appears in* [40]. *Here,* 17 *points are sampled from an annulus with topology given by* $\beta_0 = 1$, $\beta_1 = 1$, *and* $\beta_2 = 0$. *Simplicial complexes are computed for continuous values of the filtration parameter,* $\epsilon \in [0, \infty)$. *The filtration is illustrated in the upper panels as the evolution of a simplicial complex in terms of vertices, edges, and faces that are formed as the value of* ϵ *increases.* 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 at instances of the filtration connect to the bars representing the topological features that exist at the corresponding values of* ϵ. *As* ϵ *progresses, connected components merge, cycles form and then fill up. The barcode summarizes this progression of* ϵ *by tracking the "lifetimes" of all topological features according to their homology groups. Notice that there is a single* H_0 *bar that persists as* $\epsilon \to \infty$, *representing the single connected component of the annulus. There are 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 representing the cycle bounded by faces in the corresponding panel, but the length of the bar is comparatively short and likely to be a spurious topological artefact.*

Algebraically, this amounts to taking the orbit space of the action of the symmetric group on n letters on the product $\big([0, \infty) \times [0, \infty)\big)^n$, given by permuting the coordinates. We denote this orbit space by B_n.

Definition 2.3. *The barcode space* $\mathcal{B}_{\leq n}$ *consisting of bars with at most* n *intervals is given by the quotient*

$$\coprod_{n \in \mathbf{N}_{\leq n}} B_n / \sim,$$

where $\sim$ *is generated by equivalences of the form*

$$\big\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\big\} \sim \big\{(x_1, d_1), (x_2, d_2), \ldots, (x_{n-1}, d_{n-1})\big\},$$

whenever $d_n = 0$.

We may set $y_i := x_i + d_i$ to be the right endpoint of the ith interval, and equivalently consider the collection of ordered pairs (x_i, y_i) together with the diagonal $\Delta = \{(x, y) \mid x = y\}$, where each point is taken with infinite multiplicity. Here, Δ contains the bars of length zero (i.e., features that are born and die at exactly the same time and therefore have zero persistence). Plotting these ordered pairs on a scatterplot, we obtain an alternate representation of the barcode, known as a *persistence diagram*.

Metrics on barcode space. The space of barcodes can be equipped with various measures of distance, which, under mild regularity conditions (namely, local finiteness of barcodes or persistence diagrams), are also metrics. In this paper, we consider the bottleneck and Wasserstein distances, which are the distance functions used in many existing results in the field of TDA. In particular, the stability theorem [24], which provides a foundation that makes persistent homology applicable to real data analysis, is stated with respect to the bottleneck distance. Unless explicitly stated otherwise, all results in this paper hold in generality for the Wasserstein p-distance for all $p \geq 1$, as well as the bottleneck distance.

The first step in specifying the distance between any two barcodes is to define the distance between any pair of intervals, as well as the distance between any interval and the set of zero length intervals Δ. We take

$$\mathrm{d}_\infty\big((x_i, d_i), (x_j, d_j)\big) = \max\big\{|x_i - x_j|,\, |d_i - d_j + x_i - x_j|\big\}$$

to be the distance between barcodes $\{(x_i, d_i)\}$ and $\{(x_j, d_j)\}$. The distance between an interval (x, d) and the set Δ is

$$\mathrm{d}_\infty\big((x, d), \Delta\big) = \frac{d}{2},$$

which is simply the minimal d_∞ distance between $(x, x+d)$ and Δ. Now let $\mathscr{B}_1 = \{I_\alpha\}_{\alpha \in A}$ and $\mathscr{B}_2 = \{J_\beta\}_{\beta \in B}$ be two barcodes.

Definition 2.4. *For a fixed p, finite sets A and B, any bijection γ from a subset $A' \subseteq A$ to $B' \subseteq B$, and a penalty on γ set to*

$$P_p(\gamma) := \sum_{a \in A'} d_\infty\big(I_a, J_{\gamma(a)}\big)^p + \sum_{a \in A \setminus A'} d_\infty\big(I_a, \Delta\big)^p + \sum_{b \in B \setminus B'} d_\infty\big(I_b, \Delta\big)^p,$$

the Wasserstein p-distance *($p \geq 1$) between $\mathscr{B}_1$ and $\mathscr{B}_2$ is given by*

$$d_p^W(\mathscr{B}_1, \mathscr{B}_2) := \Big(\min_\gamma P_p(\gamma)\Big)^{\frac{1}{p}}.$$

Definition 2.5. *For finite subsets and sets $A' \subseteq A$ and $B' \subseteq B$, a bijection γ as previously defined above, and a penalty on γ set to*

$$P_\infty(\gamma) := \max\left\{\max_{a \in A'}\Big(d_\infty\big(I_a, J_{\gamma(a)}\big)\Big),\ \max_{a \in A \setminus A'} d_\infty(I_a, \Delta),\ \max_{b \in B \setminus B'} d_\infty(I_b, \Delta)\right\},$$

the quantity

$$d_\infty^B(\mathscr{B}_1, \mathscr{B}_2) := \min_\gamma P_\infty(\gamma),$$

defines the bottleneck distance *between $\mathscr{B}_1$ and $\mathscr{B}_2$, where the minimum is again taken over all bijections from subsets A' to B'.*

Barcodes and data. In applications, data are generated by random processes and thus their representation as simplicial complexes and their resulting barcodes are also considered to be random objects. Barcodes can therefore be seen as summary statistics of the data generating process because they allow for a dimensionality reduction of the ambient space. Therefore, in a broader and more philosophical sense, topological approaches in data science applications can be used to provide sets of principal variables for scalable inference. In practice, however, estimating persistent homology is subject to the curse of dimensionality, since it amounts to statistically estimating pairwise correspondences of sets in Hausdorff distance.

2.2. Coordinates on the space of persistence barcodes. To integrate shape information of data (via barcodes) into existing computational machinery and methodology, several vectorization techniques have been proposed. The vectorization that [51] proposes takes the form of functions, which we build upon and extend to statistical theory in this paper.

For the goal of relating persistent homology to statistical theory, studying functional vectorization techniques is preferable to studying algorithmic vectorization, because specific properties relevant to statistical inference, such as injectivity and measurability, are assessed on functions rather than algorithms. In addition to the vectorization of [51], there are other vectorization approaches based on the identification of polynomials or functions. For example, complex polynomials (where the points of a persistence diagram are the roots) have been developed in [35]; however, these are difficult to apply to parametric statistical analysis where data are primarily real-valued. Persistence landscapes in [15] are also less natural for statistical inference due to certain characteristics of Banach space (for example, the fact that norms on Banach space need not comprise an inner product structure).

Another particularly relevant approach by [4] assigns vectors directly to barcodes by defining a ring of algebraic functions. These functions are polynomials and satisfy desirable properties such as symmetry in the variables (i.e., invariance of ordering of the bars), and invariance under addition of 0-length intervals. Unfortunately, it turns out that these functions are not Lipschitz with respect to the Wasserstein p- and bottleneck distances (see Appendix A for a proof). This lack of Lipschitz continuity is problematic in real data applications that are generated by random processes because of its negative implications on stability in the target space under perturbations in the domain. More specifically, there are no guarantees that the transformation is computationally or statistically robust. As a viable solution to address such shortcomings of these formerly defined polynomials in [4], tropical functions on the space of barcodes were identified. These tropical functions possess the same desirable properties as those developed by [4], but additionally, were shown to be Lipschitz (and therefore continuous) with respect to the Wasserstein p-distance for $p \geq 1$ and the bottleneck distance [51].

We now briefly review the definition of the tropical semiring and the tropical functions that we study as coordinates on barcode space.

Fundamentals of tropical algebra. Tropical algebra is a branch of mathematics based on the study of the tropical semiring.

Definition 2.6. *The* tropical (equivalently, min-plus) semiring *is* $(\mathbb{R} \cup \{+\infty\}, \oplus, \odot)$, *with addition and multiplication being defined as*

$$a \oplus b := \min(a, b) \quad and \quad a \odot b := a + b,$$

where both operations are commutative and associative, and multiplication distributes over addition. Similarly, there exists the max-plus semiring, $(\mathbb{R} \cup \{-\infty\}, \boxplus, \odot)$, *where multiplication of two elements is defined as in the case of the tropical semiring, but addition amounts to taking the maximum instead of the minimum. Namely,*

$$a \boxplus b := \max(a, b) \quad and \quad a \odot b := a + b,$$

where again both operations are associative, commutative, and distributive as in the tropical semiring.

As in the case of ordinary polynomials that are formed by multiplying and adding real variables, max-plus polynomials can be formed in a similar fashion as follows. Let $x_1, x_2, \ldots, x_n$ be variables that are elements in the max-plus semiring. A *max-plus monomial expression* is any product of these variables, where repetition is permitted. By commutativity, we can sort the product and write monomial expressions with the variables raised to exponents:

$$p(x_1, x_2, \ldots, x_n) = a_1 \odot x_1^{a_1^1} x_2^{a_2^1} \cdots x_n^{a_n^1} \boxplus a_2 \odot x_1^{a_1^2} x_2^{a_2^2} \cdots x_n^{a_n^2} \boxplus \cdots \boxplus a_m \odot x_1^{a_1^m} x_2^{a_2^m} \cdots x_n^{a_n^m}$$
$$= \max(a_1 + a_1^1 x_1 + \cdots + a_n^1 x_n, \ldots, a_m + a_1^m x_1 + \cdots + a_n^m x_n).$$

Here, the coefficients $a_1, a_2, \ldots, a_m$ are in $\mathbb{R}$, and the exponents a_j^i are in $\mathbb{Z}^+$ for $1 \le j \le n$ and $1 \le i \le m$. Max-plus polynomial expressions do not uniquely define functions (e.g., [18]), which suggests an equivalence relation. For max-plus polynomial expressions p and q, if

$$p(x_1, x_2, \ldots, x_n) = q(x_1, x_2, \ldots, x_n)$$

for all $(x_1, x_2, \ldots, x_n) \in (\mathbb{R} \cup \{+\infty\})^n$, then p and q are said to be *functionally equivalent*, and we write $p \sim q$. Max-plus polynomials are the semiring of equivalence classes for max-plus polynomial expressions with respect to $\sim$.

Identifying tropical functions for barcodes. We now give a list of tropical functions that we use to assign vectors to barcodes, restate existing relevant results, and prove additional statements necessary for our main result (see Theorem 3.5 in section 3). For a complete discussion on the construction of the functions, and proofs related to the original definitions of the tropical functions, see the relevant previous works: [18] and [51].

Fix n and let the symmetric group S_n act on the following matrix of indeterminates by left multiplication:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}.$$

Here, the matrix X represents a barcode with n bars, where each bar is represented by a row. As previously mentioned, the action of S_n on X amounts to permuting the order of the bars within $(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)$. In parallel, also consider the collection of *exponent* matrices

$$\mathscr{E}_n = \left\{ \begin{pmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \\ \vdots & \vdots \\ e_{n,1} & e_{n,2} \end{pmatrix} \ne [0]_n^2 \mid e_{i,j} \in \{0,1\} \text{ for } i = 1, 2, \ldots, n, \text{ and } j = 1, 2 \right\}.$$

A matrix $E \in \mathscr{E}_n$, together with X, determines a max-plus monomial by

$$P(E) = x_{1,1}^{e_{1,1}} \odot x_{1,2}^{e_{1,2}} \odot \cdots \odot x_{n,1}^{e_{n,1}} \odot x_{n,2}^{e_{n,2}}.$$

Now, denote the set of orbits under the row permutation action on $\mathscr{E}_n$ by $\mathscr{E}_n/S_n$. Each orbit $\{E_1, E_2, \ldots, E_m\}$ determines an *elementary 2-symmetric* max-plus polynomial by

$$P(E_1) \boxplus P(E_2) \boxplus \cdots \boxplus P(E_m).$$

Let $E_{(e_{1,1},e_{1,2}),(e_{2,1},e_{2,2}),\ldots,(e_{n,1},e_{n,2})}$ denote the polynomial that arises from the orbit

$$\left[\begin{pmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \\ \vdots & \vdots \\ e_{n,1} & e_{n,2} \end{pmatrix} \right].$$

Example 2.7. Let $n = 3$. Here are a few examples of elementary 2-symmetric max-plus polynomials:

$$
\begin{aligned}
E_{(0,1),(0,1),(0,0)}(x_1, d_1, x_2, d_2, x_3, d_3) &= d_1 \odot d_2 \boxplus d_1 \odot d_3 \boxplus d_2 \odot d_3 \\
&= \max(d_1 + d_2,\, d_1 + d_3,\, d_2 + d_3), \\
E_{(0,1),(0,0),(0,1)}(x_1, d_1, x_2, d_2, x_3, d_3) &= d_1 \odot d_2 \boxplus d_1 \odot d_3 \boxplus d_2 \odot d_3 \\
&= \max(d_1 + d_2,\, d_1 + d_3,\, d_2 + d_3), \\
E_{(1,1),(1,1),(1,1)}(x_1, d_1, x_2, d_2, x_3, d_3) &= x_1 \odot d_1 \odot x_2 \odot d_2 \odot x_3 \odot d_3 \\
&= x_1 + d_1 + x_2 + d_2 + x_3 + d_3.
\end{aligned}
$$

In the example above, the first two polynomials are equivalent: $E_{(0,1),(0,1),(0,0)}$ is the same function as $E_{(0,1),(0,0),(0,1)}$, which in turn is the same as $E_{(0,0),(0,1),(0,1)}$. Thus, only n and the number of $(0,1)$ rows are needed to determine the 2-symmetric polynomial in this example. Since we know n, we may therefore simplify notation by writing this polynomial as $E_{(0,1)^2}$ instead of listing every row. This is the notational convention that we will adopt throughout the rest of the paper. Namely, whenever n is known, we write $E_{(e_{1,1},e_{1,2}),\ldots,(e_{n,1},e_{n,2})}$ as $E_{(1,0)^k,(0,1)^i,(1,1)^j}$—where among $(e_{1,1}, e_{1,2}), (e_{2,1}, e_{2,2}), \ldots, (e_{n,1}, e_{n,2})$, k is the number of $(1,0)$ rows, i the number of $(0,1)$ rows, and j the number of $(1,1)$ rows. The number of $(0,0)$ rows is what remains (i.e., $n - i - j - k$), which we omit from the notation. If we wish to emphasize the value of n, we write $E_{(1,0)^k,(0,1)^i,(1,1)^j}^n$.

