EBGLIDE: Empirical Bayesian Gumbel-based Loop Inference Detection

Gus Gerlach[1,*,+], **Jake Reed, PhD**[3,+], **Kevin Coombes, PhD**[2,], **and Zachary Abrams, PhD**[1,]

[1]Institute for Informatics, Division of Data Science and Biostatistics, Washington University School of Medicine, Saint Louis, MO, USA
[3]Huntsman Cancer Institute, University of Utah Health Academic Medical Center, Salt Lake City, UT, USA
[2]Department of Biostatistics, Data Science, and Epidemiology, Georgia Cancer Center at Augusta University, Augusta, GA, USA
[*]g.gerlach@wustl.edu
[+]these authors contributed equally to this work

ABSTRACT

Topological data analysis (TDA) has emerged as a powerful framework for extracting meaningful patterns from complex datasets across computational disciplines. However, distinguishing genuine topological features from noise remains a fundamental challenge in persistent homology. Here we present EBGLIDE (Empirical Bayesian Gumbel-based Loop Inference Detection), a novel statistical method for identifying significant topological loops in point cloud data. Our approach leverages the surprising universal property that transformed loop persistence values follow a left-skewed Gumbel distribution. By applying an empirical Bayesian framework to this null distribution, EBGLIDE computes posterior probabilities for loop significance with a modified credible interval approach. We validated our method on over 5,000 simulated datasets designed to mimic single-cell RNA-seq data with known ground truth topological features. EBGLIDE consistently outperformed existing hypothesis tests, achieving F1 scores of 0.815 (compared to 0.556 for the previous best method at $\alpha = 0.05$) and correctly identifying the exact number of significant loops 75% of the time. The method demonstrates robust performance across varying numbers of topological features and provides a principled statistical framework for topological inference. EBGLIDE addresses a critical gap in TDA workflows by offering researchers a reliable, accurate method for distinguishing signal from noise in persistence diagrams, with potential applications in biological regulation analysis, cancer research, and other domains requiring topological insights.

Introduction

All biological systems require some form of regulation to maintain the life of the organism. Consequently, negative feedback loops are commonly used to maintain homeostasis and prevent biological processes from running out of control. When attempting to identify regulatory connections within data, it can be beneficial to adopt a topological perspective. Looking at biological systems from a topological standpoint allows us to view regulatory cycles as two-dimensional loops. Previous research has successfully linked topological objects to relevant biological[1,2] and medical[3,4] insights, but rigorous tests for identifying significant loops have been lacking.

Viewing data topologically requires point-clouds to be viewed as geometric objects. Specifically, point-clouds are viewed as simplicial complexes, a set of points, lines, triangles, and higher dimensional simplexes. A common method for investigating these simplicial complexes is known as ‘Vietoris-Rips’ filtration (also Čech Filtration, which this paper does not deal with), a process by which simplexes are created and destroyed using an increasing scale factor[5]. The case in two dimensions can be thought of as a set of circles, centered at each point in the dataset, with an expanding radius r. As r expands, new intersections between circles are created, forming topological items which can be studied. A highly relevant result of this process is topological persistence - the longer a topological feature exists under the expanding scale factor, the more likely it is to represent the true nature of the underlying dataset. This provides the basis for the field of *Persistent Homology*, which is concerned with the significance of topological features over multiple scales.

Information extracted from this process, such as the birth and death radius of topological features is highly relevant. The topological feature of our investigation are ‘cycles’, also called ‘loops’, which are equally characterized by the voids they create. Filtration tracks the birth and death scales for these cycles, corresponding to the point at which cycles are created, and then later filled in (death). This information is graphically represented by a persistence diagram. In a persistence diagram, each point represents a topological feature - the x coordinate is birth scale, the y coordinate is death scale. The further a point is from the diagonal $y = x$, the longer the loop persists, and the more significant it can be considered. Our novel method EBGLIDE deals with identifying when a loop can be considered statistically significant. While the computation of Rips persistence diagrams

