



OPEN

# A universal null-distribution for topological data analysis

Omer Bobrowski<sup>1,2✉</sup> & Primoz Skraba<sup>2✉</sup>

One of the most elusive challenges within the area of topological data analysis is understanding the distribution of persistence diagrams arising from data. Despite much effort and its many successful applications, this is largely an open problem. We present a surprising discovery: normalized properly, persistence diagrams arising from random point-clouds obey a universal probability law. Our statements are based on extensive experimentation on both simulated and real data, covering point-clouds with vastly different geometry, topology, and probability distributions. Our results also include an explicit well-known distribution as a candidate for the universal law. We demonstrate the power of these new discoveries by proposing a new hypothesis testing framework for computing significance values for individual topological features within persistence diagrams, providing a new quantitative way to assess the significance of structure in data.

Topological Data Analysis (TDA) focuses on extracting structural information from data, in order to enhance their processing in statistics and machine learning. This field has been rapidly developing over the past two decades, bringing together mathematicians, statisticians, computer scientists, engineers, and data scientists. The motivation behind TDA is that topological methods are highly versatile, coordinate-free, and robust to various deformations<sup>1</sup>. Topological methods have been applied successfully in numerous applications, in areas such as neuroscience<sup>2–4</sup>, medicine and biology<sup>5–8</sup>, material science<sup>9,10</sup>, dynamical systems<sup>11,12</sup>, and cosmology<sup>13,14</sup>.

One of the key challenges in TDA is to distinguish between “signal”—meaningful structures underlying the data, and “noise”—features that arise from the local randomness and inaccuracies within the data<sup>15–17</sup>. The most prominent solution developed in TDA to address this issue is *persistent homology*. Briefly, it identifies structures such as holes and cavities (“air pockets”) formed by the data, and records the scales at which they are created and terminated (*birth* and *death*, respectively). The common practice in TDA has been to use this birth-death information to assess the statistical significance of topological features<sup>18–21</sup>. However, research so far has yet to provide an approach which is generic, robust, and theoretically justified. A parallel line of research has been the theoretical probabilistic analysis of persistent homology generated by random data, as means to establish a null-distribution. While this direction has been fruitful<sup>22–25</sup>, its use in practice has been limited. The main gap between theory and practice is that these studies indicate that the distribution of noise in persistent homology: (a) does not have a simple closed-form description, and (b) strongly depends on the model generating the point-cloud.

Our main goal in this paper is to refute the last premise, and to make the case that the distribution of noise in persistent homology of random point-clouds is in fact *universal*. Specifically, we claim that the limiting distribution of *persistence values* (measured using the death/birth ratio) is independent of the model generating the point-cloud. This result is loosely analogous to the central limit theorem, where sums of many different types of random variables always converge to the normal distribution. The emergence of such universality for persistence diagrams is highly surprising.

We support our universality statements by an extensive body of experiments, including point-clouds generated by different geometries, topologies, and probability distributions. These include simulated data as well as data from real-world applications (image processing, signal processing, and natural language processing). Our main goal here is to introduce the unexpected behavior of statistical universality in persistence diagrams, in order to initiate a shift of paradigm in stochastic topology that will lead to the development of a new theory. Developing this new theory, and proving the conjectures made here, is anticipated to be an exciting yet a challenging long journey, and is outside the scope of this paper. Based on our universality conjectures, we develop a powerful hypothesis testing framework for persistence diagrams, allowing us to compute numerical significance measures for individual features using very few assumptions on the underlying model.

<sup>1</sup>Viterbi Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology, Haifa, Israel. <sup>2</sup>School of Mathematical Sciences, Queen Mary University of London, London, UK. ✉email: omer@ee.technion.ac.il; p.skraba@qmul.ac.uk

## Persistent homology for point-clouds

In order to approximate the structure formed by point-clouds, a common practice in TDA is to construct a special type of hypergraphs known as *geometric simplicial complexes*, whose faces are determined by the spatial configuration of the points. We consider the two most commonly used constructions— the **Čech** and the **Vietoris-Rips** complexes, both are parameterized by a scale parameter  $r$  (see “Methods” section). The *homology groups* ( $H_k$ ) of a simplicial complex capture structural information about the complex. Loosely speaking,  $H_0$  contains information about connected components,  $H_1$ —about closed loops surrounding holes,  $H_2$ —about closed surfaces surrounding cavities (“air pockets”). Generally, we say that  $H_k$  describes ‘ $k$ -dimensional cycles.’ For more details, see “Methods”.

For both the Čech and Rips complexes,  $H_k$  is highly sensitive to the choice of the scale parameter  $r$ . To overcome this, *persistent homology* considers the entire range of scales, tracking the evolution of  $k$ -cycles as the value of  $r$  increases. In this process (called a ‘filtration’), cycles are created (born) and later filled-in (die). This information is most often represented via *persistence diagrams*, 2-dimensional scatter plots, where each point  $p = (b, d) \in \mathbb{R}^2$  represents the *birth* and *death* times (scales) of a single cycle in the filtration, see Fig. 1. We denote the persistence diagram corresponding to  $H_k$  by  $\text{dgm}_k$ .

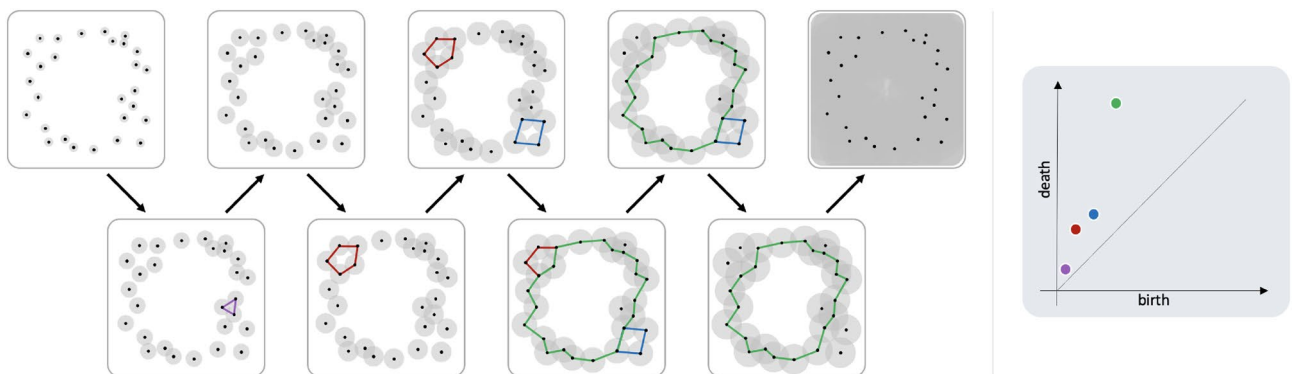
The original motivation for using persistent homology is to detect meaningful structures emerging in data. The simplest approach is by looking for points  $p \in \text{dgm}_k$  far from the diagonal ( $d = b$ ). These points represent cycles with a long lifetime ( $d - b$ ), which are “significant”, whereas points near the diagonal are due to the noisy nature of the samples (see Fig. 1). While this approach is intuitive, justifying it theoretically and providing *quantitative statements*, are among the greatest challenges in the field. For geometric complexes, a strong case can be made<sup>26</sup> that it is better to use the ratio  $\pi = d/b$  to discriminate between signal and noise in  $\text{dgm}_k$  ( $k > 0$ ), since the ratio  $\pi$  is (a) scale invariant, and (b) more robust to outliers. For further discussion see “Methods”.

**Noise distribution in persistence diagrams.** We assume that every persistence diagram can be decomposed into  $\text{dgm}_k^s \cup \text{dgm}_k^n$ , corresponding to the signal and noise parts. The signal corresponds to meaningful topological features which are latent in the data (e.g., points sampled near an annulus will always contain a hole in the middle). The noise part consists of features which are the result of randomness in the data. For formal definitions and a discussion, see Sect. 2 of the Supplementary Information. The fundamental challenge is to decide for each feature  $p$  in the diagram, whether  $p \in \text{dgm}_k^s$  or  $p \in \text{dgm}_k^n$ , and provide quantitative guarantees on this decision. The distribution of  $\text{dgm}_k^n$  thus serves as a null-distribution, and revealing it would enable us to use powerful hypothesis testing methods.

The probabilistic analysis of  $\text{dgm}_k^n$  has been fruitful<sup>22,23,25–27</sup>, a more detailed discussion is provided in Sect. 1 of the Supplementary Information. However, while the mathematical theory is quite rich, translating it into statistical tools has lagged behind. The two main reasons are: (a) these results show that various limits exist, but in most cases without any explicit description, (b) these limits strongly depend on the underlying distribution. A more practical approach<sup>13</sup> models persistence diagrams using Gibbs measures, whose parameters can be estimated from the data. Due to the limitations imposed by the theoretical analysis, the statistical literature for *topological inference* is mostly based on the premise that the distribution of persistence diagrams is inaccessible. A prominent approach in this case is based on statistical bootstrap<sup>18,19,28</sup>. Other useful methods include distance to measures<sup>29</sup>, witness complexes<sup>30</sup>, and multi-cover bifiltrations<sup>31,32</sup>.