These functions are the building blocks for the functions we use to vectorize the barcode space. Proposition 5.8 of [18] asserts that $E_{(0,1)^i,(1,1)^j,(1,0)^k}$ separate orbits under the row permutation action on $\mathbb{R}^{2n}$. In other words, if $\mathscr{B}_1 \neq \mathscr{B}_2$, then $\{i, j, k\}$ exist such that $E_{(0,1)^i,(1,1)^j,(1,0)^k}(\mathscr{B}_1) \neq E_{(0,1)^i,(1,1)^j,(1,0)^k}(\mathscr{B}_2)$. We improve this statement and show that the two-parameter family of $E_{(0,1)^i,(1,1)^j}$, parameterized by i and j alone, suffices to separate the barcodes, and that the k-factor of $(1,0)$ rows is not needed. This is an important result, as it enables us to embed the barcode space into a lower-dimensional Euclidean space, and hence tightens the main sufficiency result of this paper. We now formalize this fact in the following proposition.

Proposition 2.8. *Let* $\left[(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)\right]$ *and* $\left[(x'_1, d'_1, x'_2, d'_2, \ldots, x'_n, d'_n)\right]$ *be two orbits under the row permutation action on* $\mathbb{R}^{2n}$. *If*

$$E^n_{(0,1)^i,(1,1)^j}\left[(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)\right] = E^n_{(0,1)^i,(1,1)^j}\left[(x'_1, d'_1, x'_2, d'_2, \ldots, x'_n, d'_n)\right]$$

for all $i, j \leq n$, *then*

$$\left[(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)\right] = \left[(x'_1, d'_1, x'_2, d'_2, \ldots, x'_n, d'_n)\right].$$

In other words, $E^n_{(0,1)^i,(1,1)^j}$ *separates barcodes.*

Proof. We prove the statement by induction on the number of nonzero bars $\leq n$. If $n = 1$, then $E^1_{(0,1)}\left[(x_1, d_1)\right] = E^1_{(0,1)}\left[(x'_1, d'_1)\right]$ implies that $d_1 = d'_1$, and $(E^1_{(0,1)} - E^1_{(1,1)})\left[(x_1, d_1)\right] = (E^1_{(0,1)} - E^1_{(1,1)})\left[(x'_1, d'_1)\right]$ implies that $x_1 = x'_1$. From here it follows that $[(x_1, d_1)] = [(x'_1, d'_1)]$, and the base case holds.

Now suppose that $E^{n-1}_{(0,1)^i,(1,1)^j}$ for nonnegative integers i, j, where $i + j \leq n - 1$, separates barcodes with up to $(n-1)$ nonzero bars. We must show that $E^n_{(0,1)^i(1,1)^j}$ separates barcodes with up to n bars. Let $\mathscr{B}_1 = [(x_1, d_1, \ldots, x_n, d_n)]$ and $\mathscr{B}_2 = [(x'_1, d'_1, \ldots, x'_n, d'_n)]$ be two barcodes with up to n nonzero bars. Assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_n$ and $x'_1 \leq x'_2 \leq \cdots \leq x'_n$. We furthermore arrange that for all $i \leq n - 1$ with $x_i = x_{i+1}$, we have $d_i \leq d_{i+1}$.

If $E^n_{(0,1)}(\mathscr{B}_1) = E^n_{(0,1)}(\mathscr{B}_2)$, it follows that $\max_{1 \leq i \leq n} d_i = \max_{1 \leq i \leq n} d'_i$. Since

$$E^n_{(0,1)^2}(\mathscr{B}_1) - E^n_{(0,1)^1}(\mathscr{B}_1) = E^n_{(0,1)^2}(\mathscr{B}_2) - E^n_{(0,1)^1}(\mathscr{B}_2),$$

it follows that the second greatest element in $\{d_i \mid 1 \leq i \leq n\}$ is equal to the second greatest element in $\{d'_i \mid 1 \leq i \leq n\}$. We apply this reasoning to the remaining elements in decreasing order and finally obtain $\{d_i \mid 1 \leq i \leq n\} = \{d'_i \mid 1 \leq i \leq n\}$.

The fact that x_n equals x'_n follows by applying $E^n_{(0,1)^{n-1},(1,1)^1}$. Computing the difference

$$E^n_{(0,1)^{n-2},(1,1)^2} - E^n_{(0,1)^{n-1},(1,1)^1}$$

yields $x_{n-1} = x'_{n-1}$. We continue in this manner and finally arrive at the conclusion that $\{x_1, x_2, \ldots, x_n\}$ is a permutation of $\{x'_1, x'_2, \ldots, x'_n\}$. This means that unless

$$\left[(x_1, d_{\pi(1)}, \ldots, x_n, d_{\pi(n)})\right] = \left[(x_1, d_1, \ldots, x_n, d_n)\right]$$

for some permutation π, one of the functions among $E^n_{(0,1)^i,(1,1)^j}$ separates $\mathscr{B}_1$ and $\mathscr{B}_2$.

Assume now that we have such a π where $\left[(x_1, d_{\pi(1)}, \ldots, x_n, d_{\pi(n)})\right] = \left[(x_1, d_1, \ldots, x_n, d_n)\right]$. Let $I = \{k \in \{1, 2, \ldots, n\} \mid d_k > d_{\pi(n)}\}$. Using the fact that $x_1 \leq x_2 \leq \cdots \leq x_n$, we evaluate $E^n_{(0,1)^{|I|},(1,1)}$, where $|I|$ is the cardinality of I:

$$E^n_{(0,1)^{|I|}(1,1)}\left(\left[(x_1, d_{\pi(1)}, \ldots, x_n, d_{\pi(n)})\right]\right) = x_n + d_{\pi(n)} + \sum_{k \in I} d_k.$$

After splitting the index set according to membership in I, we arrive at

$$E^n_{(0,1)^{|I|},(1,1)}\left(\left[(x_1, d_1, \ldots, x_n, d_n)\right]\right) = \max\left(\max_{\ell \notin I}\left(x_\ell + \sum_{k \in I} d_k + d_\ell\right), \max_{\ell \in I}\left(x_\ell + \sum_{k \in I} d_k + d_{\pi(n)}\right)\right).$$

So if $d_n < d_{\pi(n)}$, then also $d_{n-i}, d_{n-i+1}, \ldots < d_{\pi(n)}$, which implies that $n-i, n-i+1, \ldots, n \notin I$. Thus if $\ell \notin I$, then there are two cases. First, $\ell < n-i$, in which case $x_\ell + d_\ell \leq x_{n-i-1} + d_{\pi(n)} < x_n + d_{\pi(n)}$, giving the inequality

$$E^n_{(0,1)^{|I|},(1,1)}\big([(x_1, d_1, \ldots, x_n, d_n)]\big) < E^n_{(0,1)^{|I|},(1,1)}\big([(x_1, d_{\pi(1)}, \ldots, x_n, d_{\pi(n)})]\big).$$

Second, if $n - i \leq \ell \leq n$, then $x_\ell + d_\ell = x_n + d_\ell < x_n + d_{\pi(n)}$ again. As for the second max, if $\ell \in I$, then necessarily $\ell < n - i$, so $x_\ell + d_{\pi(n)} < x_n + d_{\pi(n)}$. A symmetrical argument shows that one can distinguish between the barcodes with $E^n_{(0,1)^{|I|},(1,1)}$, where $I = \{k \in \{1, 2, \ldots, n\} \mid d_k > d_n\}$ if $d_n > d_{\pi(n)}$. So unless the maximal x values are not matched with the same d values in $\mathscr{B}_1$ and $\mathscr{B}_2$, we can distinguish the barcodes using elementary 2-symmetric polynomials.

Now suppose that $x'_n = x_n$ and $d'_n = d_n$. We show using the inductive hypothesis that if

$$\mathscr{B}_1 \setminus \{(x_n, d_n)\} \neq \mathscr{B}_2 \setminus \{(x_n, d_n)\},$$

then we can distinguish $\mathscr{B}_1$ and $\mathscr{B}_2$ by elementary 2-symmetric polynomials in $2n$ variables, which we construct from elementary 2-symmetric polynomials in $2(n-1)$ variables. By symmetry between $\mathscr{B}_1$ and $\mathscr{B}_2$, we may assume that there exist i and j such that

$$E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_1 \setminus \{(x_n, d_n)\}\big) > E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big).$$

Note that $E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_k \setminus \{(x_n, d_n)\}) \leq E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_k)$ for $k = 1, 2$.

If $E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_1) \neq E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2)$, we are done, since we have a function that separates the two barcodes. If $E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}) = E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2)$, then

$$E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_1) > E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2)$$

and thus $E^n_{(0,1)^i,(1,1)^j}$ distinguishes $\mathscr{B}_1$ and $\mathscr{B}_2$. It then suffices to study the situation when

$$(2.1) \qquad E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big) < E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2).$$

In this situation $E^n_{(0,1)^i,(1,1)^j}$ must depend on at least d_n or x_n (or possibly both). Then in the partition I, J such that

$$(2.2) \qquad E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2) = \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i,$$

either $n \in I$ or $n \in J$.

First we suppose that $n \in J$. We prove that then

$$E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big).$$

A short consideration of the fact that we have three possibilities (either x_n plays a role in the sum that equals $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2)$; or $x_n + d_n$; or neither appears in the sum) gives the following fact:

$$E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = \max\Big(x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}),$$
$$d_n + E^{n-1}_{(0,1)^{i-1},(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}),$$
$$E^{n-1}_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\})\Big).$$

We must show that this maximum cannot be attained in the second or third expression. We prove this statement by contradiction.

- Assume that $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = E^{n-1}_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\})$. This means that I_1, J_1 exist such that

(2.3)
$$E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i$$

and $n \notin I_1 \cup J_1$. Also, s exists such that $s \in J_1 \cup I_1 \setminus J \cup I$.

Assume that $s \in J_1$. If $x_s + d_s > x_n + d_n$, then

$$\sum_{j \in J \cup \{s\} \setminus \{n\}}(x_j + d_j) + \sum_{i \in I} d_i > \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i,$$

which contradicts the definition of $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2)$. If $x_s + d_s = x_n + d_n$, then

$$\sum_{j \in J \cup \{s\} \setminus \{n\}}(x_j + d_j) + \sum_{i \in I} d_i = \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i = E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2).$$

The sum on the left does not involve n, which would imply that

$$E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2) = E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}),$$

contradicting (2.1). Now suppose $x_s + d_s < x_n + d_n$. Then

$$\sum_{j \in J_1 \cup \{n\} \setminus \{s\}}(x_j + d_j) + \sum_{i \in I_1} d_i > \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i,$$

which contradicts (2.3).

Now assume that $s \in I_1$. Since $s \in I_1$ and not in I, and since I and I_1 have the same cardinality, there exists $t \in I \setminus I_1$. If $d_s = d_t$, we replace t with s ($s \notin I \cup J$, so this replacement is possible) and start over with I replaced by $I \cup \{s\} \setminus \{t\}$, beginning with (2.2). The key observation is that this replacement reduces the number of elements in $I_1 \setminus I$ by 1. Thus, after at most $|I|$ such replacements, we may either assume that $I = I_1$, in which case our element $s \in J_1 \cup I_1 \setminus J \cup I$ necessarily lies in J_1 and our previous arguments prove the assertion; or we reach the situation that $I \neq I_1$, and there exist $s \in I_1 \setminus (I \cup J)$ and $t \in I \setminus I_1$ with $d_s \neq d_t$.