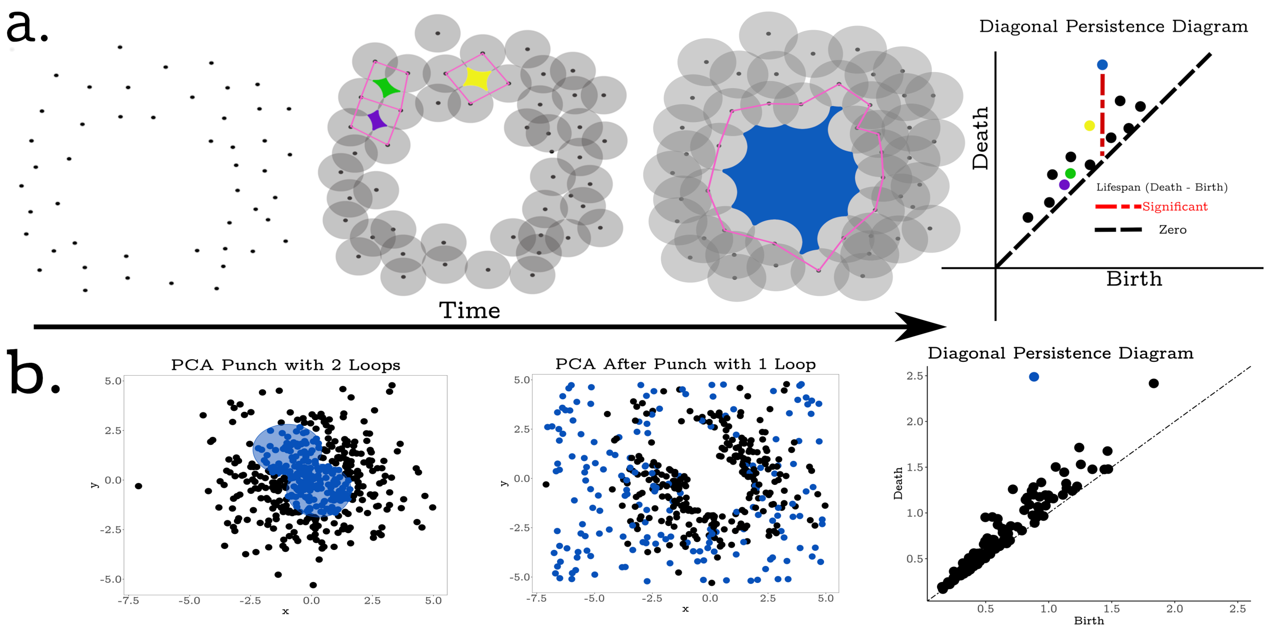

Figure 1. Vietoris-Rips filtration on a set of points in two dimensions. As time progresses the points grow in size until boundaries are formed which then produces a loop. A diagonal persistence diagram shows the "birth" and "death" of these loops. The significant loop is the most persistent loop (longest lifespan) which is highlighted by the two dashed line and blue point (a). In b, loops are simulated by "punching" holes in PCA plots. The blue points were removed from the ellipses segments and scattered around the PCA plot. Two "holes" were assigned but only one remained after considering overlap. The persistence diagram shows the significant loop in blue.

is relatively simple, determining at what rigorous threshold a topological feature may be considered ‘significant’ is an open question, with no immediately obvious answers. Various R packages provide functionality for viewing persistence diagrams graphically, but many point-clouds do not contain immediately obvious loops, or there may be multiple loops which persist for longer than the vast majority do. EBGLIDE provides a highly accurate way to distinguish between noise and signal.

Methods

The R package ‘TDA’ was used for the vast majority of our filtration and persistent homology investigations[6]. The most relevant function is ‘ripsDiag’, which computes the persistence diagram of the Rips filtration applied to the relevant dataset 1. The dimension parameter is set to 1 to indicate we are searching for loops. The ‘ripsDiag’ function produces a persistence diagram such as the one in Figure 1, and a Rips object which contains the birth and death date for each loop discovered through filtration.