## Results

**The distribution of persistent cycles.** Let  $\mathcal{S}$  be a  $d$ -dimensional metric measure space, and let  $\mathbf{X}_n = (X_1, \dots, X_n) \in \mathcal{S}^n$  be a sequence of random variables (points), whose joint probability law is denoted by  $\mathbb{P}_n$ . Let  $\mathbb{P} = (\mathbb{P}_n)_{n=1}^\infty$ , and denote  $\mathbb{S} = (\mathcal{S}, \mathbb{P})$ , which we refer to as the *sampling model*. Fix a filtration type  $\mathcal{T}$  (e.g.,



**Figure 1.** A persistence diagram generated by a point-cloud sampled from a noisy circle. The shape studied here is the union of balls (in gray), which is equivalent to the Čech complex. As the radius  $r$  increases, four different 1-cycles appear (in different color). The radii at which each cycle appears and filled are recorded by the persistence diagram on the right. The green cycle stands out, being the furthest from the diagonal, and indeed this cycle is a only “real” feature of the sample, while the others can be considered as “noise”.

Čech or Rips), and a homological degree  $k > 0$ , and consider the  $k$ -th *noise* persistence diagram  $\text{dgm}_k^N(\mathbf{X}_n; \mathcal{T})$ , which in short we denote by  $\text{dgm}_k$ . We study the distribution of the random persistence values  $\{\pi(p)\}_{p \in \text{dgm}_k}$  (where  $\pi(p) = \text{death}(p)/\text{birth}(p)$ ), and refer to them as the  $\pi$ -values of the diagram. Theoretical analysis shows that the largest  $\pi$ -value (death / birth ratio) of points in the noise ( $\text{dgm}_k^N$ ) is  $o((\log n)^{1/k})^{26}$ , while the  $\pi$ -values of the signal features ( $\text{dgm}_k^S$ ) are  $\Theta(n^{1/d})^{33}$ . Thus, the  $\pi$ -values provide a strong separation (asymptotically) between signal and noise in persistence diagrams. We stress that in this paper we study the **entire** ensemble of persistence values, not only the maximal ones.

*Weak universality.* We begin by considering the case where  $\mathbb{P}_n$  is a product measure, and the points  $X_1, \dots, X_n$  are iid (independent and identically distributed). Given  $\text{dgm}_k$  as defined above, denote the empirical measure of all  $\pi$ -values

$$\Pi_n = \Pi_n(\mathbb{S}, \mathcal{T}, k) := \frac{1}{|\text{dgm}_k|} \sum_{p \in \text{dgm}_k} \delta_{\pi(p)},$$

where  $\delta_x$  is the Dirac delta-measure at  $x$ . In Fig. 2 we present the CDF of  $\Pi_n$  for the Čech complex with various choices of  $\mathbb{S}$  and  $k$ . Similar plots are available for the Rips complex in Sect. 3 of the Supplementary Information. We observe that if we fix  $d$  (dimension of  $\mathcal{S}$ ),  $\mathcal{T}$ , and  $k$ , then the resulting CDF depends on neither the space  $\mathcal{S}$  nor the distribution  $\mathbb{P}_n$ . This leads to our first conjecture.

**Conjecture 1** Fix  $d, \mathcal{T}$ , and  $k > 0$ . For any  $\mathbb{S} \in \mathcal{I}_d$ ,

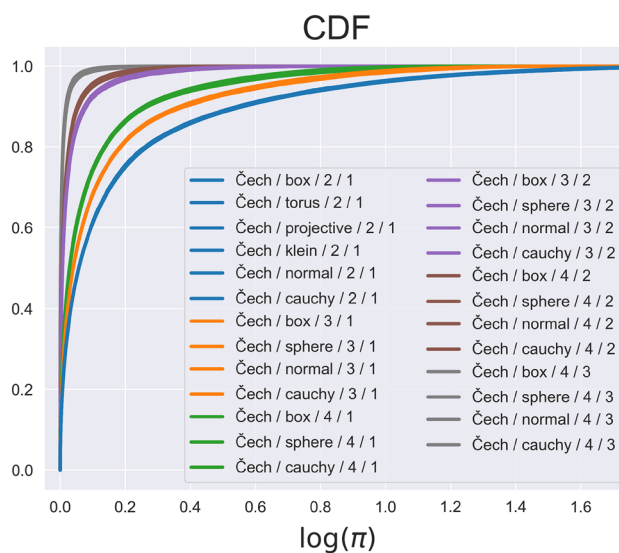
$$\lim_{n \rightarrow \infty} \Pi_n = \Pi_{d, \mathcal{T}, k}^*$$

where  $\Pi_{d, \mathcal{T}, k}^*$  is a probability distribution on  $[1, \infty)$ .

The precise notion of convergence and the extent of the class  $\mathcal{I}_d$  are to be determined as future work. We conjecture that  $\mathcal{I}_d$  is quite large. In our experiments, the space  $\mathcal{S}$  varied across a wide range of manifolds and other spaces. The distribution  $\mathbb{P}_n$  is continuous and iid, but otherwise fairly generic (possibly even without moments, see the Cauchy example in Fig. 2). We name this phenomenon “weak universality”, since on one hand the limit is independent of  $\mathbb{S}$  (hence, universal), while on the other hand it does depend on  $d, \mathcal{T}, k$  and the iid assumption. This is in contrast to the results we discuss next.

*Strong universality.* The following procedure was discovered partly by chance. While non-intuitive, the results are striking. Given a random persistence diagram  $\text{dgm}_k$ , for each  $p \in \text{dgm}_k$  apply the transformation

$$\ell(p) := A \log \log(\pi(p)) + B, \tag{1}$$



**Figure 2.** The distribution of  $\pi$ -values in the Čech complex. We take the empirical CDFs of the  $\pi$ -values (log-scale), computed from various iid samples. The legend format is  $\mathcal{T}/\mathbb{P}/d/k$ , where  $\mathcal{T}$  is the complex type,  $\mathbb{P}$  is the probability distribution,  $d$  is the dimension of the sampling space, and  $k$  is the degree of homology computed. By ‘box’, ‘torus’, ‘sphere’, ‘projective(-plane)’, and ‘Klein(-bottle)’ we refer to the uniform distribution on the respective space or its most natural parametrization, while ‘normal’ and ‘cauchy’ refer to the corresponding non-uniform distributions. See “Methods” and Sect. 3 in the Supplementary Information for further details.

where

$$A = \begin{cases} 1 & \mathcal{T} = \text{Rips}, \\ 1/2 & \mathcal{T} = \text{Čech}, \end{cases} \quad B = -\lambda - A\bar{L}, \tag{2}$$

and where  $\bar{L} = \frac{1}{|\text{dgm}_k|} \sum_{p \in \text{dgm}_k} \log \log (\pi(p))$  and  $\lambda$  is the Euler-Mascheroni constant ( $= 0.5772156649 \dots$ ). We refer to the set  $\{\ell(p)\}_{p \in \text{dgm}_k}$  as the  $\ell$ -values of the diagram. In Fig. 3 we present the empirical CDFs of the  $\ell$ -values, as well as the kernel density estimates for their PDFs, for all the iid samples that were included in Fig. 2. The plots for the Rips complex are similar, and can be found in Sect. 3 of the Supplementary Information. We observe that all the different settings  $(\mathbb{S}, \mathcal{T}, k)$  yield *exactly* the same distribution under the transformation given by (1). We refer to this phenomenon as “strong universality”.

While strong universality for iid point-clouds is by itself a very unexpected and useful behavior, a natural question is how generally it applies in other scenarios. In Fig. 3 we also include a selection of non-iid samples and real-data (see Experimental results and Methods for details). While the distribution of the  $\pi$ -values for these models is vastly different than the iid case, all of these examples exhibit the same strong universality behavior.