If $d_s > d_t$, we arrive at a contradiction to (2.2), since

$$\sum_{j \in J}(x_j + d_j) + \sum_{i \in I \cup \{s\} \setminus \{t\}} d_i > \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i.$$

Now suppose that $d_s < d_t$. If $t \notin J_1$, then we replace s by t and arrive at a contradiction, since

$$\sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1 \cup \{t\} \setminus \{s\}} d_i > \sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1} d_i.$$

We also must have that

$$(2.4) \qquad (x_n + d_n) + d_t \geq (x_t + d_t) + d_s.$$

Otherwise,

$$\sum_{j \in J \cup \{t\} \setminus \{n\}} (x_j + d_j) + \sum_{i \in I \cup \{s\} \setminus \{t\}} d_i > \sum_{j \in J} (x_j + d_j) + \sum_{i \in I} d_i,$$

which leads to a contradiction to (2.2). If $t \in J_1$, we use (2.4) to obtain

$$\sum_{j \in J_1 \cup \{n\} \setminus \{t\}} (x_j + d_j) + \sum_{i \in I_1 \cup \{t\} \setminus \{s\}} d_i \geq \sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1} d_i.$$

Now we can replace t with n to obtain

$$E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}).$$

• Assume that $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = d_n + E^{n-1}_{(0,1)^{i-1},(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\})$. This means that I_1, J_1 exist such that

$$(2.5) \qquad E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = \sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1} d_i + d_n$$

and $n \notin I_1 \cup J_1$. Take any $s \in J_1$. Since $x_n \geq x_s$, we have $(x_n + d_n) + x_s \geq (x_s + d_s) + x_n$ and we may switch s and n to get

$$\sum_{j \in J_1 \cup \{n\} \setminus \{s\}} (x_j + d_j) + \sum_{i \in I_1} d_i + d_s \geq \sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1} d_i + d_n,$$

and because the right hand side is maximal,

$$\sum_{j \in J_1 \cup \{n\} \setminus \{s\}} (x_j + d_j) + \sum_{i \in I_1} d_i + d_s = \sum_{j \in J_1} (x_j + d_j) + \sum_{i \in I_1} d_i + d_n,$$

and consequently

$$d_n + E^{n-1}_{(0,1)^{i-1},(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}) = d_n + x_n + E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}).$$

So if $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = d_n + E^{n-1}_{(0,1)^{i-1},(1,1)^{j+1}}(\mathscr{B}_2 \setminus \{(x_n, d_n)\})$, then also

$$E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = d_n + x_n + E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}).$$

We have just proved that $E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\})$. We use this fact in the first line (2.6) of the following computation. The second line (2.7) uses the inductive hypothesis and the last line (2.8) follows from the definition of the functions we use:

$$(2.6) \qquad E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_2) = x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big),$$

$$(2.7) \qquad\qquad\qquad < x_n + d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_1 \setminus \{(x_n, d_n)\}\big),$$

$$(2.8) \qquad\qquad\qquad \leq E^n_{(0,1)^i,(1,1)^{j+1}}(\mathscr{B}_1).$$

Now suppose that $n \in I$. Analogously to the previous example, we will first prove that

$$E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big).$$

Once again, we have $E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2)$ is the maximum over the following three expressions:

$$E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = \max\Big(x_n + d_n + E^{n-1}_{(0,1)^{i+1},(1,1)^{j-1}}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big),$$
$$d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big),$$
$$E^{n-1}_{(0,1)^{i+1},(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big)\Big).$$

We must show that this maximum cannot be attained in the first or third expression. We prove this statement by contradiction.

Recall the following expression from (2.2):

$$E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2) = \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i.$$

Here $n \in I$. By our current assumption, we can write

$$E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i,$$

where $n \in J_1$. Note that this only makes sense if $J_1 \neq \emptyset$ (otherwise $x_n + d_n$ will not appear in the sum). Since $n \in J_1 \setminus J$, we can pick $t \in J \setminus J_1$. Since $(x_t + d_t) + d_n \leq (x_n + d_n) + d_t$, we can rewrite $\sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i$ as $\sum_{j \in J \cup \{n\} \setminus \{t\}}(x_j + d_j) + \sum_{i \in I \cup \{t\} \setminus \{n\}} d_i$ and we would be in the situation which we previously analyzed.

• Assume that $E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = E^{n-1}_{(0,1)^{i+1},(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big)$. Recall again the following expression from (2.2):

$$E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2) = \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i,$$

where $n \in I$. Additionally, we can write $E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i$, with $n \notin J_1 \cup I_1$. Also, s exists such that $s \in J_1 \cup I_1 \setminus J \cup I$.

If $d_s > d_n$, then

$$\sum_{j \in J}(x_j + d_j) + \sum_{i \in I \cup \{s\} \setminus \{n\}} d_i > \sum_{j \in J}(x_j + d_j) + \sum_{i \in I} d_i,$$

which is a contradiction, since the expression on the right is maximal. If $d_s = d_n$, we may simply replace n by s in the sum for $E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2)$ and get that $E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2)$ does not depend on (x_n, d_n) and therefore

$$E^n_{(0,1)^i,(1,1)^j}(\mathscr{B}_2) = E^{n-1}_{(0,1)^i,(1,1)^j}(\mathscr{B}_2 \setminus \{(x_n, d_n)\}),$$

which contradicts (2.1). Finally, assume that $d_s < d_n$. If $s \in I_1$, then

$$\sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1 \cup \{n\} \setminus \{s\}} d_i > \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i,$$

which is once again a contradiction. Now assume that $s \in J_1$. Since $x_s \leq x_n$, we have $x_s + d_s < x_n + d_n$ and therefore

$$\sum_{j \in J_1 \cup \{n\} \setminus \{s\}}(x_j + d_j) + \sum_{i \in I_1} d_i > \sum_{j \in J_1}(x_j + d_j) + \sum_{i \in I_1} d_i,$$

which is a contradiction.

We conclude that

$$E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_2) = d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_2 \setminus \{(x_n, d_n)\}\big)$$
$$< d_n + E^{n-1}_{(0,1)^i,(1,1)^j}\big(\mathscr{B}_1 \setminus \{(x_n, d_n)\}\big)$$
$$\leq E^n_{(0,1)^{i+1},(1,1)^j}(\mathscr{B}_1).$$

We have shown that if $\mathscr{B}_1 \neq \mathscr{B}_2$, we can find elementary 2-symmetric max-plus polynomials that distinguish $\mathscr{B}_1$ and $\mathscr{B}_2$. This completes the induction and the proof. ∎

Regularizing subsets of barcode space. In this paper, we consider subsets of barcode space defined as follows. For a fixed $m > 0$, denote $\mathcal{B}^m_{\leq n}$ to be the subset of $\mathcal{B}_{\leq n}$ consisting of those $(x_1, d_1, x_2, d_2, \ldots, x_k, d_k)$ with $k \leq n$ and $d_i > 0$ such that

$$(2.9) \qquad\qquad\qquad\qquad x_i \leq md_i$$

for $i = 1, 2, \ldots, k$. Given a finite set of barcodes, the value of m is straightforward to determine. While it might appear restrictive at first to consider only a subset of the barcode space, it does not actually pose any limitations in practice. Indeed, for a finite data set generated by some finite process (e.g., meshes have a finite number of vertices/faces, images have limited resolution, etc.), there will only be finitely many bars, each with finite length, in the resulting barcode.

In [51], an infinite list of functions required to distinguish any two barcodes is given. Every function in this list is stable with respect to both the Wasserstein p- (for $p \geq 1$) and bottleneck distances. However, in application settings, it is not possible (and quite frankly, does not make sense) to work with infinitely many functions since, in order to perform actual computations for any real data analysis, the data set in question must be finite. Therefore, we work with subsets $\mathcal{B}^m_{\leq n}$ to ensure that we have an embedding into a finite-dimensional Euclidean space for computational feasibility. The functions $E_{m,(1,0)^i,(1,1)^j}$ that we subsequently list in (2.10), defined for a fixed m, are the finitely many functions we require to distinguish barcodes in $\mathcal{B}^m_{\leq n}$. This restriction does not affect the Lipschitz continuity of the polynomials, which always holds, both in the regularizing subset given by (2.9) as well as in general barcode space.

Tropical coordinates on barcode space. Using Proposition 2.8 and restating the proof of Theorem 6.3 in [51] for the following smaller family of functions, we establish that

$$(2.10) \quad E_{m,(0,1)^i,(1,1)^j}(x_1, d_1, x_2, d_2, \ldots, x_n, d_n) := E_{(0,1)^i,(1,1)^j}(x_1 \oplus d_1^m, d_1, \ldots, x_n \oplus d_n^m, d_n)$$

induces an injective map on $\mathcal{B}_{\leq n}^m$. This collection of functions separates nonequivalent barcodes and are Lipschitz with respect to the Wasserstein p- and bottleneck distances [51]. We now give an example (with steps) for calculating these functions and evaluating them on barcodes.

Example 2.9. Fix $n = 2$. The set of orbits under the S_2 action is

$$\mathscr{E}_2/S_2 = \left\{ \begin{array}{c} \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right], \\ \left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right] \end{array} \right\}.$$

According to Proposition 2.8, we need only consider the functions defined by the following orbits to separate barcodes when $n = 2$:

$$(2.11) \quad \left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right].$$

Suppose that we have two barcodes $\mathscr{B}_1 = [(1, 2), (3, 1)]$ and $\mathscr{B}_2 = [(2, 2)]$, where $\mathscr{B}_1, \mathscr{B}_2 \in \mathcal{B}_{\leq 2}$.

1. Compute m: For intervals $(1, 2)$, $(3, 1)$, $(2, 2)$, find the smallest m such that $x_i \leq m d_i$ for all i. This m is determined by rearranging (2.9) as $x_i \leq m d_i \Leftrightarrow \frac{x_i}{d_i} \leq m$, and must be greater than the quotients $\frac{x_i}{d_i}$ for $i = 1, 2, 3$. These quotients are $\frac{1}{2}, \frac{3}{1}, 1$, and therefore we take $m = 3$. Having determined m, it follows that $\mathscr{B}_1, \mathscr{B}_2 \in \mathcal{B}_{\leq 2}^3$.

2. Determine the 2-symmetric max-plus polynomials from the subcollection of orbits (2.11):

$$\begin{aligned} E_{3,(0,1)}(x_1, d_1, x_2, d_2) &= E_{3,(0,1),(0,0)}(x_1, d_1, x_2, d_2) \\ &= d_1 \boxplus d_2 \\ &= \max\{d_1, d_2\}, \\ E_{3,(0,1)^2}(x_1, d_1, x_2, d_2) &= E_{3,(0,1),(0,1)}(x_1, d_1, x_2, d_2) \\ &= d_1 d_2 \\ &= d_1 + d_2, \\ E_{3,(1,1)}(x_1, d_1, x_2, d_2) &= E_{3,(0,0),(1,1)}(x_1, d_1, x_2, d_2) \\ &= (x_2 \oplus d_2^3) d_2 \boxplus (x_1 \oplus d_1^3) d_1 \\ &= \max\left\{ \min\{x_2, 3d_2\} + d_2, \min\{x_1, 3d_1\} + d_1 \right\}, \\ E_{3,(0,1),(1,1)}(x_1, d_1, x_2, d_2) &= (x_1 \oplus d_1^3) d_1 d_2 \boxplus (x_2 \oplus d_2^3) d_2 d_1 \\ &= \max\left\{ \min\{x_1, 3d_1\} + d_1 + d_2, \min\{x_2, 3d_2\} + d_2 + d_1 \right\}, \\ E_{3,(1,1)^2}(x_1, d_1, x_2, d_2) &= E_{3,(1,1),(1,1)}(x_1, d_1, x_2, d_2) \\ &= (x_1 \oplus d_1^3) d_1 (x_2 \oplus d_2^3) d_2 \\ &= \min\{x_1, 3d_1\} + d_1 + \min\{x_2, 3d_2\} + d_2. \end{aligned}$$

3. Evaluate on $\mathscr{B}_1$ and $\mathscr{B}_2$.

The Euclidean-space vector representation of $\mathscr{B}_1$ and $\mathscr{B}_2$ is $(2, 3, 4, 6, 7)$ and $(2, 2, 4, 4, 4)$, respectively.

3. Sufficient statistics and likelihoods for barcodes. In this section, we motivate the need for sufficient statistics for persistent homology. We also present our main result that tropical functions, defined according to the max-plus and min-plus semirings and listed in (2.10), are sufficient statistics for persistence barcodes. Lastly, we discuss the implications of our result on statistical parametric inference by imposing the structure of exponential family distributions and defining likelihood functions for the tropical coordinatized representation of barcodes.

3.1. Sufficient statistics and persistent homology. Although a well-defined metric space, the space of barcodes exhibits a complex geometry with respect to the L_2-Wasserstein distance, it is an Alexandrov space with curvature bounded from below [83]. In the more conventional sense, this structure is theoretically not restrictive since under general Wasserstein p-distances, it is also a Polish space and admits well-defined characterizations of expectations, variances, and conditional probabilities [63]. This means that formal probability measures and distributions may be formulated for persistence diagrams and barcodes. Probability measures on the space of persistence diagrams have been of key interest for conducting statistical inference in TDA (e.g., [6, 14, 17]). However, formulating explicit, parametric probability distributions directly on the space of persistence diagrams—under either the Wasserstein p-distance or the bottleneck distance as the limiting case of the Wasserstein distance—remains a significant challenge. Although nonparametric methods for persistent homology exist (e.g., [21, 76]), in classical statistics, parametric modeling is the canonical starting point for data analysis and inference. Fitting nonparametric models can be more computationally intensive than parametric methods, which is another drawback given that computing persistent homology does not scale well for very large data sets. In particular, the upper size bound for Vietoris–Rips complexes based on N vertices is $2^{\mathcal{O}(N)}$, and the computational complexity for persistent homology implemented by the standard algorithm [32, 91] is cubic in the number of simplices (for a detailed survey on computing persistent homology, see [66]). Furthermore, the existence of Fréchet means has been proven, but these are not unique [83]. It is therefore difficult to perform statistical analyses, particularly parametric inference, directly on the space of persistence diagrams.