Our method for determining significance begins with this Rips object: a set of loops $\{c\}_{i=1}^{n}$ where each $c_i = (b_i, d_i)$ where b is birth radius and d is death radius. $d_i - b_i$ then gives the duration for a particular topological feature. Following Bobrwski and Skraba[7], we apply a surprising transformation: define $\pi(c_i) = \frac{d_i}{b_i}$, the multiplicative duration. Bobrowski and Skraba[7] argue that the ratio is more robust than $d - b$ for distinguishing between signal and noise, for two reasons. First, the ratio is scale invariant, so cycles which are structurally equivalent but exist in different scales are scored the same. Second, the process of filtration may reveal outliers which have large $d - b$ values. The ratio should keep these outliers in check by weighting to how dense the point cloud surrounding the loop is. So, we proceed with our transformation, using $\pi(c) = \frac{d}{b}$ as our metric for persistence.

Then, for every $c_i \in c$,

$$\ell(c) := \log\log(\pi(c)) + B \tag{1}$$

where $B = -\lambda - \bar{L}$, $\bar{L} = \frac{1}{|c|}\sum_{c_i \in c} \log\log(\pi(c_i))$, and λ is the Euler-Mascheroni constant (≈ 0.5772). The resulting distribution of all $\ell-$values universally follows the left-skewed Gumbel distribution[7], henceforth referred to as "LGumbel." Bobrowski and Skraba test a variety of datasets, including real world data, and the striking fit to the LGumbel distribution holds *universally.*[7] This null distribution provides a basis for our significance test. LGumbel has PDF and CDF defined below:

$$f(x) = e^{x-e^x}, F(x) = 1 - e^{-e^x} \tag{2}$$

The expected value of this function is λ, the Euler-Mascheroni constant, which we use in 1.

Using the resulting distribution of ℓ values, we employ an Empirical bayes method[8] to calculate posterior probabilities. We assume that there are only two types of loops: "significant" and "not significant". Define the prior probabilities as $p_1 = \text{Prob}\{\text{Significant}\}$ and $p_0 = 1 - p_1$. The prior probabilities have corresponding density functions, in this case defined as: $f_1(\ell)$ if loop$_i$ is significant and $f_0(\ell)$ otherwise. The combined distribution is then viewed as a combination of the known distribution, LGumbel, and an unknown "interesting" distribution. The mixture density function is defined as

$$f(\ell) = p_0 f_0(\ell) + p_1 f_1(\ell) \tag{3}$$

Applying Bayes' theorem then gives our posterior probability of being significant:

$$p_1(\ell) := \text{Prob}\{\text{Significant} \mid \ell_i = \ell\} = 1 - p_0 \cdot \frac{f_0(\ell)}{f(\ell)}$$

f_0 is the theoretical distribution, namely the probability density function of the LGumbel distribution equation 2. f_0 is estimated on a large set of midway points between the minimum ℓ and the maximum ℓ. We estimate the mixture density function $f(\ell)$ by using a density interpolation approach provided by the 'approx()' function in R. Other methods for estimating the combined density function were tested, including the use of splines and local polynomial regression fitting. The density interpolation approach consistently provided by the best results. The prior value p_0 is impossible to estimate without a strong understanding of the underlying point cloud. Following Efron and Tibshirani[8], we estimate this prior probability by minimizing p_0 to be the smallest positive value which keeps all posterior probabilities non-negative. Together, these give us an estimation of $p_1(\ell)$:

$$\hat{p}_1(\ell) = 1 - p_0 \cdot \frac{f_0(\ell)}{\hat{f}(\ell)} \tag{4}$$

Posterior probabilities are calculated for every loop discovered in the process of filtration, resulting in a new distribution of posterior probabilities, with a set of p-values $\{p\}_{i=1}^n$ corresponding to an estimate of the probability of each individual loop being significant. We could use these probabilities and a chosen α-level to call loops significant, however, as we will see later, these estimates alone are generally not sufficient to determine the significance of loops. To proceed to calling loops significant or not, we implement a modified version of a credible interval. Find the top α percentile $p-$value in the set $\{p\}_{i=1}^n$, called p_α. Here we implement a sanity check: only loops with a value > 1 may be considered significant. The necessity for this check results from the method not being unidirectional: loops with extremely small multiplicative ratios $(\pi(c) \approx 1)$ which produce very negative ℓ values are likely to be a part of the "interesting" distribution and not a part of the known LGumbel distribution. The check for which loops are significant is then twofold: loops must have a $\ell-$value ≥ 1 and a posterior probability $p \geq p_\alpha$. Figure 2 summarizes the entire process, from initial point cloud to final determination of a loop's significance.