To summarize, our experiments highly indicate that persistent  $\ell$ -values have a universal limit for a wide class of sampling models  $\mathbb{S}$ , denoted by  $\mathcal{U}$ . For our main conjecture, we consider the empirical measure of *all*  $\ell$ -values,

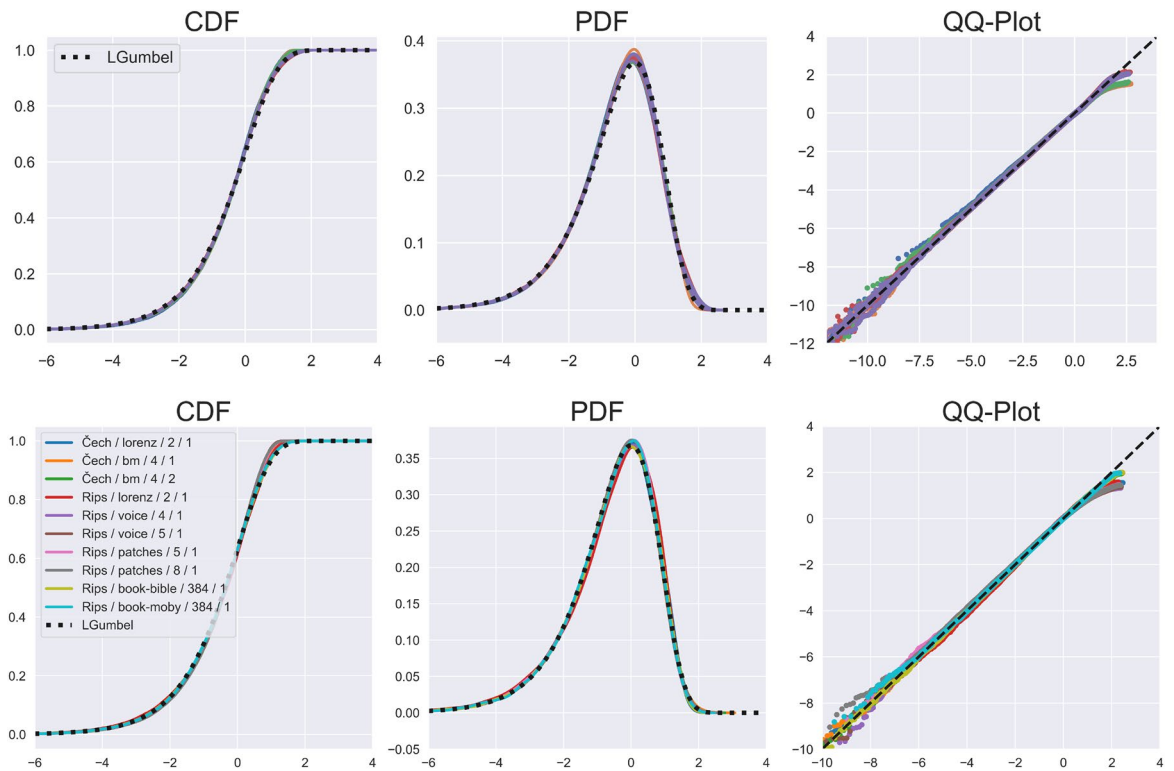
$$\mathcal{L}_n = \mathcal{L}_n(\mathbb{S}, \mathcal{T}, k) := \frac{1}{|\text{dgm}_k|} \sum_{p \in \text{dgm}_k} \delta_{\ell(p)}.$$

**Conjecture 2** For any  $\mathbb{S} \in \mathcal{U}$ ,  $\mathcal{T}$ , and  $k \geq 1$ ,

$$\lim_{n \rightarrow \infty} \mathcal{L}_n = \mathcal{L}^*,$$

where  $\mathcal{L}^*$  is independent of  $\mathbb{S}$ ,  $\mathcal{T}$ , and  $k$ .

Observe that in this Conjecture, the only dependence on the distribution generating the point-cloud is in the value of  $B$  (2) (similar to the role the mean and the variance play in the central limit theorem). In Sect. 5 of the Supplementary Information, we examine the value of  $B$  for different iid settings. As suggested by Conjecture 1, our experiments confirm that the value of  $B$  (for the iid case) depends on  $d, \mathcal{T}, k$ , but is otherwise independent of  $\mathbb{S}$ . Revealing the exact relationship between all parameters remains future work.



**Figure 3.** The distribution of  $\ell$ -values. (top) All the iid samples included in Fig. 2 (26 curves). (bottom) A selection non-iid and real-data point clouds. The left column shows the empirical CDF, and the middle column is the kernel density estimate for the PDF, with the LGumbel distribution shown as the dashed line. The right column shows the QQ plots compared this distribution.

**Remark** We note that models with homogeneous spacing between the points, such as perturbed lattice models, or repulsive point processes, do not follow Conjecture 2. See Sect. 2.4 in the Supplementary Information.

*A candidate distribution.* A natural question is whether the observed limiting distribution  $\mathcal{L}^*$  is a familiar one, and in particular, if it has a simple expression. Surprisingly, it seems that the answer might be yes. We denote the *left-skewed Gumbel distribution* by LGumbel, whose CDF and PDF are given by

$$F(x) = 1 - e^{-e^x}, \quad \text{and} \quad f(x) = e^{x-e^x}. \quad (3)$$

The expected value of this distribution is the Euler-Mascheroni constant ( $\lambda$ ) used in (2). In Fig. 3, the black dashed lines represent the CDF and PDF of the LGumbel distribution. In addition, the right column presents the QQ-plots of all the different models compared to the LGumbel distribution. These plots provide very strong evidence for the validity of our final conjecture.

**Conjecture 3**  $\mathcal{L}^* = \text{LGumbel}$ .

**Remark** The Gumbel distribution often emerges as the limit of extreme value distributions (i.e., minima or maxima). We wish to emphasize that the limiting LGumbel distribution in Conjecture 3 describes the **entire ensemble** of all  $\ell$ -values and so it **does not** describe any extreme values. Consequently, while there may exist an indirect connection to extreme value theory, it is by no means evident or straightforward. The appearance of the LGumbel distribution in this context, is therefore quite surprising.

**Remark** The plots in Fig. 3 exhibit a remarkable fit with the LGumbel distribution, with one minor exception observed in the deviation of the tails of the QQ-plots. This slight deviation can be attributed to slow convergence rates. Existing theoretical results<sup>26</sup> indicate that the largest  $\pi$ -value tends to infinity with  $n$ , but its growth rate is logarithmic. Consequently, the largest  $\ell$ -value is of order  $\log \log \log(n)$ , and so particularly the tails will exhibit a slow rate of convergence. Since the point-clouds we are able to process consist of at most millions of points, our ability to accurately capture the tail of the distribution is quite limited. It is important to note that this limitation holds regardless of the validity of Conjecture 3. While the limiting distribution has a non-compact support, our experiments cover a restricted range of death/birth values. Therefore, if we draw the QQ-plot against any other distribution with non-compact support, we will observe similar deviations.

**Experimental results.** We present a large body of experimental evidence collected to support our conjectures. As the results do not seem to be significantly impacted by the choice of  $n$ , we leave this detail to the Supplementary Information (Section 3). Complementing the statistical plots presented here and in the Supplementary Information, we also performed a Kolmogorov-Smirnov goodness-of-fit test for our entire collection of point-clouds. The details are provided in Sect. 3.4 of the Supplementary Information. The test did not detect significant difference between the distribution of  $\ell$ -values and the LGumbel distribution in any of the point-clouds, providing further support for the validity of Conjectures 1, 2, 3.

*Sampling from iid distributions.* We began by considering samples from the uniform distribution on various compact manifolds, with diverse geometry and topology. Next, we tested non-uniform distributions in  $\mathbb{R}^d$ , by taking products of well-known distributions on  $\mathbb{R}$ . We attempted to test a wide range of settings. The beta distribution has a compact support ( $[0, 1]$ ), while the normal and Cauchy distributions are supported on  $\mathbb{R}$ . The standard normal distribution is an obvious choice and Cauchy was chosen as a heavy-tailed distribution without moments. Finally, we considered more complex models—sampling from the configuration space of a closed five-linkage, and stratified spaces (intersecting manifolds of different dimensions). The results for many of the experiments are in Figs. 2 and 3 (see Sect. 3 of the Supplementary Information for the complete set of experiments). All of the iid sampling models we tested support Conjectures 1, 2, 3.

*Non-iid distributions.* To better understand the extent of universality, as well as to consider more realistic models, we tested more complex cases. We tested two vastly different models: sampling points from the path of a  $d$ -dimensional Brownian motion, and a discrete-time sample of the Lorenz dynamical system—a well-studied chaotic system. The results in Fig. 3 confirm that these non-iid models exhibit strong universality as well. Surprisingly, the results for the Brownian motion demonstrate the best fit with the LGumbel distribution, among all the settings we tested (see Figs. 10 and 11 in the Supplementary Information). This could be related to the fractal, or self-similarity, behavior of the Brownian motion, but remains a topic for future study.