Statistical sufficiency. The difficulty of defining explicit probability distributions directly on the space of barcodes is our main motivation for the present work. To solve this problem, our strategy is to achieve sufficiency for barcodes via Euclidean coordinatization. This way we can establish distributional forms for the coordinatized equivalents of barcodes, as long as our mapping is a sufficient statistic. Given that the coordinatization is a topological embedding and the mappings are continuous and injective between barcode and Euclidean space, any parametric inference performed on these coordinatized equivalents naturally preserves all key properties in barcode space. The property of injectivity also ensures the existence of an inversion. Therefore, the results of parametric analyses on Euclidean space can theoretically be relayed back into barcode space by the inverse mappings. Note, however, that this is not necessary in practice, because the sufficiency guarantees that (for analytical purposes) all of the information from the original barcodes is retained by their Euclidean coordinates.

Intuitively, a statistic is said to be *sufficient* for a family of probability distributions if the original sample from which the statistic was computed provides no more additional information on the underlying probability distribution than does the statistic itself. A subtlety to remark here is that the term "statistic" has a dual meaning: the same term is used to refer to the point estimate, i.e., evaluation of a function, as well as to the function itself (as an example, the mean is often used to refer to both the value obtained from computing the empirical average of a finite set of data points as well as the function itself, which is the sum of data values divided by the number of data points). The concept of sufficiency retaining all information applies to both the statistic as a function of the data mapping to a codomain as well as the statistic as an evaluation of such a function. In the former case, sufficiency refers to a property of the statistic with respect to probability measures on the corresponding domain and codomain, where the exact distribution may be unknown but the spaces nevertheless admit well-defined probability measures. In the latter case, sufficiency refers to the purpose of determining the exact probability distribution that generated the observed data, which is a key aim of parametric inference. In this case, knowing a sufficient statistic contains all of the information needed to compute any estimator of the parameters for the assumed underlying distribution.

Definition 3.1. *Let X be a vector of observations of size n whose components X_i are independent and identically distributed (i.i.d.). Let ϑ be the parameter that characterizes the underlying probability distribution that generates X_i. A statistic $T(X)$ is* sufficient *for the parameter ϑ if the conditional probability of observing X, given $T(X)$, is independent of ϑ. Equivalently,*

$$\mathbb{P}(X = x \mid T(X) = t, \vartheta) = \mathbb{P}(X = x \mid T(X) = t).$$

The above definition applies to point evaluations of the statistic $T(X)$. From the measure-theoretic perspective, a statistic $T(\cdot)$ is sufficient for a given set $\mathcal{M}$ of probability measures on a σ-algebra $\mathbb{S}$ if the conditional probability of some subset $E \subseteq \mathbb{S}$, given a value y of $T(\cdot)$, is invariant for every probability measure in $\mu \in \mathcal{M}$.

We now recall an important result in the mathematical statistics literature for assessing statistical sufficiency: the factorization criterion for sufficient statistics [37, 64]. The following theorem is classically used as a definition for checking sufficiency, in particular, for statistics that are point estimates.

Theorem 3.2 (Fisher–Neyman factorization criterion). *If the probability density function for the observed data is $f(x; \vartheta)$, then the statistic $T = T(x)$ is sufficient for $\vartheta \in \Theta$ if and only if nonnegative functions g and h exist such that*

$$(3.1) \qquad\qquad f(x; \vartheta) = h(x)g\big(T(x); \vartheta\big).$$

Namely, the density $f(\cdot)$ can be factored into a product where one factor $h(\cdot)$ is independent of ϑ, and the other factor $g(\cdot)$ depends on ϑ and only depends on x only through $T(\cdot)$.

This factorization need not be unique for any given distribution. For example, to see that the identity function $T(X) = X$ is a sufficient statistic, we may factorize by setting $h(x) = 1$, and then $f(x; \vartheta) = g\big(T(x); \vartheta\big)$ for all x. In this case, $T(X)$ is the trivial sufficient statistic and is interpreted as the data themselves being sufficient for the data.

Example 3.3 (normal distribution: unknown mean μ, known variance σ^2). We denote the known parameter by the subscript $\vartheta = \sigma$. The normal probability density is

$$f_\sigma(x; \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

By setting

$$h_\sigma(x) := \frac{1}{\sqrt{2\pi\sigma^2}}\left\{-\frac{x^2}{2\sigma^2}\right\}, \quad T_\sigma(x) := \frac{x}{\sigma}, \quad g_\sigma(y) := \exp\left\{\frac{\mu}{\sigma}\cdot y - \frac{\mu^2}{2\sigma^2}\right\}$$

and substituting in (3.1) with $y = T_\sigma(x)$, we get

$$f_\sigma(x; \mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\cdot\exp\left\{\frac{\mu}{\sigma}\cdot\frac{x}{\sigma} - \frac{\mu^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

as desired.

Theorem 3.2 is valid for arbitrary families of distributions when studying sufficiency of point estimates. The following measure-theoretic reinterpretation asserts the generality of the theorem, particularly in the case where sufficiency refers to statistics on probability measures [44]. Essentially, the result depends on the assumption that the family of probability measures $\mathcal{P}_\vartheta$ is dominated by a single σ-finite measure. In the case of discrete variables, this amounts to the existence of a countable set S, which does not depend on ϑ, for which $\mathcal{P}_\vartheta(S) = 1$ for all $\vartheta \in \Theta$. In the case of absolutely continuous variables, this amounts to the measurability of the statistic.

Theorem 3.4 ([44]). *A necessary and sufficient condition that the statistic $T(\cdot)$ be sufficient for a dominated set $\mathcal{M}$ of measures on a σ-algebra $\mathbb{S}$ is that for every $\mu \in \mathcal{M}$, the density $f_\mu := \mathrm{d}\mu/\mathrm{d}\lambda$ (where $\mathrm{d}\mu/\mathrm{d}\lambda$ is the Radon–Nikodym derivative) admits the factorization*

$$f_\mu(x) = h(x)g_\mu\big(T(x)\big),$$

where $h(\cdot) \neq 0$ is a measurable function with respect to $\mathbb{S}$, and $g_\mu(\cdot)$ is a measurable function with respect to the codomain $\mathcal{T}$ of $T(\cdot)$.

As we will show, in the settings where the statistic is given by tropical polynomials in (2.10), $\mathbb{S}$ corresponds to the space of persistence barcodes $\mathcal{B}^m_{\leq n}$, where $\mathcal{T}$ is Euclidean space. Although the above theorem holds in broader generality, the case of absolutely continuous variables is sufficient for our setting since the tropical polynomials are Lipschitz continuous and are therefore absolutely continuous as well.

Sufficient statistics for persistence barcodes. Given the existence of parametric distributions and probability measures on barcode space [63], our main theorem states that the map from $\mathcal{B}^m_{\leq n}$, as an embedding into Euclidean space via tropical functions, is a sufficient statistic. Sufficiency is classically asserted by the injectivity of continuous mappings (e.g., [29, 44]). We follow this same approach in our proof.

Theorem 3.5. *Consider the subset $\mathcal{B}^m_{\leq n}$ of $\mathcal{B}_{\leq n}$ consisting of barcodes $[(x_1, d_1, \ldots, x_n, d_n)]$ with $d_i > 0$ such that $x_i \leq md_i$ for some fixed m and $i = 1, 2, \ldots, n$. Let $\mathcal{P}_\Theta$ be the family of probability measures on $\mathcal{B}^m_{\leq n}$. The map defined by*

$$T : \mathcal{B}^m_{\leq n} \to \mathbb{R}^d$$

(3.2)
$$\mathscr{B} \mapsto \left(E_{m,(0,1)^i,(1,1)^j}(x_1, d_1, x_2, d_2, \ldots, x_n, d_n) \right)_{i+j \in \mathbb{N}_{\leq n}}(\mathscr{B}),$$

where d is the number of orbits used to define separating functions, is a sufficient statistic for all $P_\vartheta \in \mathscr{P}_\Theta$.

Proof. Denote the image of $\mathcal{B}^m_{\leq n}$ under $(E_{m,(0,1)^i,(1,1)^j}(x_1, d_1, x_2, d_2, \ldots, x_n, d_n))_{i+j \in \mathbb{N}_{\leq n}}$ by $\mathcal{T} \subseteq \mathbb{R}^d$. On $\mathcal{B}^m_{\leq n}$, we take the bottleneck distance, and on $\mathcal{T}$, we take the standard Euclidean distance. We can view $\mathcal{B}^m_{\leq n}$ as a measurable space by taking the Borel σ-algebra to be the smallest σ-algebra that contains all open sets in the metric space $(\mathcal{B}^m_{\leq n}, d^B_\infty)$ with topology generated by open balls. The same goes for $\mathcal{T}$ with the standard Euclidean distance ([69, Chapter 1]). Both metric spaces are separable [63] and complete ($\mathcal{B}^m_{\leq n}$ is complete since it is a closed subspace of the barcode space).

To apply Theorem 3.4 and show that T generates sufficient statistics for probability measures on $\mathcal{B}^m_{\leq n}$, we use the property that the map T is injective, proven in Proposition 2.8. By definition, this means that there exists some $\varphi \colon \mathcal{T} \to \mathcal{B}^m_{\leq n}$ such that $\varphi \circ T = \mathrm{Id}_{\mathcal{B}^m_{\leq n}}$ and $T \circ \varphi = \mathrm{Id}_{\mathcal{T}}$. For notational simplicity, we denote $f_\vartheta(\mathscr{B}) := f(\mathscr{B}; \vartheta)$ and $g_\vartheta(T(\mathscr{B})) := g(T(\mathscr{B}); \vartheta)$. The condition of Theorem 3.4 holds for T where $g_\vartheta = f_\vartheta \circ \varphi$:

$$\begin{aligned}
f_\vartheta(\mathscr{B}) &= f_\vartheta\big(\varphi\big(T(\mathscr{B})\big)\big), \\
&= f_\vartheta \circ \varphi\big(T(\mathscr{B})\big), \\
&= g_\vartheta\big(T(\mathscr{B})\big).
\end{aligned}$$

We now proceed to verify that all the relevant functions are measurable, i.e., that for each of them, the preimage of any measurable set is measurable. The function $h(\mathscr{B}) \equiv 1$ is a constant and therefore continuous, from where it follows that it is measurable ([69, Theorem 1.5 in Chapter 1]). To show that f_ϑ is measurable, notice that $g_\vartheta = f_\vartheta \circ \varphi$. Since g_ϑ is measurable by assumption (Theorem 3.4), it will be sufficient to show $\varphi \colon \mathcal{T} \to \mathcal{B}^m_{\leq n}$ is measurable. Since the map T is Lipschitz continuous with respect to the bottleneck, Wasserstein p-, and the standard Euclidean distances, it is measurable with respect to Borel σ-algebras as specified previously ([69, Theorem 1.5 in Chapter 1]). By the Kuratowski theorem ([69, Corollary 3.3 in Chapter 3]), the inverse of an injective, measurable map between separable metric spaces equipped with Borel σ-algebras is measurable, which leads to the conclusion that φ is measurable.

This proves the result when $\mathcal{B}^m_{\leq n}$ is equipped with the bottleneck distance. A verbatim proof provides the same result for all Wasserstein p-distances. ∎

As previously mentioned, the dimension d of the codomain $\mathbb{R}^d$ for the tropical mappings in (3.2) need not equal the number of all orbits given by the symmetric group action, as a result of Proposition 2.8. The dimension d is determined by the restriction that $i + j \in \mathbb{N}_{\leq n}$, where n is the number of bars in the largest barcode, which simplifies to $2d = 2n + n(n+1)$.

3.2. Exponential families and likelihood functions for persistent homology. Sufficient statistics are strictly parametric by nature and thus are most relevant in studies centered around inference. Parametric inference is carried out under the assumption that the parameter

space is necessarily finite and therefore bounded with respect to sample size. The only class of distributions that satisfies this restriction is known as the *exponential family* [28, 55, 74], generated by a particular case of Theorem 3.2.

For many applications, data can be appropriately modeled using exponential family distributions. This is especially true for biological applications, which is the focus of the case study presented in section 4. Here, Gaussian distributions are commonly assumed for many quantitative traits such as human height (e.g., [82, 89]) and weight (e.g., [86]), while dichotomous phenotypes, such as disease status, are usually modeled as Bernoulli random variables (e.g., [80]). Counts, such as gene expression data, are often modeled as Poisson or negative binomial distributions (e.g., [8, 60]). These distributions are all members of the exponential family; indeed, many more important distributions such as the exponential, geometric, gamma, beta, chi-squared, and Dirichlet distributions are also members of this class. The application of such distributions also extend beyond biology and arise in many other scientific disciplines (e.g., quality control).