To test our method against other significance tests, we devised a ground truth and then repeatedly tested each method on simulated data. For the simulation of data, we started with data randomly generated by the 'zinbwave' package in R[9]. This package generates data with a zero-inflated negative binomial distribution, which is a common distribution for single-cell RNA-seq data. We then applied principal component analysis (PCA) to the data to reduce the dimensionality. The first two principal components were selected for further analysis, thereby reducing the dimensionality of the data to two. This was done to simulate a point cloud in two dimensions, which is a common input for topological data analysis. We generated 5000 simulations with this method composed of 250 points which is a common application of topological data analysis. The purpose of this was to approximate single-cell RNA-seq data. Each of these simulations was composed of 250 'cells' each with 450 'genes'. The topological structures were then procedurally generated using custom R scripts.

Briefly, a random number of 'holes', between 0 and 3, were 'punched' in the data which varied in size and location. All points within these ellipses were moved outside the respective ellipse. This created a dataset with topological features, a loop or loops, which was then analyzed using multiple methods. Each method was tested by performing TDA analysis followed by comparing the number of identified significant loops to the known number of 'holes'. The purpose of this was to evaluate the robustness of each method in detecting multiple 1-dimensional topological features (loops) from noisy data.

To further investigate the different methods' veracity at identifying known topological features, another 500 simulations were produced using the previous method (zinbwave and PCA). However, instead of 'punching' one to three 'holes', for 125 of the simulations, we punched 5 equivalent sized 'holes' in the data, and for another 125 simulations, we punched 10 equivalent sized 'holes' in the data. For the remaining 250 simulations, we 'punched' 1-2 'holes' that increased in size incrementally, generating a series of 5 simulations with 1 or 2 holes of increasing size. This was done to test the methods' ability to detect multiple loops in the presence of noise.

We tested EBGLIDE against existing methods for determining loop significance. We devised a performance score reminiscent of an F1 score, highlighting both the precision and recall of each significance test (henceforth referred to as 'F1'). We calculate the number of true positives, false positives, and false negatives based on difference between guess and truth. For instance, if there are two actual loops and the method calls three significant, then the method gets two true positives and one false positive. We then calculate precision, recall, and F1 in the standard way.

We also devised a performance metric entitled 'Binary F1'. In this case, we award only one 'point' for each simulated dataset. If the true number of loops equals the number called significant, one is added for true positives. If the true number of loops is less than the guess, we add one in false positives and, if the true number of loops is more than the guess, one is added in false negatives (and no other additions are made). Precision, Recall, and F1 are then calculated as before.

The process of filtration applied to each dataset produces a list of hundreds of loops, the vast majority of which are not significant. We summed, on an individual loop level, the number of loops which are correctly or incorrectly called as significant or not significant. These are reported in the True Positives, False Positives, True Negatives, and False Negatives rows. Finally, we also calculate the number of times each method got it exactly right - when the number of loops called significant exactly aligned with the true number of loops (Percent Correct).