*Testing on real data.* The most important test for Conjectures 2 and 3 is with real world data. We tested three different examples (see Methods for mode details). (1) Natural images: We sampled  $3 \times 3$  patches from natural gray-scale images taken from van Hateren and van der Schaaf dataset<sup>34</sup>. We applied the dimension reduction procedure proposed by Lee et al.<sup>35</sup>, which results in a point-cloud on a 7-dimensional sphere embedded in  $\mathbb{R}^8$ . We tested both the 7-dimensional point-cloud, as well as its lower-dimensional projections. (2) Audio recording: We applied the time-delay embedding transformation<sup>36</sup> to an arbitrary speech recording to create a  $d$ -dimensional point-cloud. (3) Sentence embeddings: We used a pretrained sentence transformer<sup>37</sup>, to convert the entire

text in a book into a 384-dimensional point-cloud. Our experiments using both the Čech and Rips complexes, show a remarkable matching to the universal distribution (see Fig. 3 for a subset).

**Application: hypothesis testing.** Based on Conjectures 2–3, we present a hypothesis testing framework for individual cycles in persistence diagrams. We address finite and infinite cycles separately.

*Finite cycles.* Given a persistence diagram  $\text{dgm} = \{p_1, \dots, p_m\}$ , our goal is to determine for each point  $p_i$  whether it is signal or noise. This can be modelled as a multiple hypothesis testing problem with the  $i$ -th null-hypothesis, denoted  $H_0^{(i)}$ , is that  $p_i$  is a noisy cycle. Assuming Conjectures 2 and 3, we can formalize the null hypothesis in terms of the  $\ell$ -values (1) as

$$H_0^{(i)} : \ell(p_i) \sim \text{LGumbel}.$$

In other words, cycles that deviate significantly from the LGumbel distribution should be declared as *signal*. If the observed persistence  $\ell$ -value is  $x$ , then its corresponding  $p$ -value is computed via

$$p\text{-value}_i = \mathbb{P}(\ell(p_i) \geq x \mid H_0^{(i)}) = e^{-e^x}. \tag{4}$$

Since we are testing multiple cycles simultaneously, we applied the Bonferroni correction to the  $p$ -values, which sufficed for our experiments. The signal part of a diagram (for significance level  $\alpha$ ) can thus be recovered, via

$$\text{dgm}_k^s(\alpha) = \left\{ p \in \text{dgm}_k : e^{-e^{\ell(p)}} < \frac{\alpha}{|\text{dgm}_k|} \right\}.$$

*Infinite cycles.* Computing persistent homology for an entire filtration is often intractable. The common practice is to fix a threshold  $\tau$ , and compute  $\text{dgm}_k(\tau)$  for the partial filtration. This often introduces cycles that are “infinite”—i.e., born prior to  $\tau$ , but die after  $\tau$ . The question we address here is how to *efficiently* determine whether such cycles are statistically significant. Let  $p = (b, d) \in \text{dgm}_k(\tau)$  be an infinite cycle, i.e.,  $b \leq \tau$  is known and  $d > \tau$  is *unknown*. While we do not know  $\ell(p)$ , we observe that  $\ell(p) > \tau/b$ , which gives an upper bound for the  $p$ -value,

$$p\text{-value}_i < p\text{-value}_i(\tau) := e^{-e^{\tau/b}}.$$

If  $p\text{-value}_i(\tau)$  is below the required significance value (e.g.  $\alpha/|\text{dgm}_k(\tau)|$ ), we can declare  $p$  as significant, despite not knowing the true death-time. Otherwise, we can determine the minimal value  $\tau^*$  required so that  $p\text{-value}_i(\tau^*)$  is below the significance value. We then compute  $\text{dgm}_k(\tau^*)$ , and if the cycle represented by  $p$  remains infinite (i.e.  $d > \tau^*$ ), we declare it significant. We observe that for measuring significance, we do not need to know the exact value of  $d$ , only whether it is smaller or larger than  $\tau^*$ , and we need only to compute the filtration up to  $\tau^*$ , rather than the actual death time  $d$ . The key point is that the death time  $d$  may be much larger than  $\tau^*$ .

The procedure we just described works well for studying a single infinite cycle. However, it is likely that  $\text{dgm}_k(\tau)$  contains multiple infinite cycles. Moreover, increasing the threshold may result in new infinite cycles emerging as well. We therefore propose the iterative procedure described in Algorithm 1. Briefly, at every step the algorithm picks one infinite cycle, and chooses the next threshold  $\tau$  so that we can determine if it is significant or not. The value  $\pi_{\min}(x)$  in the Algorithm 1, is the minimum  $\pi$ -value required so that the resulting  $p$ -value (4) is smaller than  $x$ . Formally,

$$\pi_{\min}(x) = \ell^{-1}(F^{-1}(1-x)) = \ell^{-1}(\log \log(1/x)),$$

where  $F$  is the CDF of the LGumbel distribution. In the algorithm, we choose the earliest-born infinite cycle ( $\min(I)$ ), while we could have chosen the latest-born ( $\max(I)$ ), or any intermediate value. This choice represents a trade-off between the number of iterations needed and the overestimation of  $\tau$ . Choosing the earliest born cycle results in the smallest threshold, but with potentially more iterations, while choosing the last cycle will have fewer iterations with a possible overestimation of  $\tau$ .

Algorithm 1 Finding the threshold for infinite cycles	
$\tau \leftarrow \tau_0$	
<b>do</b>	
$D \leftarrow \text{dgm}_k(\tau)$	
$I \leftarrow \{(b, d) \in D, d = \infty, \text{ and } \tau/b < \pi_{\min}(\alpha/ D )\}$	
$\tau \leftarrow \begin{cases} \min(I) \cdot \pi_{\min}(\alpha/ D ) & I \neq \emptyset \\ \tau & I = \emptyset \end{cases}$	
<b>while</b> $ I  > 0$	
<b>return</b> $\tau$	

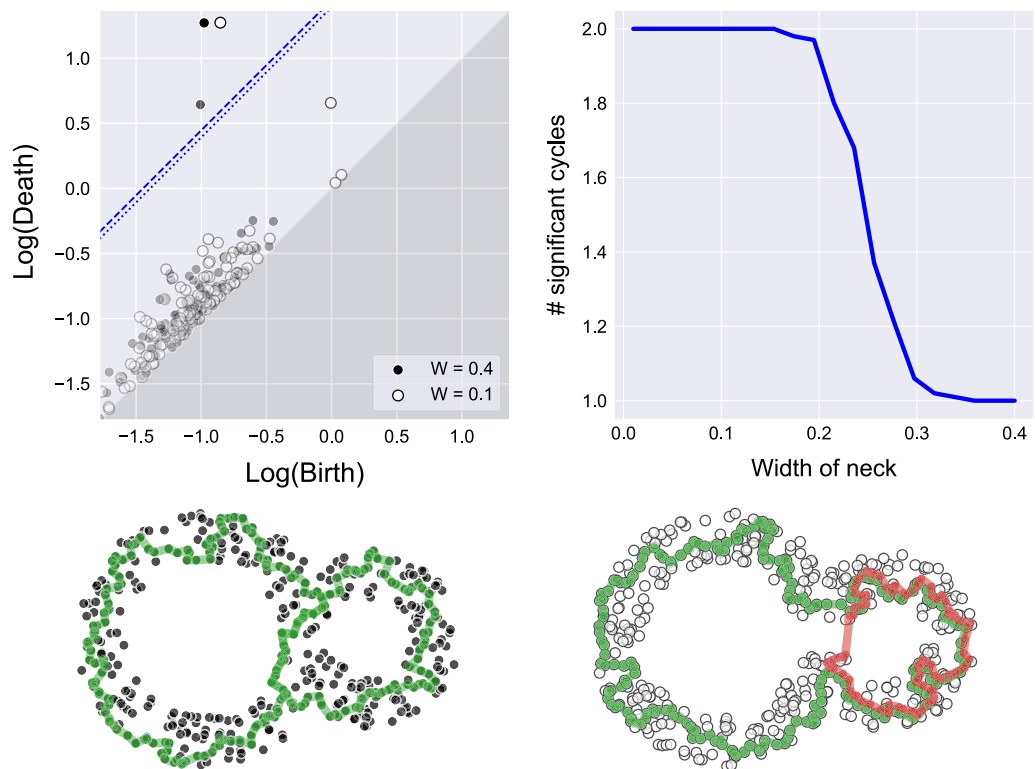
*Examples.* We present two examples for our hypothesis testing framework. In all our experiments, we set the desired significance level to be  $\alpha = 0.05$ .

- Computing  $p$ -values:** We begin with a toy example, sampling 1000 points on an 8-shape (a wedge of circles) in  $\mathbb{R}^2$  (see Fig. 4), where we vary the width of the neck. We expect one cycle to always be significant (the outer one), but the significance of the second cycle depends on the width of the neck. For each width value we computed the persistence diagram, and checked how many cycles were significant (i.e.,  $p$ -value  $< \frac{\alpha}{|\text{dgm}_k|}$ ). The results are presented in Fig. 4. For very small neck widths, our sample is indistinguishable from the 8-shape where the neck is fully connected, and hence both cycles are detected.