Definition 3.6. *Let the following restriction be applied to* (3.1) *in Theorem* 3.2,

$$(3.3) \qquad g(T(x); \vartheta) = \exp\big\{\langle \eta(\vartheta), T(x)\rangle - A(\vartheta)\big\},$$

where $\langle \cdot, \cdot \rangle$ *denotes the usual inner product in Euclidean space, and* $\eta(\vartheta)$, $T(x)$, *and* $A(\vartheta)$ *are known functions. The resulting class of probability distributions spanned by* (3.1), *with* $g(\cdot)$ *given by* (3.3), *is known as the* exponential family *of probability densities.*

Though less general than the form in (3.1), the exponential family is an important class of distributions that is comprised of many of the most common distributions in statistics. Given an observation x, the *likelihood function* $\mathscr{L}(\cdot)$ is a function of the parameter ϑ from the probability distribution f_ϑ that generates the observation within the sample. Namely,

$$
\begin{aligned}
\mathscr{L}(\vartheta \mid x) &= f_\vartheta(x) \\
&= h(x)g(T(x); \vartheta),
\end{aligned}
$$
(3.4)

with some sufficient statistic $T(x)$. For an exponential family model, if $g(\cdot)$ is given by (3.3) and we denote $a(\vartheta) := \exp\{-A(\vartheta)\}$, then we obtain

$$(3.5) \qquad \mathscr{L}(\vartheta \mid x) = h(x)a(\vartheta)\exp\big\{\langle \eta(\vartheta), T(x)\rangle\big\}.$$

Joint likelihoods for persistence barcodes. We now extend the above formulation to the multivariate case and, in particular, we consider the case for multiple barcodes. In this setting, we denote collections of K barcodes by $\boldsymbol{\mathscr{B}} = (\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_K)$, the tropicalized versions of these collections by $T(\boldsymbol{\mathscr{B}})$, and the collection of distribution parameters by $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)$. We will use $\mathscr{B}$, $T(\mathscr{B})$, and ϑ to notate singular (or univariate) elements of these vectors, respectively. For K barcodes marginally distributed according to $f_{\boldsymbol{\vartheta}}$, as defined by (3.4), their joint likelihood may be written as

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\vartheta} \mid \boldsymbol{\mathscr{B}}) &= f_{\boldsymbol{\vartheta}}(\boldsymbol{\mathscr{B}}) \\
&= h(\boldsymbol{\mathscr{B}})g(T(\boldsymbol{\mathscr{B}}); \boldsymbol{\vartheta}),
\end{aligned}
$$
(3.6)

with expectation $\mathbb{E}[\mathscr{B}] = \boldsymbol{\mu}$ and some general covariance structure $\mathbb{V}[\mathscr{B}] = \boldsymbol{\Sigma}$ between the barcodes. When $f_{\boldsymbol{\vartheta}}$ belongs to the exponential family, as denoted in (3.5), the joint likelihood for the K observations may be written as the following:

$$(3.7) \qquad\qquad \mathscr{L}(\boldsymbol{\vartheta} \mid \mathscr{B}) = a(\boldsymbol{\vartheta})^K \exp\left\{\left\langle \eta(\boldsymbol{\vartheta}), T(\mathscr{B}) \right\rangle\right\}.$$

Since the form of our tropical sufficient statistic (3.2) is a vector in Euclidean space for each barcode, this quantity is well defined. Most importantly, the inner product is readily computable in Euclidean space with respect to the natural Euclidean metric (as opposed to the space of barcodes which does not yield a natural definition for an inner product due to its complex geometry, as discussed above).

4. Application: Studying evolutionary behavior in viral data sets.
We now demonstrate the utility of our sufficiency result in applications regarding two infectious viral diseases: HIV and avian influenza. More specifically, we study both dimensions 0 and 1 of persistent homology. In dimension 0, we concretely demonstrate sufficiency in the preservation of biological characteristics (in the form of clustering behavior) of HIV. In dimension 1, we impose the parametric structure of an exponential distributive family to carry out pairwise comparisons between the marginal distributions of intra- and intersubtype reassortment in avian influenza.

4.1. Demonstrating sufficiency in HIV transmission clustering.
HIV transmission clusters represent groups of epidemiologically related individuals who share a common recent viral ancestor [45, 50]. Recently, [70] reported the existence of a prominent HIV transmission cluster affecting more than 100 people from a population of men who have sex with men (MSM) in Valencia, Spain.

Since epidemiologically related groups are often characterized by the identification of small pairwise genetic distances, sequences derived from a transmission cluster can be easily differentiated from those derived from unrelated patients by statistically assessing their clustering behavior (see Figure 2). Using persistent homology, dimension 0 (PH_0) captures the clustering of sequences based on vertical evolution (i.e., the amount of evolution explained solely by mutation [20]). This is shown via bottleneck distances in Figure 3(a). This particular application serves as a real-data example to demonstrate that tropical coordinatization retains all information contained in barcodes. Specifically, we show that no recombination events occurred in this HIV data set, indicated by the similar clustering pattern shown among the phylogenetic trees obtained for each independent segment of analyzed viral samples. More generally, persistent homology not only provides information on clusters that would also be recovered by single-linkage clustering in dimension 0, but it also provides relevant information on the potential for higher dimensional persistence intervals to appear. Higher dimensional persistence intervals would suggest that there exists recombination or reassortment events. Such information cannot be obtained from single-linkage clustering alone.

In order to demonstrate sufficiency and verify that the tropical representations of the barcodes retain the same biologically-relevant clustering information in dimension 0, we generated 10 random subsets by uniform sampling without replacement of 30 polymerase sequences from (i) the MSM transmission cluster, (ii) the reference data set (i.e., the unrelated patients), and (iii) sequences from both the transmission cluster and the reference data set. In total, 10×30 sequences were analyzed. Matrices of pairwise genetic distances were obtained from each
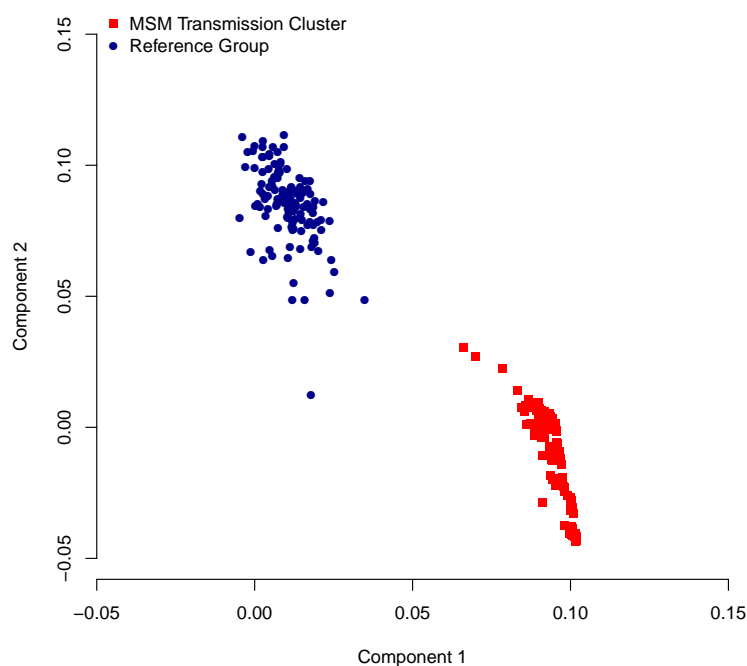
Figure 2. **Principal component analysis plot of genetic clustering for HIV sequences.** *Principal components calculated from pairwise genetic distances for HIV transmission cluster versus unrelated reference cluster.*
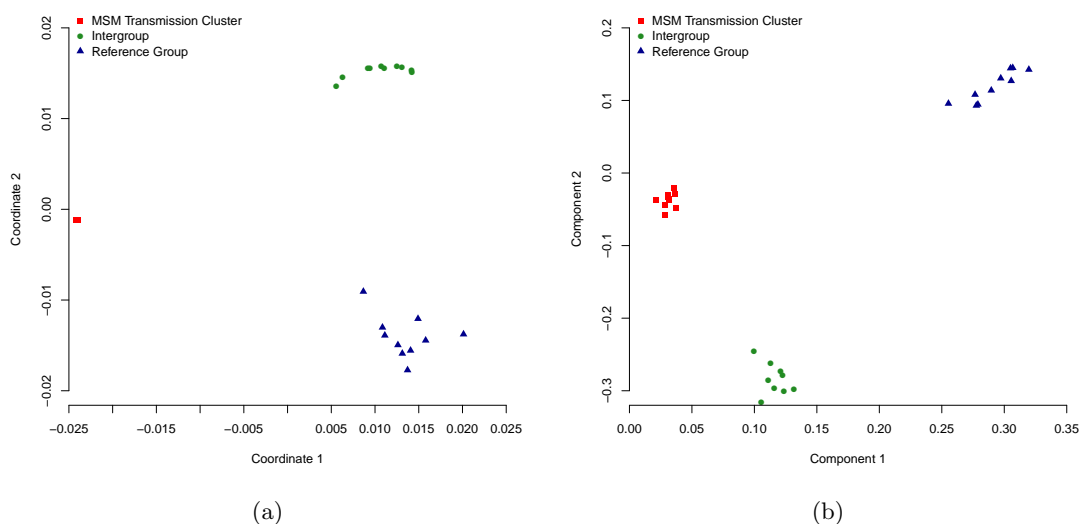

Figure 3. **Projection plots for pairwise bottleneck distances and Euclidean distances for HIV sequences.** **(a)** *Metric multidimensional scaling (MDS) calculated from pairwise bottleneck distances calculated from dimension 0 persistence barcodes.* **(b)** *Principal component analysis (PCA) calculated from pairwise Euclidean distances of tropical coordinatized barcodes.*

subset. The entries of these matrices are evolutionary distances between sequences, calculated from observed differences between sequences using evolutionary models (see Appendix C for further details).

Obtaining phylogenetic trees from distance-based methods as described above can be viewed analogously to embedding a general metric space (which in our case, is spanned by distances between gene loci) into tree metrics. This is a specific case of a more general research problem in the theory of finite metric spaces [1]. As in previous work [20], the underlying topological space we consider is the phylogenetic tree generated by the genetic distance matrix. A phylogenetic tree is a specific subgraph defined by a set of vertices called *leaves* and a set of edges called *branches*—each with positive length, representing evolutionary time. A leaf is a vertex with degree 1, a root is a vertex with degree 2, and nonleaf vertices have degree of at least 2. The graph in consideration is the topological space arising from a usual graph $G = (V, E)$, where V denotes the set of vertices and E is the set of edges. A graph as a topological space can be obtained from a usual graph by replacing the vertices by points and the edges $e = st$ by a copy of the unit interval $[0, 1]$, where 0 identifies with the point associated to the vertex s and 1 identifies with the point associated to the vertex t. As a topological space, a graph is the simplicial 1-complex generated by G [46]. Further details in mathematical phylogenetics from the perspective of algebraic statistics can be found in [67], and from the perspective of topology and metric geometry in [59]. Since the subgraph of a graph (as topological space) is itself a topological space, a phylogenetic tree can be viewed as the underlying topological space [20]. In our setting, leaves in the phylogenetic tree are viral sequences sampled from individuals, and edges are genetic distances between them.

We obtained the dimension 0 Vietoris–Rips persistence barcodes using Ripser [10] and evaluated them via the tropical functions described in subsection 2.2. The resulting barcodes (and their post-transformed Euclidean representations) from subsets (i) and (ii) provide information on intragroup distances, while those from subset (iii) provide information on intergroup distances (i.e., transmission cluster versus unrelated patients). Figure 2 and Figure 3 show that the three subsets can be easily differentiated, thus demonstrating that the tropical coordinatization of the barcodes are sufficient statistics retaining the clustering behavior of epidemiologically related individuals.

4.2. Analyzing intra- and intersubtype reassortment in avian influenza. The influenza virus presents a genome that consists of eight different ribonucleic acid (RNA) molecules (segments). Consequently, genetic reassortment—that is, the process of swapping gene segments—is an important evolutionary process that redistributes genetic variability. Subtype classification of influenza is based on the analysis of two surface proteins: hemagglutinin (HA) and neuraminidase (NA). There exist 18 types of HA (indexed from 1 to 18) and 11 types of NA (indexed from 1 to 11). Gene reassortment can occur either between viruses from the same subtype (i.e., intrasubtype reassortment [61]) or between viruses from different subtypes (i.e., intersubtype reassortment [65]). From a broad perspective, detecting gene reassortment is key to understanding mutations within the evolutionary dynamics of viruses.

Traditionally, the detection of reassortment events is based on molecular phylogenetic analyses, in which different genomic regions (i.e., segments) are analyzed independently. Specific details on this procedure given in Appendix B. As mentioned previously, if no reassortment

has occurred, phylogenetic trees obtained for each independent segment show a similar clustering pattern of the viral samples that are analyzed. This indicates that all segments share the same evolutionary history. Alternatively, when a reassortment event occurs, it is detected by a segment generating a phylogenetic tree with a differing clustering pattern. This indicates the existence of segment swapping [73]. The detection and quantification of reassortment events with phylogenetic approaches is a slow and tedious process, partly due to the fact that obtaining accurate phylogenetic trees can take days, or even weeks, when hundreds or thousands of samples are analyzed [61]. The use of persistent homology in detecting genetic reassortment events poses a clear advantage over phylogenetic analysis, not only due to the increased computational speed, but also because of the enhanced interpretability: dimension 1 persistence intervals provide explicit information on the genetic divergence of the sequences involved in the reassortment event [20, 33].

Given that viruses from the same subtype are more closely related to each other than they are to viruses from different subtypes, it is expected that lengths of intervals in PH_1 barcodes derived from events of intrasubtype reassortment will be shorter than those from intersubtype reassortment [33]. We will demonstrate this parametrically using our sufficiency result. In this context, sufficiency means preserving the information that intra- and intersubtype samples should not cluster together.