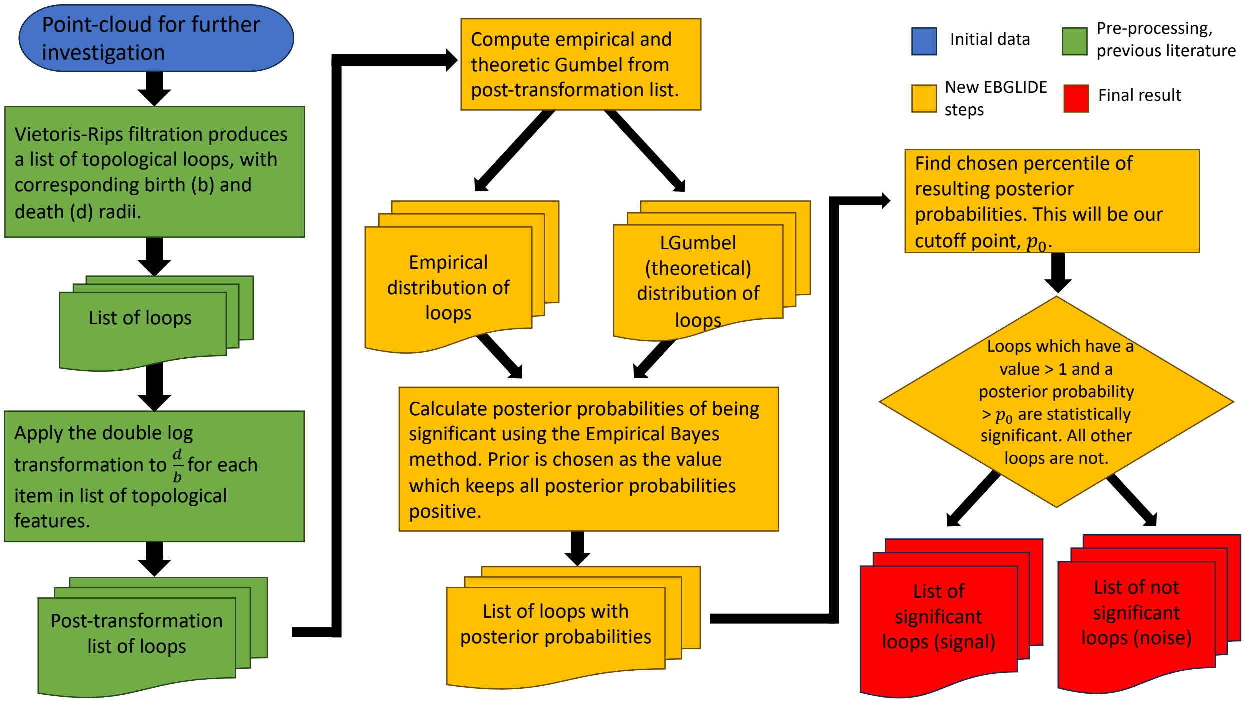
Point-cloud for further investigation
Vietoris-Rips filtration produces a list of topological loops, with corresponding birth (b) and death (d) radii.
List of loops
Apply the double log transformation to $\frac{d}{b}$ for each item in list of topological features.
Post-transformation list of loops
Compute empirical and theoretic Gumbel from post-transformation list.
Empirical distribution of loops
LGumbel (theoretical) distribution of loops
Calculate posterior probabilities of being significant using the Empirical Bayes method. Prior is chosen as the value which keeps all posterior probabilities positive.
List of loops with posterior probabilities
Initial data
Pre-processing, previous literature
New EBGLIDE steps
Final result
Find chosen percentile of resulting posterior probabilities. This will be our cutoff point, p_0.
Loops which have a value > 1 and a posterior probability > p_0 are statistically significant. All other loops are not.
List of significant loops (signal)
List of not significant loops (noise)

Figure 2. Workflow diagram from initial data to final result, a determination of which loops are significant and which are not. Yellow represents the unique steps of the EBGLIDE method

Results

Table 1 depicts the final scores after applying each method to thousands of datasets, all with exact ground truth. The methods entitled 'alpha' refers to a method which simply calls loops significant or not based on their posterior probabilities, which are calculated during the EBGLIDE method. Preliminary testing revealed $\alpha = 0.3$ to be the best option to maximize both F1 and binary F1 performance scores. We also tested $\alpha = 0.5$ as in preliminary testing it was found to closely mirror the EBGLIDE Recall scores. The 'null' method refers to the method[7] which uses LGumbel as a null distribution, and calculates the probability of an individual loop's post transformation persistence value differing from that distribution. We tested two statistical alpha levels: 0.05 and 0.5. 0.05 was chosen to compare directly with EBGLIDE's effectiveness at that alpha level. 0.5 was chosen after preliminary testing revealed it to maximize the F1 score. Although this is an unrealistic alpha level for any dataset other than strictly simulated data, we include it here. 'EBGLIDE' is the aforementioned method, with the alpha level corresponding to the percentile of the posterior probability that becomes the cutoff point for significant loops.