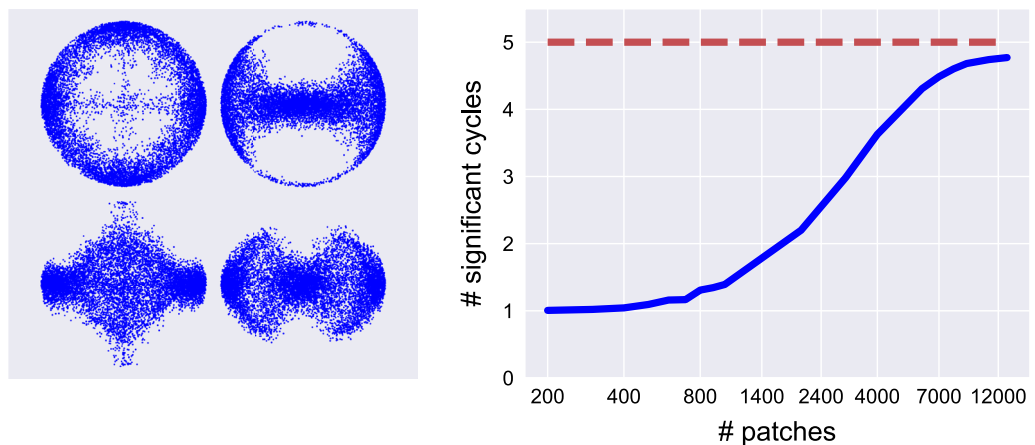
Next, we apply this method to a real-world dataset, specifically the van-Hateren natural image database mentioned earlier. The main claim by de-Silva & Carlsson<sup>30</sup> is that the space of  $3 \times 3$  patches has a 3-circle structure in  $\mathbb{R}^8$ , leading to the conclusion that the patches are concentrated around a Klein-bottle<sup>38</sup>. This was supported by five relatively long 1-cycles in the persistence diagram computed over the patches. To provide quantitative statistical support for this claim, we randomly selected a subset of the patches, processed them<sup>30</sup>, and computed  $p$ -value for all cycles (using the Rips complex). We repeated this experiment for varying numbers of patches, and computed the average number of detected signal cycles over 250 trials. The results are presented in Fig. 5. Firstly, we observe that there exists a single 1-cycle that is nearly always detected (the primary circle), while other cycles appear as we increase the sample size. Secondly, we observe that the fifth cycle is intermittently detected. Plotting a 2-dimensional projection of the points, we see that this cycle contains very sparse areas, increasing its birth time and consequently the  $p$ -value.

To conclude, using this approach, we are able to correctly detect the signal cycles discovered by de-Silva & Carlsson<sup>30</sup>, as well as *quantitatively declare the significance level* for each cycle.

- Infinite cycles:** To test Algorithm 1, we used a point-cloud on a 2-dimensional torus in  $\mathbb{R}^3$ . Computing the  $p$ -values of the two signal 1-cycles requires a massive complex and becomes computationally intractable for even a few thousand points. Using Algorithm 1, the filtration size is incrementally increased until the signal was detected. At 50k points, this saves approximately 80% of the edges that would otherwise be needed. This ratio would significantly increase for higher dimensional simplexes. See Sect. 7 in the Supplementary Information for the complete details.



**Figure 4.** Computing  $p$ -values for the 8-shape (a wedge of circles). (top-left) Persistence diagrams for two instances of the 8-shape with different neck gaps, with the significance lines shown for  $\alpha = 0.05$  (the dotted and dashed lines correspond to  $W = 0.1$  and  $W = 0.4$ , respectively). (top-right) The average number of signal cycles detected in 100 repetitions, as a function of the neck gap. (bottom) In green we show significant cycles. On the left ( $W = 0.1$ ) we see two significant cycles ( $p$ -values = 0.0005, 0.012), and on the right ( $W = 0.4$ ) only the outer cycle is significant ( $p$ -value=0.0015).



**Figure 5.** Testing the 3-circle model in the natural image patches. (left) Four different 2-dimensional projections of the 8-dimensional patches point-cloud. The projection on the top-right shows a 1-cycle that is quite “thin” and contains large gaps. (right) The  $p$ -value curve for the patches dataset as a function of the number of samples. We observe that we are not detecting 5-cycles in 100% of the cases. This is most likely due to the “thin” cycle.

## Discussion

In this work, we revealed a highly surprising phenomena: viewed through the lens of the  $\ell$ -values, almost all persistence diagrams follow the same universal distribution. This claim is supported by an exhaustive set of experimental results, ranging from iid samples to real data. We expect two major outcomes of this paper: (1) revealing the full extent of Conjectures 1, 2, 3 will provide fertile grounds for a whole new line of theoretical research in stochastic topology; (2) the universality properties enable the development of powerful statistical tools in TDA with hypothesis testing being only the tip of the iceberg. For example, the conjectures may lead to stronger bounds on distances between persistence diagrams coming from data, or further algorithmic advances based on the universal distribution.

## Methods

**Persistent homology.** In order to turn point-clouds into shapes that can be studied topologically, the common practice in TDA is to use their geometry to generate simplicial complexes. A simplicial complex can be thought of as a “high-dimensional graph” where in addition to vertices and edges, we include triangles, tetrahedra, and higher dimensional simplexes. Formally, we say that  $X$  is an *abstract simplicial complex* over a set  $S$ , if  $X$  consists of finite nonempty subsets of  $S$ , and is closed under inclusion (i.e.,  $A \in X$  and  $B \subset A$  implies  $B \in X$ ). We refer to elements in  $X$  of size  $k + 1$  as either  $k$ -simplexes or  $k$ -faces.

For every simplicial complex  $X$  we can compute its *homology groups*, denoted  $H_k(X)$ . Homology is a classical construction in algebraic topology<sup>39</sup>, describing the structure of simplicial complexes in terms of *chains*. Chains are formal linear combinations of simplexes of the same dimension  $-\sum_i \lambda_i \sigma_i$ , where  $\sigma_i$  are  $k$ -simplexes, and  $\lambda_i$  are some coefficients. Commonly, the coefficients are taken in a finite field (e.g.,  $\mathbb{Z}_2$ ), in which case, the space of  $k$ -chains forms a vector space, denoted  $C_k$ . The key ingredient in defining homology is the boundary operator  $\partial_k : C_k \rightarrow C_{k-1}$ . This is a linear map describing how  $(k - 1)$ -simplexes are attached to  $k$ -simplexes. Elements in the kernel of  $\partial_k$  are called  $k$ -cycles (i.e., chains with zero boundary), while elements in the image of  $\partial_{k+1}$  are called  $k$ -boundaries. The  $k$ -th homology is then defined as the space of all  $k$ -cycles which are not  $k$ -boundaries, formally given as the quotient  $H_k = \ker \partial_k / \text{im } \partial_{k+1}$ .

Loosely speaking, the homology groups  $H_k(X)$  ( $k \geq 0$ ) represent structural information about the complex  $X$ . For example, the basis of  $H_0(X)$  corresponds to the connected components, the basis of  $H_1(X)$ —to closed loops surrounding holes, and the basis of  $H_2(X)$ —to closed surfaces surrounding cavities (can be thought of as “bubbles” or “air-pockets”). For a more formal definitions and background, we refer the reader to the literature on algebraic topology<sup>39,40</sup>.

In this paper we are interested in *geometric complexes*, i.e., abstract simplicial complexes whose vertex set is a point-cloud, and simplexes are determined by the geometric configuration of the points. The two most common constructions in TDA are both parameterized by a scale parameter (radius)  $r > 0$ , and defined as follows:

- **Vietoris-Rips complex:** We include a subset of  $(k + 1)$  points as a  $k$ -simplex, if all points are within distance  $r$  from each other.
- **Čech complex:** We place a ball of radius  $r$  around each point, and then include a subset of  $(k + 1)$  points as a  $k$ -simplex, if the intersection of the corresponding balls is non-empty.