Since we are interested in analyzing reassortment events in this example, we focus on avian influenza type A and study dimension 1 persistence. As discussed above, in the same manner that phylogenetic trees are specific subgraphs and therefore are topological spaces; here the subgraph representing reassortment events is a directed acyclic graph and is also a topological space [20]. By uniform sampling without replacement, as above, we generated 100 random subsets of 56 concatenated HA and NA sequences representing: (i) only H5N1 sequences, and (ii) sequences from all subtypes H5Nu and HvN1 where $u = 1, 2, \ldots, 12$ and $v = 1, 2, \ldots, 9$, respectively. In total, 100×56 sequences were analyzed. H5N1 is of particular interest because of its characteristics as a highly pathogenic subtype of avian influenza with over 60% mortality and its high potential to cause a pandemic affecting the human population [54]. The resulting barcodes from subsets (i) provide information on intrasubtype reassortment events, while those from subsets (ii) provided information on intersubtype reassortment events.

Sufficiency is inherently a parametric construct, and since we have sufficient statistics for the family of probability measures on the space of barcodes, the parametric structure of probability distributions on and between the two spaces is thus preserved [13, 58]. We assume that the Euclidean representation of barcodes, from each of the intra- and intersubtype populations, jointly come from a multivariate normal distribution:

$$T(\mathscr{B}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with K-dimensional mean vector $\boldsymbol{\mu}$, and $K \times K$ symmetric and positive semidefinite covariance matrix $\boldsymbol{\Sigma}$. This implies that we may assume each barcode to have a marginal univariate normal distribution:

$$(4.1) \qquad T(\mathscr{B}_k) \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad k = 1, 2, \ldots, K,$$

with scalar mean and variance $\vartheta = \{\mu_k, \sigma_k^2\}$, and $K = \{43, 100\}$ for the intra- and intersubtype reassortment events, respectively. The data on intra- and intersubtype reassortment events

were obtained from the publicly available data source (see Appendix C for details); data on intrasubtype reassortment are more limited than those on intersubtype reassortment. The distribution plots for the data of each class of reassortment events are shown in Figure 4.
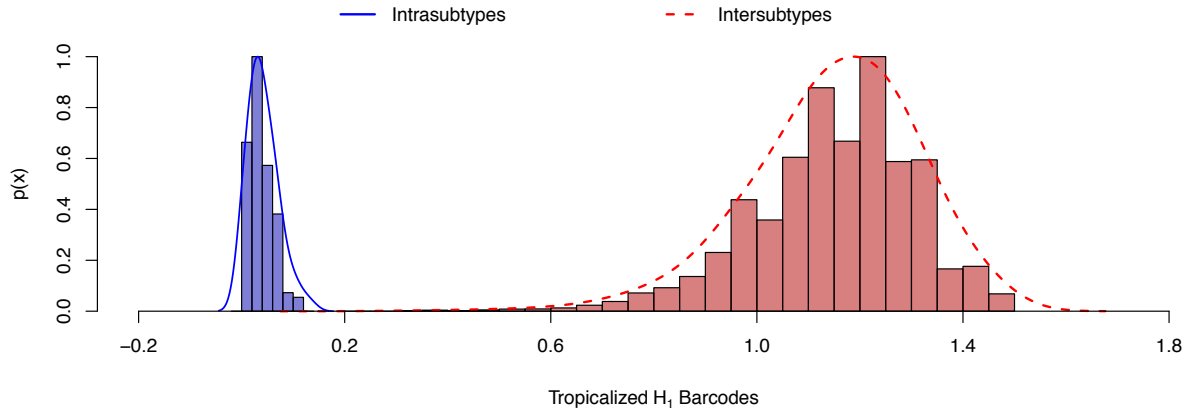

Figure 4. Marginal distribution plot of intra- and intersubtype reassortment events for avian influenza. *The marginal distributions for square-root transformations of the Euclidean barcode representations (via tropical coordinatization) were calculated for both intra- and intersubtype reassortment for avian influenza and then fitted with a smooth density function.*

In this parametric setting, we may study pairwise comparisons between the marginal distributions of each $T(\mathscr{B}_k)$ in order to assess the following clustering behavior: samples that cluster together will tend to have marginal distributions that are more similar or, in other words, have shorter distances. A function commonly used to measure differences between two probability distributions is an f-divergence. Two particular instances of an f-divergence are the Kullback–Leibler divergence [56] and Hellinger distance [47], which we now briefly define.

Definition 4.1. Assume that $T(\mathscr{B}_i)$ and $T(\mathscr{B}_j)$ are probability measures that are absolutely continuous with respect to a third probability measure, λ. The Kullback–Leibler divergence *from $T(\mathscr{B}_i)$ to $T(\mathscr{B}_j)$ is defined to be the integral*

$$(4.2) \qquad KLD\big(T(\mathscr{B}_i) \,\|\, T(\mathscr{B}_j)\big) = \int \frac{\mathrm{d}T(\mathscr{B}_i)}{\mathrm{d}\lambda} \log\left(\frac{\mathrm{d}T(\mathscr{B}_i)}{\mathrm{d}\lambda} \Big/ \frac{\mathrm{d}T(\mathscr{B}_j)}{\mathrm{d}\lambda}\right) \mathrm{d}\lambda.$$

Alternatively, the Hellinger distance *between the measures may be computed by solving the following*

$$(4.3) \qquad H^2\big(T(\mathscr{B}_i), T(\mathscr{B}_j)\big) = \frac{1}{2} \int \left(\sqrt{\frac{\mathrm{d}T(\mathscr{B}_i)}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}T(\mathscr{B}_j)}{\mathrm{d}\lambda}}\right)^2 \mathrm{d}\lambda,$$

where $\mathrm{d}T(\mathscr{B}_i)/\mathrm{d}\lambda$ and $\mathrm{d}T(\mathscr{B}_j)/\mathrm{d}\lambda$ are the Radon–Nikodym derivatives of $T(\mathscr{B}_i)$ and $T(\mathscr{B}_j)$, respectively.

Intuitively, $\mathrm{KLD}(\cdot, \cdot)$ measures the relative entropy (or information gain) when comparing statistical models of inference; while, $\mathrm{H}^2(\cdot, \cdot)$ is the probabilistic analog of the Euclidean

distance between two distributions. The latter is a natural measure for the Euclidean-space vector representation of barcodes. The Kullback–Leibler divergence and Hellinger distance are nonnegative quantities and, in this context, take the value of zero if and only if two probability distributions are exactly equivalent. Derivable from the Cauchy–Schwarz inequality, the Hellinger distance also has the desirable property of satisfying the unit interval bound for all probability distributions. Namely,

$$0 \leq \mathrm{KLD}\big(T(\mathscr{B}_i) \,\|\, T(\mathscr{B}_j)\big) < \infty; \qquad 0 \leq \mathrm{H}^2\big(T(\mathscr{B}_i), T(\mathscr{B}_j)\big) \leq 1.$$

Note that it can be shown that when the probability measures $T(\mathscr{B}_i)$ and $T(\mathscr{B}_j)$ stem from the exponential family, the solutions of both integrals in (4.2) and (4.3) have a closed form (e.g., [31, 85]). More specifically, for two normal random variables $T(\mathscr{B}_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and $T(\mathscr{B}_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$, we have the following:

$$(4.4) \qquad \mathrm{KLD}\big(T(\mathscr{B}_i) \,\|\, T(\mathscr{B}_j)\big) = \log \frac{\sigma_j}{\sigma_i} + \left[\frac{\sigma_i^2 + (\mu_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2} \right],$$

$$(4.5) \qquad \mathrm{H}^2\big(T(\mathscr{B}_i), T(\mathscr{B}_j)\big) = 1 - \sqrt{\frac{2\sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2}} \exp\left\{ -\frac{(\mu_i - \mu_j)^2}{4(\sigma_i^2 + \sigma_j^2)} \right\}.$$

Based on the marginal normality assumption in (4.1) with parameter values μ_i, μ_j and σ_i, σ_j estimated empirically from the data, we show that the tropical coordinatization of persistence barcodes sufficiently preserves the fact that intra- and intersubtype samples do not cluster together. This was done by calculating (4.4) and (4.5) for every pair $T(\mathscr{B}_i)$ and $T(\mathscr{B}_j)$ and encoding this information as similarity matrices $\mathbf{K}$ and $\mathbf{H}$, respectively. Note that the key reasoning behind considering two different types of divergence measures is to illustrate the robustness of the results. Figure 5 plots a transformed version of the matrix $\mathbf{K}^* = \mathbf{11}^\intercal - \exp\{-\mathbf{K}\}$ (for the purpose of exact comparisons and interpretability), where $\mathbf{1}$ is a vector of ones, and Hellinger distance matrix $\mathbf{H}$. These results can be interpreted as follows: a value of 0 represents exact likeness, while 1 represents complete dissimilarity. As expected from the literature, the intersubtypes are seen to be more homogeneous and demonstrate a higher level of similarity, while the intrasubtypes are comparatively much more random (e.g., [61, 62, 88]). Moreover, there is very little overlap between the two groups. These results are consistent for both f-divergence measures.

5. Discussion. In this paper, we explored a functional vectorization method for persistence barcodes that is a topological embedding into Euclidean space based on tropical algebraic geometry. We proved that it generates sufficient statistics for the family of all probability measures on the space of barcodes under mild regularity conditions. The statistical and data analytic utility of this result lies in the fact that sufficiency allows parametric probabilistic assumptions to be imposed, which previously has been difficult due to the prohibitive geometry of the space of barcodes. This allows for the application of classical parametric inference methodologies to persistence barcodes, which we demonstrated in a concrete example in dimension 1 persistence applied to avian influenza data. Our methodological approach is based on the assumption that the random barcodes we obtain from real data may be represented
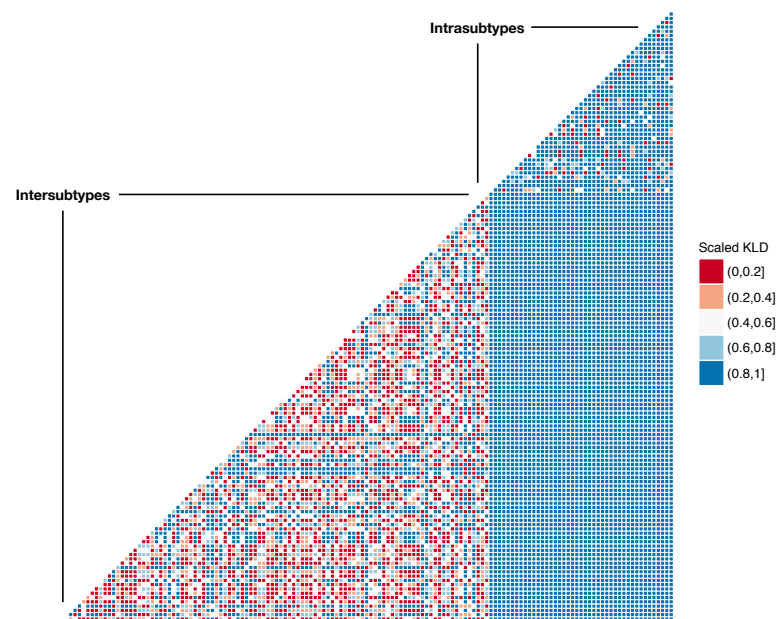
(a) Scaled Kullback–Leibler Divergence
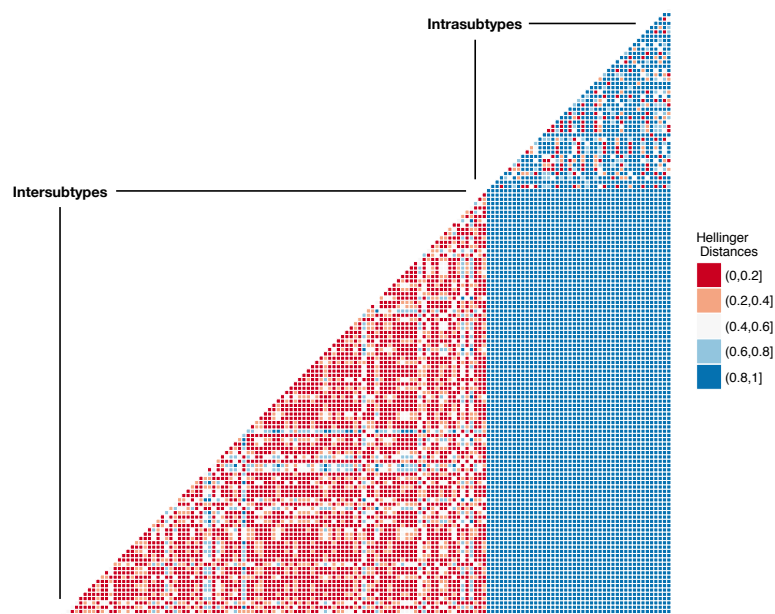
(b) Hellinger Distance

Figure 5. **Visual pairwise comparisons of transformed Kullback–Leibler divergence and Hellinger distances between Euclidean barcode representations.** *The entries are color-coded according to values of the corresponding similarity matrices* **(a)** $\mathbf{K}^* = \mathbf{11}^\mathsf{T} - \mathbf{K}$, *where* $\mathbf{1}$ *is a vector of ones, and* **(b)** $\mathbf{H}$. *A value of* 0 *represents exact likeness, while* 1 *represents complete dissimilarity. Red values represent those that are more similar, while blue values are more dissimilar. This plot shows that intersubtypes are more homogeneous (a higher level of similarity, i.e., more red), while the intrasubtypes are comparatively much more random. There is very little overlap between the two groups.*

by a parametric model. Given the complex representation of persistent homology, we recognize that such an assumption may not always be reasonable—in which case, nonparametric methods as a more flexible alternative option, e.g., [7]. However, note that the trade-off in relinquishing any parametric assumption is the extent to which inference may be carried out, which is a well-known problem in statistical modeling, e.g., [26].