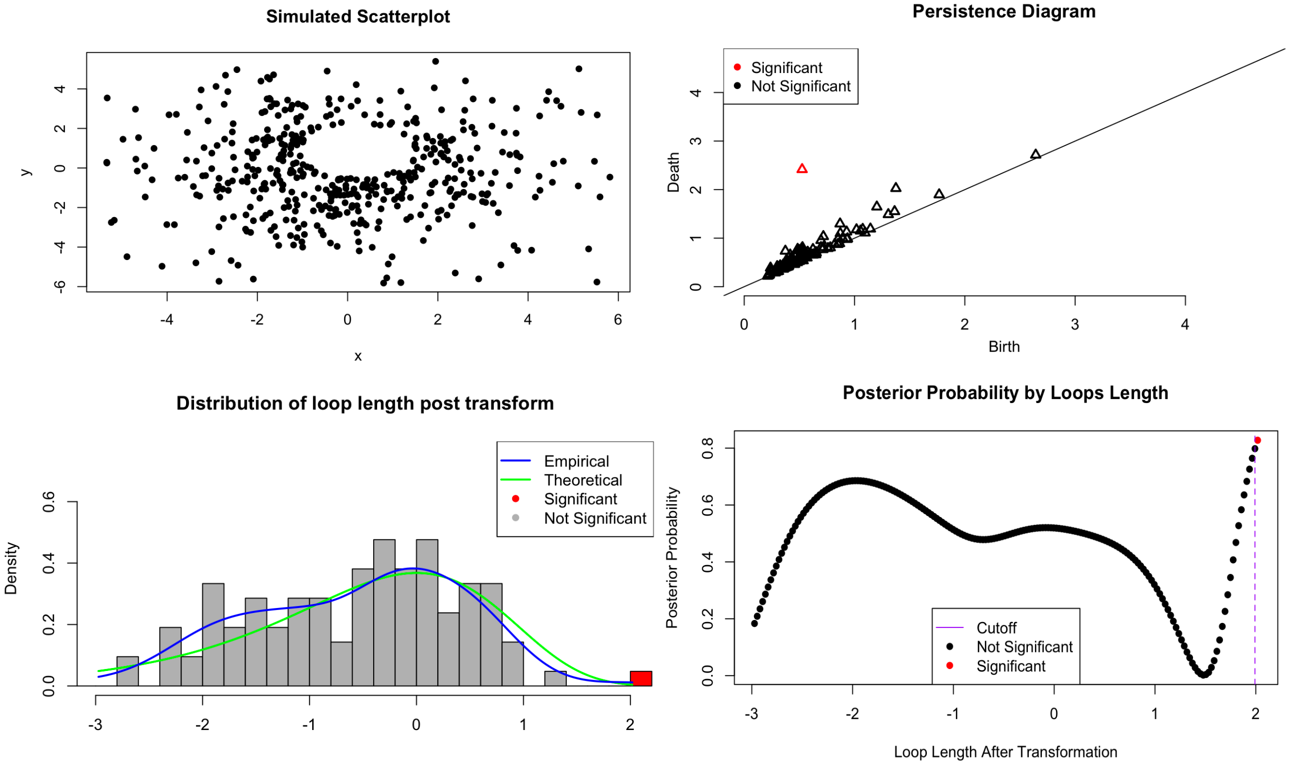

Figure 3. Example of EBGLIDE's determination of significance. We start with a dataset, represented by the simulated scatter plot. There is one significant loop in this case. The filtration process is applied to the simulated data, creating the persistence diagram in the upper right. The triangle far away from the $y = x$ diagonal is our one significant loop. The double log transform is applied to the loop persistence values, and the empirical and theoretical (LGumbel) distributions are calculated. We use the Empirical Bayes approach to calculate a resulting distribution of posterior probabilities, seen in the bottom right. We use the previously-described algorithm to determine a cutoff p value, which finally indicates our dataset contains one significant loop.