See Fig. 2 in the Supplementary Information for examples. Denoting by  $X_r$  either of the above complexes, an important observation is that this construction is “increasing”, in the sense that for  $r < s$ , we have  $X_r \subseteq X_s$ . Such sequences are also known as *filtrations*. While we can study the homology of each individual complex



$X_r$  separately, a more powerful approach is to consider the entire range of parameters (in this case  $r \in [0, \infty)$ ) and track the *evolution* of  $H_k(X_r)$  along this range. A key property of homology is that it is *functorial*, i.e., every map between simplicial complexes induces a linear map between their respective homology groups. Using the inclusion maps  $i : X_r \hookrightarrow X_s$ , we have the corresponding linear maps in homology  $i_* : H_k(X_r) \rightarrow H_k(X_s)$  for all  $r < s$ . *Persistent homology*<sup>41,42</sup> uses this collection of linear maps to describe how homology groups change over the filtration. This is done by assembling all the homology groups and the maps between them into an algebraic structure called a *persistence module*. A key result in topological data analysis<sup>43</sup> is that persistence modules can be *uniquely* decomposed into a direct-sum of “basis-like” elements. These elements correspond to  $k$ -cycles that appear (born) at some parameter value, and disappear (die) at a later value in the filtration. Given this decomposition, we can summarize the structure of the persistence module, using a collection of points  $(b, d) \in \mathbb{R}^2$ , corresponding to the birth and death times of these basis elements. This representation is known as a *persistence diagram* (see Fig. 1). Persistent homology has been extensively studied, and employed in numerous applications.

**Multiplicative persistence.** The significance of topological features is often measured by the lifetimes of persistent cycles, i.e.,  $\Delta = (\text{death} - \text{birth})$ . This method is intuitive in toy examples (see Fig. 1), as it captures the geometric “size” of topological features. However, a strong case can be made<sup>26</sup> that for geometric complexes the ratio

$$\pi = (\text{death} / \text{birth})$$

is in fact a more robust statistic to discriminate between signal and noise in persistence diagrams (for  $k > 0$ ). There are two main justifications for this statement. Firstly, the ratio  $\pi$  is scale invariant, so that cycles that have exactly the same structure but exist at different scales are weighed the same. Secondly, datasets often contain outliers that may generate cycles with a large diameter, and consequently their lifetime  $\Delta$  will also be large. However, the value of  $\pi$  for such outliers should remain low, compared to features that occur in dense regions (see Fig. 6).

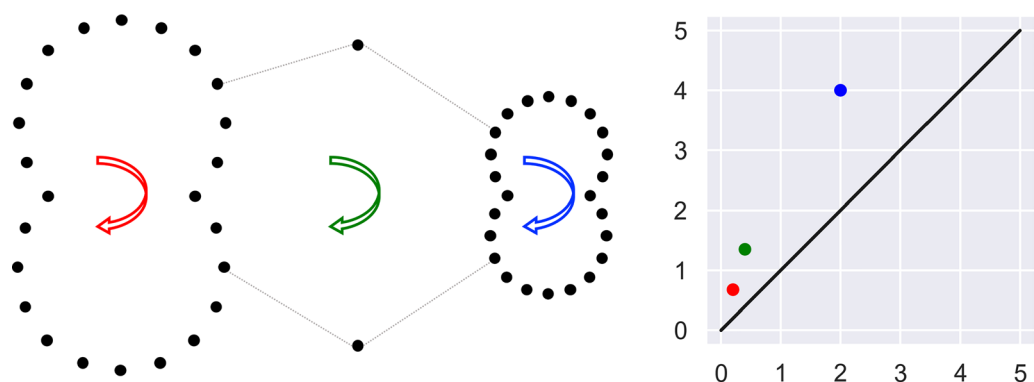
In previous work<sup>26</sup>, the asymptotic scaling for the largest  $\pi$ -value was studied. Denoting  $\pi_k^{\max} := \max_{p \in \text{dgm}_k} \pi(p)$ , the main result (Theorem 3.1) shows that with high probability

$$A \left( \frac{\log n}{\log \log n} \right)^{1/k} \leq \pi_k^{\max} \leq B \left( \frac{\log n}{\log \log n} \right)^{1/k},$$

for some constants  $A, B > 0$ . This is in contrast to *signal* cycles, who we know<sup>33</sup> are of the order of  $n^{1/d}$ . Thus, the  $\pi$ -values provide a strong separation (asymptotically) between signal and noise in persistence diagrams.

**Remark** These previous results are phrased for special cases (i.e., the uniform distribution on a  $d$ -dimensional box<sup>26</sup> and a  $d$ -dimensional flat torus<sup>33</sup>). Nevertheless, the proofs can be adapted to a wide class of compact spaces and distributions, with the same asymptotic rates.

**Experimented point-clouds.** In this section we provide the details on the models generating the various point-clouds tested. Tables summarizing all the tested settings, as well as the full results, are available in Sect. 3 of the Supplementary Information. We first discuss a few methodological points with respect to the experiments.



**Figure 6.** Motivation for multiplicative persistence. The point-cloud on the left generates three 1-dimensional cycles (marked by the arrows). On the right we see the corresponding persistence diagram for the Čech filtration with matching cycle colors. The coordinates of the points are  $p_1 = (0.2, 0.675)$  (red),  $p_2 = (0.4, 1.35)$  (green), and  $p_3 = (2, 4)$  (blue). Their corresponding lifetimes are  $\Delta_1 = 0.475$ ,  $\Delta_2 = 0.95$ , and  $\Delta_3 = 2$ . This might lead us to declare the blue cycle as the most significant, while it seems to be generated by outliers. In addition, while the structure of the other two cycles (red and green) is identical, their lifetime is different. Both issues are resolved if we take the death/birth ratio instead. In this case, we have  $\pi_1 = \pi_2 = 3.375$ , while  $\pi_3 = 2$ .

- (1) Most experiments on the Čech complex (particularly in 2 or 3-dimensions), were done in practice using the much more computationally efficient Delaunay-Alpha complex (especially for lower dimensions), which has an identical persistence diagram.
- (2) The experiments were run primarily on the Apocrita HPC cluster, using several different software packages including Gudhi<sup>44</sup>, Ripser<sup>45</sup>, Eirene<sup>46</sup>, Dionysus<sup>47</sup>, and Diode<sup>48</sup>.
- (3) The persistence diagrams for these experiments and the associated code can be found at <https://doi.org/10.7910/DVN/CNIHYV> and <https://github.com/primozskraba/PersistenceUniversality> respectively.

**Remark** In all the considered settings, it is necessary to assume that the diagrams we analyze consist only of noisy cycles, without any signal. This assumption is essential, particularly when estimating the CDF or the value of the normalization parameter  $B$ . In cases where the signal is known (e.g., an annulus), we can manually exclude the signal features. However, in cases where the signal is unknown (e.g., audio sampling, sentence embeddings), we cannot do so. Nevertheless this issue is not significant in practice. The diagrams we generate typically contain thousands to millions of points. Thus, unless we anticipate a massive amount of signal features, the inclusion of signal cycles in our estimates will have a negligible effect, especially in the log log scale. This observation is particularly important for the applicability of our hypothesis testing framework.

**Sampling from iid distributions.** The first setting tested is where points are sampled from a random model in an iid fashion. Let  $X = (X^{(1)}, \dots, X^{(d)})$  denote a single point in one of the iid samples. We describe briefly how  $X$  is generated in each of the models.

- **Box:**  $X$  is uniformly distributed in  $[0, 1]^d$ .
- **Ball:**  $X$  is uniformly distributed in a unit  $d$ -dimensional ball. Sampling was done using the rejection-sampling method.
- **Annulus:**  $X$  is uniformly distributed in a  $d$ -dimensional annulus with radii in the range  $[1/2, 1]$ . Sampling was done using the rejection-sampling method.
- **Sphere:**  $X$  is uniformly distributed on a  $d$ -dimensional unit sphere, embedded in  $\mathbb{R}^{d+1}$ . Sampling was done by generating a standard  $(d + 1)$ -dimensional normal variable, and projecting it on the unit sphere.
- **Beta(a, b):** The coordinates  $X^{(1)}, \dots, X^{(d)}$  are sampled independently from the Beta(a,b) distribution.
- **Cauchy:** The coordinates  $X^{(1)}, \dots, X^{(d)}$  are sampled independently from the Cauchy distribution.
- **Normal:** The coordinates  $X^{(1)}, \dots, X^{(d)}$  are sampled independently from the standard normal distribution.
- **Torus:** We generate points on the 2-dimensional torus embedded in  $\mathbb{R}^3$  as follows. We generate two independent variables  $\phi$  and  $\theta$  uniformly in  $[0, 2\pi]$ . Then we take the coordinates to be:

$$X^{(1)} = (R_1 + R_2 \cos(\phi)) \cos(\theta), \quad X^{(2)} = (R_1 + R_2 \cos(\phi)) \sin(\theta), \quad X^{(3)} = R_2 \sin(\phi).$$

We used  $R_1 = 2$  and  $R_2 = 1$ .