Our sufficiency also result provides the foundations for future research towards a better understanding of parametric probabilistic behavior on the space of barcodes. Since the tropical vectorization method was shown to be an injective function, we may take a class of parametric probability distributions (e.g., the exponential family) and calculate its image via the inverse of the tropical functions in order to explore what functional form the exponential family assumes on the space of barcodes. Such a study would be algorithmic in nature, since it would entail exploring collections of maps given by subsets of the orbits under the symmetric group action.

In terms of an increased utility in statistical analyses, the tropical functions are limited since the vector representation is unique up to barcodes only and not on the level of bars. This is restrictive in questions motivated by the application of our paper. For example, when predicting or modeling genetic outbreaks of infectious diseases, individual sequences (or at least the persistence generators) would need to be identifiable. Future research towards these efforts may begin with the development of a vectorization method that traces back to the individual bars of a barcode.

Software and data availability. Software to compute and evaluate the tropical functions on barcodes is publicly available in C++ code, coauthored by Melissa McGuirl and Steve Oudot and located on the Tropix GitHub repository at https://github.com/lorinanthony/Tropix. The persistence barcodes were calculated using Ripser [10], which is written in C++ and is freely available at https://github.com/Ripser/ripser. The bottleneck distances were computed using Hera [53], which is also written in C++ and freely available at https://bitbucket.org/grey_narn/hera. The data used in this paper were obtained from publicly available sources and preprocessed, as detailed in Appendix C. The final version used in the analyses in this paper are also publicly available on the Tropix GitHub repository.

Appendix A. Mathematical supplement.

Claim A.1. The polynomials that are a ring of algebraic functions on the space of barcodes, identified by [4], are not Lipschitz with respect to the Wasserstein p- and bottleneck distances.

Proof. Consider the following counterexample. Take one barcode to be the diagonal Δ (i.e., intervals of length 0) denoted by $\mathscr{B}_0$ and consider the family of barcodes with a single interval $\mathscr{B}_x = \{[x, x+1]\}$. The Wasserstein p- and bottleneck distances between these $\mathscr{B}_0$ and $\mathscr{B}_x$ is the same for all x, namely, $1/2$. Consider a polynomial given by [4], e.g.,

$$p_{2,1}(x_1, y_1, x_2, y_2, \ldots) = \sum_i (x_i + y_i)^2 (y_i - x_i)^1,$$

then

$$p_{2,1}(\mathscr{B}_0) = 0 \quad \text{and} \quad p_{2,1}(\mathscr{B}_x) = (2x + 1)^2.$$

Now, the difference $p_{2,1}(\mathscr{B}_x) - p_{2,1}(\mathscr{B}_0)$ tends to infinity as x tends to infinity. The bottleneck (and Wasserstein p-) distance, meanwhile, is constant at value $1/2$ the entire time. Therefore,

by definition, this function is not Lipschitz. This particular case is problematic and shows that other functions proposed in [4] are also not Lipschitz with respect to the Wasserstein p- and bottleneck distances. ∎

Appendix B. Molecular phylogenetic analysis.

A phylogenetic analysis consists of studying the evolutionary relationships of a set of individuals (i.e., viral sequences), with the aim of extracting and analyzing their diversification history. This type of analysis is usually performed by using an alignment of sequences as input. Given this input, a *phylogenetic tree* is then produced—that is, a graphical representation of the evolutionary relationships among the individuals. Common statistical methods of phylogenetic inference are maximum likelihood and Bayesian phylogenetics, which aim to maximize a statistical parameter (either the likelihood of the alignment or its posterior probability) by modeling the molecular evolutionary processes considered in the phylogenetic reconstruction using stochastic processes (e.g., [43, 77]). Other methods are distance-based, which may be used on the trees themselves as a means of reconstruction (e.g., [38, 78]), or on spaces of trees, which are commonly used to study and compare various trees, e.g., [12]. See [67] and [59] for further detail and a discussion on challenges of computational phylogenetic analysis.

Appendix C. Data sourcing and preprocessing.

The respective virus data sets in this paper were obtained from public sources. We give the details on their sources and preliminary data processing procedures below.

HIV. HIV polymerase sequences derived from patients included in the MSM HIV transmission cluster [71] were retrieved from supplementary material made public on GenBank. Sequences derived from unrelated patients were obtained from the Los Alamos HIV database in October 2016. Only sequences from the same subtype (HIV subtype B) spanning the polymerase region were considered. In order to ensure that these sequences were not epidemiologically related, redundant sequences were removed after conducting an initial clustering analysis with a specified genetic distance threshold of 5%, using CD-HIT [49]. All sequences were aligned using MAFFTv7 [52].

Avian influenza. HA and NA genes of avian influenza A were downloaded from the Influenza Virus Database of the National Center for Biotechnology Information (NCBI). The resulting gene data sets were aligned with MAFFTv7 [52]. Concatenated sequences of both genes (derived from the same sample) were generated with the package `ape` written in R [68]. The multiple sequence alignments were trimmed with TrimAl [16] in order to remove regions of sparse homology (i.e., biologically shared ancestry).

In both viral examples, pairwise distances were obtained using PAUP* [79] and were calculated by using the GTR + GAMMA (4 CAT) model, which is commonly used for studying HIV and influenza data [81, 87]. Briefly, the generalized time reversible (GTR) model is an evolutionary model that considers variable base frequencies, where each pair of nucleotide substitutions occur at different rates [30]. Combined with a gamma distribution, it also accounts for rate variation among sites [90]. The use of a substitution model when calculating genetic distances, as carried out according to these procedures, leads to estimates that are assumed to be more biologically accurate.

REFERENCES

[1] I. Abraham, Y. Bartal, and O. Neiman, *Embedding metrics into ultrametrics and graphs into spanning trees with constant average distortion*, in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, pp. 502–511.

[2] H. Adams and G. Carlsson, *Evasion paths in mobile sensor networks*, Int. J. Robotics Res., 34 (2015), pp. 90–104.

[3] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, *Persistence images: A stable vector representation of persistent homology*, J. Mach. Learn. Res., 18 (2017), pp. 218–252, http://dl.acm.org/citation.cfm?id=3122009.3122017.

[4] A. Adcock, E. Carlsson, and G. Carlsson, *The ring of algebraic functions on persistence bar codes*, Homology Homotopy Appl., 18 (2016), pp. 381–402.

[5] A. Adcock, D. Rubin, and G. Carlsson, *Classification of hepatic lesions using the matching metric*, Comput. Vision Image Understanding, 121 (2014), pp. 36–42, http://www.sciencedirect.com/science/article/pii/S1077314213002221.

[6] R. Adler and J. Taylor, *Topological Complexity of Smooth Random Functions: École d'Été de Probabilités de Saint-Flour XXXIX-2009*, Lecture Notes in Mathematics, Springer, Berlin, 2011.

[7] R. J. Adler, S. Agami, and P. Pranav, *Modeling and replicating statistical topology and evidence for CMB nonhomogeneity*, Proc. Natl. Acad. Sci., (2017), http://www.pnas.org/content/early/2017/10/23/1706885114.

[8] S. Anders and W. Huber, *Differential expression analysis for sequence count data*, Genome Biol., 11 (2010), R106, https://doi.org/10.1186/gb-2010-11-10-r106.

[9] M. S. Bartlett, *Properties of sufficiency and statistical tests*, Proc. Roy. Soc. London Ser. A Math. Phys. Sci., 160 (1937), pp. 268–282.

[10] U. Bauer, *Ripser*, 2015, https://github.com/Ripser/ripser.

[11] P. Bendich, S. P. Chin, J. Clark, J. Desena, J. Harer, E. Munch, A. Newman, D. Porter, D. Rouse, N. Strawn, and A. Watkins, *Topological and statistical behavior classifiers for tracking applications*, IEEE Trans. Aerospace Electronic Syst., 52 (2016), pp. 2644–2661, https://doi.org/10.1109/TAES.2016.160405.

[12] L. J. Billera, S. P. Holmes, and K. Vogtmann, *Geometry of the Space of Phylogenetic Trees*, Adv. Appl. Math., 27 (2001), pp. 733–767, http://www.sciencedirect.com/science/article/pii/S0196885801907596.

[13] P. Billingsley, *Probability and Measure*, John Wiley & Sons, New York, 1979.

[14] A. J. Blumberg, I. Gal, M. A. Mandell, and M. Pancia, *Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces*, Found. Comput. Math., 14 (2014), pp. 745–789.

[15] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, J. Mach. Learn. Res., 16 (2015), pp. 77–102.

[16] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses*, Bioinformatics, 25 (2009), pp. 1972–1973, http://dx.doi.org/10.1093/bioinformatics/btp348.

[17] G. Carlsson, *Topological pattern recognition for point cloud data*, Acta Numerica, 23 (2014), pp. 289–368.

[18] G. Carlsson and S. Kališnik Verovšek, *Symmetric and r-symmetric tropical polynomials and rational functions*, J. Pure Appl. Algebra, 220 (2016), pp. 3610–3627, http://www.sciencedirect.com/science/article/pii/S0022404916300251.

[19] M. Carrière, S. Y. Oudot, and M. Ovsjanikov, *Stable Topological Signatures for Points on* 3D *Shapes*, in Proceedings of the Eurographics Symposium on Geometry Processing 2015, 34 (2015), pp. 77–102.

[20] J. M. Chan, G. Carlsson, and R. Rabadán, *Topology of viral evolution*, in Proceedings of the National Academy of Sciences, 110 (2013), pp. 18566–18571, http://www.pnas.org/content/110/46/18566.abstract.

[21] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman, *Stochastic convergence of persistence landscapes and silhouettes*, in Proceedings of the Thirtieth Annual Symposium on Computational Geometry, SOCG'14, New York, 2014, ACM, pp. 474:474–474:483, http://doi.acm.org/10.1145/2582112.2582128.

[22] Y.-C. Chen, D. Wang, A. Rinaldo, and L. Wasserman, *Statistical analysis of persistence intensity functions*, preprint, arXiv, (2015), arXiv:1510.02502.

[23] M. K. Chung, P. Bubenik, and P. T. Kim, *Persistence diagrams of cortical surface data*, in Information Processing in Medical Imaging, J. L. Prince, D. L. Pham, and K. J. Myers, eds., Springer, Berlin, 2009, pp. 386–397.

[24] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of persistence diagrams*, Discrete Comput. Geometry, 37 (2007), pp. 103–120, https://doi.org/10.1007/s00454-006-1276-5.

[25] L. Crawford, A. Monod, A. X. Chen, S. Mukherjee, and R. Rabadán, *Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform*, preprint, arXiv, (2016), arXiv:1611.06818.

[26] L. Crawford, K. C. Wood, X. Zhou, and S. Mukherjee, *Bayesian approximate kernel regression with variable selection*, J. Amer. Stat. Assoc., (2018), pp. 1–12, https://doi.org/10.1080/01621459.2017.1361830.

[27] C. Curto, V. Itskov, A. Veliz-Cuba, and N. Youngs, *The neural ring: An algebraic tool for analyzing the intrinsic structure of neural codes*, Bull. Math. Biol., 75 (2013), pp. 1571–1611.

[28] G. Darmois, *Sur les lois de probabilités à estimation exhaustive*, C.R. Acad. Sci. Paris (in French), 200 (1935), pp. 1265–1266.

[29] P. Diaconis, *Sufficiency as statistical symmetry*, in Proceedings of the AMS Centennial Symposium, 1988, pp. 15–26.

[30] P. Donnelly and S. Tavaré, *Coalescents and genealogical structure under neutrality*, Ann. Rev. Genetics, 29 (1995), pp. 401–421, https://doi.org/10.1146/annurev.ge.29.120195.002153. PMID: 8825481.

[31] J. Duchi, *Derivations for Linear Algebra and Optimization*, Technical report, Berkeley, California, 2007, https://web.stanford.edu/~jduchi/projects/general_notes.pdf.

[32] H. Edelsbrunner, D. Letscher, and A. J. Zomorodian, *Topological persistence and simplification*, Discrete Comput. Geometry, 28 (2002), pp. 511–533.

[33] K. Emmett, D. Rosenbloom, P. Cámara, and R. Rabadán, *Parametric inference using persistence diagrams: a case study in population genetics*, preprint, arXiv, (2014), arXiv:1406.4582.

[34] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, *Confidence sets for persistence diagrams*, Ann. Stat., 42 (2014), pp. 2301–2339, https://doi.org/10.1214/14-AOS1252.

[35] M. Ferri and C. Landi, *Representing size functions by complex polynomials*, Proc. Math. Met. Pattern Recognition, 9 (1999), pp. 16–19.

[36] M. Ferri and I. Stanganelli, *Size Functions for the Morphological Analysis of Melanocytic Lesions*, J. Biomedical Imaging, 2010 (2010), pp. 5:1–5:5, http://dx.doi.org/10.1155/2010/621357.