Table 1. Results from applying the various significance tests to the simulated data and comparing to ground truth. In addition to calculating both versions of our F1 score, we also calculate percent correct and the raw number of correct predictions at an individual loop level. These are the most important statistics for performance determination.

	alpha_0.3	alpha_0.5	EBGLIDE_0.01	EBGLIDE_0.05	null_0.05	null_0.5
Precision	0.428	0.128	0.851	0.697	0.999	0.986
Recall	0.584	0.803	0.781	0.800	0.386	0.826
F1	0.494	0.221	0.815	0.745	0.556	0.899
Binary Precision	0.947	0.713	0.900	0.873	0.999	0.989
Binary Recall	0.675	0.803	0.819	0.829	0.540	0.868
Binary F1	0.788	0.756	0.858	0.851	0.701	0.925
True Positives	2347	3225	3139	3215	1550	3318
False Positives	3127	21832	548	1397	1	47
True Negatives	516896	498191	519475	518626	520022	519976
False Negatives	1668	790	876	800	2465	697
Percent Correct	0.651	0.607	0.751	0.740	0.540	0.860

The results for the simulated data containing 5 and 10 significant loops are shown in Table 2. In general, all methods performed worse on this data compared to the original simulated data. However, EBGLIDE's performance decreased less than the other methods, indicating its versatility. Notably, the null method with $\alpha = 0.5$ experienced a significant drop in F1 score from 0.899 to 0.173, suggesting potential overfitting to the original dataset. In contrast, EBGLIDE with $\alpha = 0.01$ maintained a relatively higher F1 score of 0.204, demonstrating its robustness across varying numbers of significant loops. These findings further support EBGLIDE as a more effective and adaptable method for determining loop significance.

Table 2. Simplified results from testing the methods on simulated data with 5 and 10 loops. These results are primarily used to test for a method's flexibility outside the bounds of the original simulated data.

	alpha_0.3	alpha_0.5	EBGLIDE_0.01	EBGLIDE_0.05	null_0.05	null_0.5
Precision	0.674	0.558	1.00	1.00	1.00	1.00
Recall	0.078	0.275	0.114	0.206	0.001	0.094
F1	0.140	0.368	0.204	0.342	0.002	0.173
True Positives	147	516	214	388	2	178
False Positives	71	408	0	0	0	0
True Negatives	23176	22839	23247	23247	23247	23247
False Negatives	1728	1359	1661	1487	1873	1697

Discussion

The consideration of topological persistence as a data analysis tool is a relatively new phenomenon and as such, the rigorous tests for when a loop has persisted long enough to be considered 'significant' are few. In many cases, loops are considered 'significant' simply by looking at the persistence diagram and guessing which loops are significant and which are not. EBGLIDE provides a rigorous, consistent way to determine which loops are significant and which are not. Our discussion of results uses Table 1 as the primary gauge of effectiveness. In comparison to the other existing methods for determining significance, EBGLIDE scores higher.

Comparing EBGLIDE to the previously discussed alpha method, EBGLIDE outperforms in almost every metric. As expected, the alpha method with $\alpha = 0.3$ was the better performer of the alpha methods. However, when comparing that iteration with either implementation of EBGLIDE, we observe the comparative strength of EBGLIDE. EBGLIDE with $\alpha = 0.01$ achieves a much higher F1 score and slightly higher binary F1 score compared to the alpha method. Moreover, EBGLIDE consistently attains higher precision and recall scores, regardless of the alpha levels being compared. Additionally, EBGLIDE correctly identifies the number of loops significantly more often ($\geq$10%) than the best performing alpha version. The only noteworthy advantage of the 'alpha0.3' method is its slightly higher Binary Precision compared to the EBGLIDE implementations. Nonetheless, overall, EBGLIDE proves to be a much preferable method.