- **Klein:** We sample an embedding of the Klein bottle into  $\mathbb{R}^4$  as follows. We generate two independent variables  $\phi$  and  $\theta$  uniformly in  $[0, 2\pi]$ . The value of  $X$  is then computed as

$$\begin{aligned} X^{(1)} &= (1 + \cos(\theta)) \cos(\phi), & X^{(2)} &= (1 + \cos(\theta)) \sin(\phi), \\ X^{(3)} &= \sin(\theta) \cos\left(\frac{\phi}{2}\right), & X^{(4)} &= \sin(\theta) \sin\left(\frac{\phi}{2}\right). \end{aligned}$$

- **Projective:** We sample an embedding of the real projective plane into  $\mathbb{R}^4$ . We generate independent variables  $U_0, V_0$  and  $W_0$  from the standard normal distribution. Next we take  $(U, V, W) = \frac{(U_0, V_0, W_0)}{\sqrt{U_0^2 + V_0^2 + W_0^2}} \in \mathbb{S}^2$ , and define

$$X^{(1)} = UV, \quad X^{(2)} = UW, \quad X^{(3)} = V^2 - W^2, \quad X^{(4)} = 2VW.$$

- **Linkage:** This model samples the configuration space of unit pentagonal linkages, i.e. a pentagon where adjacent edges are of unit length. To sample this space, we first fix two vertices at  $p_1 = (0, 0)$  and  $p_2 = (1, 0)$  respectively. Next, we generate two independent variables  $\phi$  and  $\theta$  uniformly in  $[0, 2\pi]$ , and set

$$p_5 = (\cos(\phi), \sin(\phi)), \quad p_3 = (1 + \cos(\theta), \sin(\theta)).$$

If  $\|p_3 - p_5\| > 2$ , the sample is rejected as there is no linkage with the chosen angles. Otherwise, there are two possible choices of the last point  $p_4$ . Let  $(q^{(1)}, q^{(2)})$  denote the midpoint of  $\overline{p_3 p_5}$ . Then

$$p_4 = q + S \frac{\sqrt{1 - \|q - p_5\|^2}}{\|q - p_5\|} (p_5^{(2)} - q^{(2)}, q^{(1)} - p_5^{(2)})$$

where  $S$  is independent of  $\phi, \theta$ , and  $\mathbb{P}(S = 1) = \frac{1}{2}$  and  $\mathbb{P}(S = -1) = \frac{1}{2}$ . See Fig. 1 in the Supplementary Information for an example.

- **Neptune:** We construct a sample from the surface of the statue of Neptune. We used a triangulation of the surface<sup>49</sup>, consisting of 4,007,872 triangles. To generate a sample, we first compute the area of each triangle

and then choose a triangle at random with probability inversely proportional to the area of the triangle. We then pick a point uniformly in the chosen triangle. See Fig. 1 in the Supplementary Information.

- **Henneberg:** We construct a sample of the Henneberg surface in  $\mathbb{R}^3$  (see Fig. 1 in the Supplementary Information). We start by generating two independent variables  $\phi$  and  $\theta$  uniformly in  $[0, 2\pi]$ . We then construct the sample by

$$\begin{aligned} X^{(1)} &= 2 \cos(\theta) \sinh(\phi) - \frac{2}{3} \cos(3\theta) \sinh(3\phi), \\ X^{(2)} &= 2 \sin(\theta) \sinh(\phi) + \frac{2}{3} \sin(3\theta) \sinh(3\phi), \\ X^{(3)} &= 2 \cos(2\theta) \cosh(2\phi). \end{aligned}$$

- **Stratified spaces:** To construct  $X$ , we consider two spaces  $M_1 \subset M_2$  such that the dimension of  $M_1$  is less than  $M_2$ . Then one of the two spaces is chosen with some probability  $p$  (which is a parameter of the model), and the chosen space is sampled uniformly. Although several models were tried, in the examples we show a plane  $[-1, 1]^2$  embedded in the middle of a cube  $[-1, 1]^3$ . In other words, if the point  $(x, y)$  is chosen from the plane, the coordinates would be  $(x, y, 0)$ .

**Sampling from non-iid distributions.** In addition to the iid setting, we also tested two examples of non-iid point-clouds. The first example is sampling the path of a  $d$ -dimensional Brownian motion  $W_t$ . To sample  $W_t$  at times  $t = 1, \dots, n$  we use the fact that  $W_t$  has stationary independent increments. We start by taking  $Z_1, \dots, Z_n$  to be iid  $d$ -dimensional standard normal variables, and then we define  $X_i = W_{t=i} = \sum_{i=1}^n Z_i$ .

The second example we examined is a discrete-time sample of the Lorenz dynamical system, which is generated as follows. We start by picking a random initial point, uniformly in  $[0, 1]^3$ . We then generate the sample using the differential equations,

$$\begin{aligned} \frac{dX^{(1)}}{dt} &= \sigma (X^{(2)} - X^{(1)}), \\ \frac{dX^{(2)}}{dt} &= X^{(1)} (\rho - X^{(3)}) - X^{(2)}, \\ \frac{dX^{(3)}}{dt} &= X^{(1)} X^{(2)} - \beta X^{(3)}. \end{aligned}$$

We use  $\sigma = 45$ ,  $\rho = 54$ , and  $\beta = 10$ , and a numerical approximation with  $dt = 0.1$  to generate a trajectory for the number of samples required. Note that each instance is a single trajectory. See Fig. 9 in the Supplementary Information for examples.

**Testing on real data.** Finally, we tested the strong-universality conjectures against two examples of real data—patches of natural images, and sliding windows of voice recordings. In this section we provide more details about these examples.

- **Image patches:** The images were taken from van Hateren and van der Schaaf<sup>34</sup> database. This database contains a collection of about 4,000 gray-scale images. We follow the procedure described in Lee et al.<sup>35</sup>. We randomly select patches of size  $3 \times 3$  from the entire dataset. This gives us a point-cloud in  $\mathbb{R}^9$ . Let  $\mathbf{x} = (x_1, \dots, x_9) \in \mathbb{R}^9$  represent the log-values of a single patch, and follow the following steps:

- (1) Compute the average pixel value  $\bar{\mathbf{x}}$ , and subtract it from all 9 pixels in the patch, i.e.  $\mathbf{y} = \mathbf{x} - \bar{\mathbf{x}}$
- (2) Compute the “D-norm” (a measure for contrast),  $\|\mathbf{y}\|_D = \sqrt{\mathbf{y}^T D \mathbf{y}}$ .
- (3) Use the D-norm to normalize the pixel values,  $\mathbf{z} = \mathbf{y} / \|\mathbf{y}\|_D$ .
- (4) Use the Discrete Cosine Transform (DCT) basis to change the coordinate system,  $\mathbf{v} = \Lambda A^T \mathbf{z}$ .

The values of  $D, A, \Lambda$ , as well as more details about this procedure can be found in the literature<sup>35,38</sup>. The process above results in a point-cloud lying on the unit 7-dimensional sphere in  $\mathbb{R}^8$ .

For the topological analysis<sup>38</sup>, the point-cloud is further filtered, in order to focus on the “essential” information captured by the patches. This is done in two steps:

- (1) Keep only “high-contrast” patches—whose D-norm is in the top 20%.
- (2) Compute the distance of each of the remaining patches to their  $k$ -nearest neighbor (with  $k = 15$ ), and keep only the patches in the bottom 15%.

We repeat the exact procedure performed Carlsson et al.<sup>38</sup> for two main reasons. Firstly, we wanted to test our hypothesis testing framework to assign  $p$ -values to the cycles found using this procedure. Secondly, this procedure makes the sample distribution more intricate, and adds dependency between the sample points. We wanted to challenge our conjectures with data as complex as possible. In addition to taking the original 8-dimensional point-cloud, we also examined its lower dimensional projections for dimensions  $d = 3, \dots, 7$ . The sample size used was  $n = 50,000$  for all dimensions.

- **Sound recordings:** We took an arbitrary audio recording (the voice of one of the authors), and applied the time-delay embedding transformation<sup>36</sup> to convert the temporal signal into a  $d$ -dimensional point-cloud. The voice recording is a 20 seconds excerpt, sampled at 16KHz, 48Kbps. We denote the corresponding discrete time signal as  $V_t$ . To convert the signal into a point-cloud we used the following:

$$X_i = (V_{i\Delta}, V_{i\Delta+\tau}, \dots, V_{i\Delta+(d-1)\tau}) \in \mathbb{R}^d, \quad i = 1, \dots, n.$$

The values we took for the example here were  $\Delta = 3$ , and  $\tau = 7$ . This, in particular, generates overlap between the windows, which guarantees strong dependency both between the points, and between the coordinates of each point. The sample size taken was  $n = 50,000$ .