[37] R. A. Fisher, *On the mathematical foundations of theoretical statistics*, Philosoph. Trans. Roy. Soc. London A Math. Phys. Engrg. Sci., 222 (1922), pp. 309–368.

[38] W. M. Fitch and E. Margoliash, *Construction of Phylogenetic Trees*, Science, New York, 155 (1967), pp. 279–284, http://www.jstor.org/stable/1720651.

[39] P. Frosini and C. Landi, *Size theory as a topological tool for computer vision*, Pattern Recognition Image Anal., 9 (1999), pp. 596–603.

[40] R. Ghrist, *Barcodes: The persistent topology of data*, Bull. Amer. Math. Soc. (N.S.), 45 (2008), pp. 61–75, http://dx.doi.org/10.1090/S0273-0979-07-01191-3.

[41] R. Ghrist and V. de Silva, *Coordinate-free coverage in sensor networks with controlled boundaries via homology*, Int. J. Robotics Res., 25 (2006), pp. 1205–1222.

[42] C. Giusti, E. Pastalkova, C. Curto, and V. Itskov, *Clique topology reveals intrinsic geometric structure in neural correlations*, Proc. Natl. Acad. Sci., 112 (2015), pp. 13455–13460, http://www.pnas.org/content/112/44/13455.

[43] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, *New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml* 3.0, Syst. Biol., 59 (2010), pp. 307–321, https://doi.org/10.1093/sysbio/syq010.

[44] P. R. Halmos and L. J. Savage, *Application of the Radon–Nikodym Theorem to the Theory of Sufficient Statistics*, Ann. Math. Stat., 20 (1949), pp. 225–241, https://doi.org/10.1214/aoms/1177730032.

[45] A. S. Hassan, O. G. Pybus, E. J. Sanders, J. Albert, and J. Esbjornsson, *Defining HIV-1 transmission clusters based on sequence data*, AIDS, 31 (2017), pp. 1211–1222.

[46] A. Hatcher, *Algebraic Topology*, Algebraic Topology, Cambridge University Press, Cambridge, 2002.

[47] E. Hellinger, *Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.*, J. Reine Angew. Math., 136 (1909), pp. 210–271.

[48] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, *Deep learning with topological signatures*, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Curran Associates, 2017, pp. 1634–1644, http://papers.nips.cc/paper/6761-deep-learning-with-topological-signatures.pdf.

[49] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, *CD-HIT Suite: A web server for clustering and comparing biological sequences*, Bioinformatics, 26 (2010), pp. 680–682.

[50] S. Hué, D. Pillay, J. P. Clewley, and O. G. Pybus, *Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups*, Proc. Natl. Acad. Sci., 102 (2005), pp. 4425–4429, http://www.pnas.org/content/102/12/4425.

[51] S. Kališnik, *Tropical coordinates on the space of persistence barcodes*, Found. Comput. Math., (2018), pp. 1–29.

[52] K. Katoh and D. M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability*, Molecular Biol. Evolution, 30 (2013), pp. 772–780, https://doi.org/10.1093/molbev/mst010.

[53] M. Kerber, D. Morozov, and A. Nigmetov, *Geometry Helps to Compare Persistence Diagrams*, J. Experimental Algorithmics, 22 (2017), pp. 1–20, https://doi.org/10.1145/3064175.

[54] A. M. Kilpatrick, A. A. Chmura, D. W. Gibbons, R. C. Fleischer, P. P. Marra, and P. Daszak, *Predicting the global spread of H5N1 avian influenza*, Proc. Natl. Acad. Sci., 103 (2006), pp. 19368–19373, https://doi.org/10.1073/pnas.0609227103.

[55] B. O. Koopman, *On distributions admitting a sufficient statistic*, Trans. Amer. Math. Soc., 39 (1936), pp. 399–409.

[56] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Stat., 22 (1951), pp. 79–86, https://projecteuclid.org:443/euclid.aoms/1177729694.

[57] R. Kwitt, S. Huber, M. Niethammer, W. Lin, and U. Bauer, *Statistical Topological Data Analysis – A Kernel Perspective*, in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, 2015, pp. 3070–3078.

[58] L. Le Cam, *Sufficiency and approximate sufficiency*, Ann. Math. Stat., 35 (1964), pp. 1419–1455, http://www.jstor.org/stable/2238284.

[59] B. Lin, A. Monod, and R. Yoshida, *Tropical Foundations for Probability & Statistics on Phylogenetic Tree Space*, preprint, arXiv, (2018), arXiv:1805.12400.

[60] M. I. Love, J. B. Hogenesch, and R. A. Irizarry, *Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation*, Nature Biotechnology, 34 (2016), pp. 1287–1291, http://dx.doi.org/10.1038/nbt.3682.

[61] I. MALJKOVIC BERRY, M. C. MELENDREZ, T. LI, A. W. HAWKSWORTH, G. T. BRICE, P. J. BLAIR, E. S. HALSEY, M. WILLIAMS, S. FERNANDEZ, I.-K. YOON, L. D. EDWARDS, R. KUSCHNER, X. LIN, S. J. THOMAS, AND R. G. JARMAN, *Frequency of influenza H3N2 intra-subtype reassortment: Attributes and implications of reassortant spread*, BMC Biology, 14 (2016), p. 117, https://doi.org/10.1186/s12915-016-0337-3.

[62] N. MARSHALL, L. PRIYAMVADA, Z. ENDE, J. STEEL, AND A. C. LOWEN, *Influenza virus reassortment occurs with high frequency in the absence of segment mismatch*, PLoS Pathog., 9 (2013), e1003421.

[63] Y. MILEYKO, S. MUKHERJEE, AND J. HARER, *Probability measures on the space of persistence diagrams*, Inverse Problems, 27 (2011), p. 124007.

[64] J. NEYMAN, *Su un teorema concernente le cosiddette statistiche sufficienti*, Giornale Dell'Istituto Italiano degli Attuari, 6 (1935), pp. 320–334.

[65] T. H. NGUYEN, V. T. THAN, H. D. THANH, V.-K. HUNG, D. T. NGUYEN, AND W. KIM, *Intersubtype Reassortments of H5N1 Highly Pathogenic Avian Influenza Viruses Isolated from Quail*, PLoS ONE, 11 (2016), pp. 1–15, https://doi.org/10.1371/journal.pone.0149608.

[66] N. OTTER, M. A. PORTER, U. TILLMANN, P. GRINDROD, AND H. A. HARRINGTON, *A roadmap for the computation of persistent homology*, EPJ Data Science, 6 (2017), p. 17, https://doi.org/10.1140/epjds/s13688-017-0109-5.

[67] L. PACHTER AND B. STURMFELS, *Algebraic Statistics for Computational Biology*, Vol. 13, Cambridge University Press, Cambridge, 2005.

[68] E. PARADIS, J. CLAUDE, AND K. STRIMMER, *APE: Analyses of Phylogenetics and Evolution in R language*, Bioinformatics, 20 (2004), pp. 289–290, https://doi.org/10.1093/bioinformatics/btg412.

[69] K. PARTHASARATHY, *Probability and mathematical statistics: A series of monographs and textbooks*, in Probability Measures on Metric Spaces, Probability and Mathematical Statistics: A Series of Monographs and Textbooks, Academic Press, 1967, ii p, https://doi.org/10.1016/B978-1-4832-0022-4.50001-6.

[70] J. Á. PATIÑO-GALINDO, M. TORRES-PUENTE, M. A. BRACHO, I. ALASTRUÉ, A. JUAN, D. NAVARRO, M. J. GALINDO, C. GIMENO, E. ORTEGA, AND F. GONZÁLEZ-CANDELAS, *Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain*, PLoS ONE, 12 (2017), e0171062.

[71] J. Á. PATIÑO-GALINDO, M. TORRES-PUENTE, M. A. BRACHO, I. ALASTRUÉ, A. JUAN, D. NAVARRO, M. J. GALINDO, D. OCETE, E. ORTEGA, C. GIMENO, J. BELDA, V. DOMÍNGUEZ, R. MORENO, AND F. GONZÁLEZ-CANDELAS, *The molecular epidemiology of HIV-1 in the Comunidad Valenciana (Spain): analysis of transmission clusters*, Scientific Reports, 7 (2017), p. 11584, https://doi.org/10.1038/s41598-017-10286-1.

[72] J. A. PEREA AND G. CARLSSON, *A Klein-Bottle-Based Dictionary for Texture Representation*, Int. J. Comput. Vision, 107 (2014), pp. 75–97, https://doi.org/10.1007/s11263-013-0676-2.

[73] M. PÉREZ-LOSADA, M. ARENAS, J. C. GALÁN, F. PALERO, AND F. GONZÁLEZ-CANDELAS, *Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences*, Infection, Genetics and Evolution, 30 (2015), pp. 296–307, http://www.sciencedirect.com/science/article/pii/S156713481400478X.

[74] E. J. G. PITMAN, *Sufficient statistics and intrinsic accuracy*, Math. Proc. Cambridge Philosoph. Soc., 32 (1936), pp. 567–579.

[75] J. REININGHAUS, S. HUBER, U. BAUER, AND R. KWITT, *A stable multi-scale kernel for topological machine learning*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4741–4748.

[76] A. ROBINSON AND K. TURNER, *Hypothesis testing for topological data analysis*, J. Appl. Comput. Topology, 1 (2017), pp. 241–261, https://doi.org/10.1007/s41468-017-0008-7.

[77] F. RONQUIST, M. TESLENKO, P. VAN DER MARK, D. L. AYRES, A. DARLING, S. HÖHNA, B. LARGET, L. LIU, M. A. SUCHARD, AND J. P. HUELSENBECK, *MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space*, Syst. Biol., 61 (2012), pp. 539–542, https://doi.org/10.1093/sysbio/sys029.

[78] N. SAITOU AND M. NEI, *The Neighbor-Joining Method: A new method for reconstructing phylogenetic trees*, Molecular Biol. Evolution, 4 (1987), pp. 406–425, https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[79] D. L. SWOFFORD, *PAUP*: Phylogenetic Analysis Using Parsimony (and other methods)* 4.0.*b*5, 2001.

[80] THE WELLCOME TRUST CASE CONTROL CONSORTIUM, *Genome-wide association study of* 14,000 *cases of seven common diseases and* 3,000 *shared controls*, Nature, 447 (2007), pp. 661–6678, http://dx.doi.org/10.1038/nature05911.

[81] H. TIAN, S. ZHOU, L. DONG, T. P. VAN BOECKEL, Y. CUI, S. H. NEWMAN, J. Y. TAKEKAWA, D. J. PROSSER, X. XIAO, Y. WU, B. CAZELLES, S. HUANG, R. YANG, B. T. GRENFELL, AND B. XU, *Avian influenza H*5*N*1 *viral and bird migration networks in Asia*, Proc. Natl. Acad. Sci., 112 (2015), pp. 172–177, http://www.pnas.org/content/112/1/172.abstract.

[82] M. C. TURCHIN, C. W. CHIANG, C. D. PALMER, S. SANKARARAMAN, D. REICH, J. N. HIRSCHHORN, AND G. I. OF ANTHROPOMETRIC TRAITS (GIANT) CONSORTIUM, *Evidence of widespread selection on standing variation in Europe at height-associated SNPs*, Nature Genetics, 44 (2012), 1015.

[83] K. TURNER, Y. MILEYKO, S. MUKHERJEE, AND J. HARER, *Fréchet means for distributions of persistence diagrams*, Discrete Comput. Geometry, 52 (2014), pp. 44–70.

[84] K. TURNER, S. MUKHERJEE, AND D. M. BOYER, *Persistent homology transform for modeling shapes and surfaces*, Inform. Inference, 3 (2014), pp. 310–344.

[85] A. W. VAN DER VAART, *Asymptotic Statistics*, Vol. 3, Cambridge University Press, Cambridge, 2000.

[86] C. J. WILLER, E. K. SPELIOTES, R. J. LOOS, S. LI, C. M. LINDGREN, I. M. HEID, S. I. BERNDT, A. L. ELLIOTT, A. U. JACKSON, AND C. LAMINA, *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation*, Nature Genetics, 41 (2009), p. 25.

[87] M. WOROBEY, T. D. WATTS, R. A. MCKAY, M. A. SUCHARD, T. GRANADE, D. E. TEUWEN, B. A. KOBLIN, W. HENEINE, P. LEMEY, AND H. W. JAFFE, 1970*s and 'Patient 0' HIV-*1 *genomes illuminate early HIV/AIDS history in North America*, Nature, 539 (2016), pp. 98–101.

[88] S. YAN AND G. WU, *Possible reason for cross-species and cross-subtype reassortment in polymerase basic protein* 2 *from influenza A virus*, Protein and Peptide Letters, 18 (2011), pp. 434–439.

[89] J. YANG, T. FERREIRA, A. P. MORRIS, S. E. MEDLAND, P. A. MADDEN, A. C. HEATH, N. G. MARTIN, G. W. MONTGOMERY, M. N. WEEDON, AND R. J. LOOS, *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits*, Nature Genetics, 44 (2012), p. 369.

[90] Z. YANG, *A space-time process model for the evolution of DNA sequences*, Genetics, 139 (1995), pp. 993–1005.

[91] A. J. ZOMORODIAN AND G. CARLSSON, *Computing persistent homology*, Discrete Comput. Geometry, 33 (2005), pp. 249–274.