Comparing EBGLIDE with the previous "best" test, the 'null' significance test, at the $\alpha = 0.05$ level, it is observed that EBGLIDE outperforms the null method in all key summary statistics. Specifically, EBGLIDE achieves substantially higher scores in F1, Binary F1, and Percent Correct. Despite the null method being highly precise and achieving near-perfect Precision and Binary Precision values at both α-levels, it suffers in terms of recall score. At the $\alpha = 0.05$ level, the null method only calls one false positive. However, when considering aggregate scores, EBGLIDE performs better in all aspects. Preliminary testing revealed that the null method with $\alpha = 0.5$ attained extremely high scores in F1, Binary F1, and Percent Correct. To validate these findings, we included this method in our simulated runs, which confirmed the initial results. On the simulated data, the null method with $\alpha = 0.5$ obtained an F1 value of 0.899, Binary F1 of 0.868, and correctly identified the exact number of significant loops 86% of the time. At first glance, these scores might suggest that the null method with this specific α level is superior, rather than EBGLIDE. However, we consider an α level of 0.5 to be highly unrealistic for real data and interpret these results as indicative of overfitting the α level to this particular dataset. Importantly, when we standardize the α level between EBGLIDE and the null method, EBGLIDE significantly outperforms the null method. Nevertheless, for the sake of transparency, we include the results from $\alpha = 0.5$. To verify our hypothesis of overfitting, we conducted additional simulations on datasets containing a higher number of significant loops (5 and 10) to observe the responses of each method. While both methods exhibited worse performance compared to the original testing, the F1 score for the null method dropped significantly more than EBGLIDE 2. This further supports the notion that EBGLIDE is a more effective and versatile method compared to the null method.

In comparison to other methods for determining loop significance, EBGLIDE has been demonstrated to offer higher accuracy scores and greater versatility. However, it is worth noting that all methods tend to exhibit lower performance on datasets that feature a high number of significant loops (as indicated in Table 2). Therefore, there is a need for further development and refinement of methods in this domain. Nevertheless, EBGLIDE plays a crucial role in establishing a robust workflow for topological data analysis by offering a reliable and accurate score for hypothesis testing.

References

1. Palande, S. *et al.* Topological data analysis reveals a core gene expression backbone that defines form and function across flowering plants. *PLOS Biol.* **21**, 1–20, DOI: 10.1371/journal.pbio.3002397 (2023).

2. Bhaskar, D., Zhang, W. Y., Volkening, A., Sandstede, B. & Wong, I. Y. Topological data analysis of spatial patterning in heterogeneous cell populations: clustering and sorting with varying cell-cell adhesion. *npj Syst. Biol. Appl.* **9**, 43, DOI: 10.1038/s41540-023-00302-8 (2023).

3. Skaf, Y. & Laubenbacher, R. Topological data analysis in biomedicine: A review. *J. Biomed. Informatics* **130**, 104082, DOI: https://doi.org/10.1016/j.jbi.2022.104082 (2022).

4. McGee, R. L. *et al.* Topological Structures in the Space of Treatment-Naïve Patients with Chronic Lymphocytic Leukemia. *Cancers* **16**, 2662, DOI: 10.3390/cancers16152662 (2024).

5. Choudhary, A. Approximation algorithms for vietoris-rips and Čech filtrations. *Saarland Univ.* (2017).

6. Fasy, B. T., Kim, J., Lecci, F. & Maria, C. Introduction to the r package tda (2015). 1411.1830.

7. Bobrowski, O. & Skraba, P. A universal null-distribution for topological data analysis. *Sci. Reports* **13**, DOI: https://doi.org/10.1038/s41598-023-37842-2 (2023).

8. Efron, B. & Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* DOI: https://doi.org/10.1002/gepi.1124 (2002).

9. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284, DOI: 10.1038/s41467-017-02554-5 (2018). Publisher: Nature Publishing Group.

Acknowledgements (not compulsory)

Gus Gerlach completed part of this research thanks to BIDS@I2DB summer research program at Washington University in St. Louis.

Author contributions statement

Gus Gerlach and Jake Reed contributed equally to this work. Gus Gerlach and Jake Reed conceived of the method while Jake Reed performed all the simulations. Gus Gerlach implemented the method in R. Gus Gerlach wrote the initial draft of the manuscript. Kevin Coombes and Zachary Abrams provided critical feedback and helped shape the research, analysis and manuscript. All authors reviewed the manuscript.

Additional information

The authors declare no competing interests.