- **Sentence embeddings:** Choosing several freely available texts of sufficient length—we chose the Bible, Ulysses by James Joyce, and Moby Dick by Herman Melville, obtained from Project Gutenberg. Each book was first tokenized into sentences using the Natural Language Toolkit (NLTK)<sup>50</sup>. Each sentence was then embedded into a 384 dimensional space using a sentence transformer. These are pretrained models which can be used to compute the embeddings. We used the **all-MiniLM-L6-v2** model<sup>37</sup>. This model has been trained on over a billion sentences. The persistence diagram was then computed for resulting point cloud using Euclidean distance. The texts were chosen to so that they generated between 9000-30,000 sentences.

## Data availability

The experimental data and code for all the results shown can be found at <https://github.com/primozskraba/PersistenceUniversality> and <https://doi.org/10.7910/DVN/CNIHYV>, respectively.

Received: 27 October 2022; Accepted: 28 June 2023

Published online: 28 July 2023

## References

1. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **46**(2), 255–308 (2009).
2. Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**(7895), 123–128 (2022).
3. Giusti, C., Pastalkova, E., Curto, C. & Itskov, V. Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Natl. Acad. Sci.* **112**(44), 13455–13460 (2015).
4. Curto, C. What can topology tell us about the neural code?. *Bull. Am. Math. Soc.* **54**(1), 63–78 (2017).
5. Chan, J. M., Carlsson, G. & Rabadan, R. Topology of viral evolution. *Proc. Natl. Acad. Sci.* **110**(46), 18566–18571 (2013).
6. Ichinomiya, T., Obayashi, I. & Hiraoka, Y. Protein-folding analysis using features obtained by persistent homology. *Biophys. J.* **118**(12), 2926–2937 (2020).
7. Lawson, P., Sholl, A. B., Quincy, B. J., Fasy, B. T. & Wenk, C. Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Sci. Rep.* **9**(1), 1–15 (2019).
8. Masoomy, H., Askari, B., Tajik, S., Rizi, A. K. & Jafari, G. R. Topological analysis of interaction patterns in cancer-specific gene regulatory network: Persistent homology approach. *Sci. Rep.* **11**(1), 1–11 (2021).
9. Hirata, A., Wada, T., Obayashi, I. & Hiraoka, Y. Structural changes during glass formation extracted by computational homology with machine learning. *Commun. Mater.* **1**(1), 1–8 (2020).
10. Lee, Y. *et al.* Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Commun.* **8**(1), 1–8 (2017).
11. Kramár, M. *et al.* Analysis of kolmogorov flow and rayleigh-bénard convection using persistent homology. *Phys. D Nonlinear Phenom.* **334**, 82–98 (2016).
12. Xian, L., Adams, H., Topaz, C. M. & Ziegelmeier, L. Capturing dynamics of time-varying data via topology. *Found. Data Sci.* **4**(1), 1–36 (2022).
13. Adler, R. J., Agami, S. & Pranav, P. Modeling and replicating statistical topology and evidence for CMB nonhomogeneity. *Proc. Natl. Acad. Sci.* **114**(45), 11878–11883 (2017).
14. Sousbie, T. The persistent cosmic web and its filamentary structure-I. Theory and implementation. *Mon. Not. R. Astron. Soc.* **414**(1), 350–383 (2011).
15. Chazal, F., Cohen-Steiner, D. & Lieutier, A. A sampling theory for compact sets in Euclidean space. *Discrete Comput. Geom.* **41**(3), 461–479 (2009).
16. Chazal, F., Guibas, L. J., Oudot, S. Y. & Skraba, P. Persistence-based clustering in Riemannian manifolds. *J. ACM (JACM)* **60**(6), 41 (2013).
17. Niyogi, P., Smale, S. & Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39**(1–3), 419–441 (2008).
18. Blumberg, A. J., Gal, I., Mandell, M. A., Pancia, M. Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Found. Comput. Math.* **14**, 745–789 (2013).
19. Fasy, B. T. *et al.* Confidence sets for persistence diagrams. *Ann. Stat.* **42**(6), 2301–2339 (2014).
20. Reani, Y., & Bobrowski, O. Cycle registration in persistent homology with applications in topological bootstrap. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(5), 5579–5593 (2022).
21. Vejdemo-Johansson, M. & Mukherjee, S. Multiple hypothesis testing with persistent homology. *Found. Data Sci.* **4**(4), 667–705 (2022).
22. Hiraoka, Y., Shirai, T. & Trinh, K. D. Limit theorems for persistence diagrams. *Ann. Appl. Probab.* **28**(5), 2740–2780 (2018).
23. Owada, T. & Adler, R. J. Limit theorems for point processes under geometric constraints (and topological crackle). *Ann. Probab.* **45**(3), 2004–2055 (2017).
24. Yogeshwaran, D. & Adler, R. J. On the topology of random complexes built over stationary point processes. *Ann. Appl. Probab.* **25**(6), 3338–3380 (2015).
25. Yogeshwaran, D., Subag, E., Adler, R. J. Random geometric complexes in the thermodynamic regime. *Probab. Theory Relat. Fields* **167**(1), 107–142 (2016).
26. Bobrowski, O., Kahle, M. & Skraba, P. Maximally persistent cycles in random geometric complexes. *Ann. Appl. Probab.* **27**(4), 2032–2060 (2017).
27. Kahle, M. & Meckes, E. Limit theorems for Betti numbers of random simplicial complexes. *Homol. Homotopy Appl.* **15**(1), 343–374 (2013).
28. Chazal, F. *et al.* Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18**(1), 5845–5884 (2017).

29. Chazal, F., Cohen-Steiner, D. & Mérigot, Q. Geometric inference for probability measures. *Found. Comput. Math.* **11**(6), 733–751 (2011).
30. de Silva, V. & Carlsson, G. Topological estimation using witness complexes. *SPBG* **4**, 157–166 (2004).
31. Corbet, René, Kerber, M., Lesnick, M., Osang, G. Computing the multicover bifiltration. *Discrete Comput. Geom.* (2023).
32. Sheehy, D. R. A multicover nerve for geometric inference, in *CCCG*, 309–314 (2012).
33. Bobrowski, O. & Skraba, P. Homological percolation: The formation of giant  $k$ -cycles. *Int. Math. Res. Not.* **2022**(8), 6186–6213 (2022).
34. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **265**(1394), 359–366 (1998).
35. Lee, A. B., Pedersen, K. S. & Mumford, D. The nonlinear statistics of high-contrast patches in natural images. *Int. J. Comput. Vis.* **54**(1–3), 83–103 (2003).
36. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence* Springer lecture notes in Mathematics. **898**, 366–381 (1981).
37. Reimers, N., Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11 (Association for Computational Linguistics, 2019).
38. Carlsson, G., Ishkhanov, T., de Silva, V. & Zomorodian, A. On the local behavior of spaces of natural images. *Int. J. Comput. Vis.* **76**(1), 1–12 (2008).
39. Hatcher, A. *Algebraic Topology* (Cambridge University Press, 2002).
40. Munkres, J. R. *Elements of Algebraic Topology* Vol. 2 (Addison-Wesley Reading, 1984).
41. Edelsbrunner, H. & Harer, J. L. *Computational Topology: An Introduction* (AMS Bookstore, 2010).
42. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**(2), 249–274 (2005).
43. Crawley-Boevey, William. Decomposition of pointwise finite-dimensional persistence modules. *J. Algebra Appl.* **14**(05), 1550066 (2015).
44. Maria, C., Boissonnat, J.-D., Glisse, M., Yvinec, M. The gudhi library: Simplicial complexes and persistent homology, in *International Congress on Mathematical Software*, 167–174 (Springer, 2014).
45. Bauer, U. Ripser: Efficient computation of vietoris-rips persistence barcodes. *J. Appl. Comput. Topol.* **5**(3), 391–423 (2021).
46. Henselman, G., Ghrist, R. Matroid filtrations and computational persistent homology. *ArXiv e-prints* (2016).
47. Morozov, D. Dionysus 2. <http://mrzv.org/software/dionysus2>.
48. Morozov, D. Diode. <https://github.com/mrzv/diode>.
49. Falcidieno, B. Aim@ shape project presentation, in *Proceedings. Shape Modeling International 2004*, 329–329 (IEEE Computer Society, 2004).
50. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media Inc, 2009).

## Acknowledgements

The authors are grateful to the following people for their helpful comments and suggestions: Henry Adams, Robert Adler, Matthew Kahle, Anthea Monod, Vidit Nanda, Erik Novak, and Jasna Urbancic. We would also like to thank the anonymous referees for their useful comments. O.B. was partially supported by the Israel Science Foundation grant #1965/19.

## Author contributions

O.B. and P.S. contributed equally to the planning of the experiments, implementation of the experiments, and writing up the results.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-37842-2>.

**Correspondence** and requests for materials should be addressed to O.B. or P.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023