The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

TopoLoop: A new tool for chromatin loop detection in live cells via single-particle tracking

Cite as: J. Chem. Phys. 161, 204105 (2024); doi: 10.1063/5.0236090
Submitted: 30 August 2024 • Accepted: 28 October 2024 •
Published Online: 22 November 2024

Aryan Kokkanti,[1] Andrew Atanasiu,[1] Daniel Kolbin,[1] David Adalsteinsson,[2] Kerry Bloom,[1] and Paula A. Vasquez[3,a)]

AFFILIATIONS

[1]Department of Biology, University of North Carolina at Chapel Hill, 622 Fordham Hall, CB3280, Chapel Hill, North Carolina 27599, USA
[2]Department of Mathematics, University of North Carolina at Chapel Hill, 120 E Cameron Avenue, CB3250, Chapel Hill, North Carolina 27599, USA
[3]Department of Mathematics, University of South Carolina, 1523 Greene St., LC417, Columbia, South Carolina 29208, USA

Note: This paper is part of the JCP Special Topic on Chromatin Structure and Dynamics: Recent Advancements.
[a)]**Author to whom correspondence should be addressed:** paula@math.sc.edu

ABSTRACT

We present a novel method for identifying topological features of chromatin domains in live cells using single-particle tracking and topological data analysis (TDA). By applying TDA to particle trajectories, we can effectively detect complex spatial patterns, such as loops, that are often missed by traditional time series analysis. Using simulations of polymer bead–spring chains, we have validated the accuracy of our method and determined its limitations for detecting loops. Our approach offers a promising avenue for exploring the topological complexity of chromatin in living cells using TDA techniques.

INTRODUCTION

The discovery of cross-linking proteins that extrude chromatin loops has revolutionized our understanding of higher order chromatin organization. Loops are typically detected through the analysis of chromatin conformational states from fixed timepoints. Here, we propose a new methodology to identify loops from snapshots of single-particle positions or time-lapse series of single-particle trajectories in living cells.

Traditional time series analysis represents a particle's trajectory as a sequence of discrete points. While this approach effectively captures linear motion and overall displacement, it often falls short in detecting complex spatial patterns, such as loops. To address this challenge, we turn to topological data analysis (TDA). In contrast to traditional time-series analysis, TDA considers the entire particle trajectory as a continuous path embedded in a four-dimensional space (x, y, z, t) and then uses persistent homology for the detection of topological features such as holes or voids.

TDA is an emerging analytical technique that uses algebraic topology to uncover the underlying shape or structure inherent in data. At its core, TDA represents data as geometric objects and examines their topological features. By focusing on the data's shape, TDA can identify both local and global structures across multiple scales. This approach has found applications in diverse fields, including biology, medicine, economics, and image analysis. For instance, researchers are using TDA to investigate molecular structures, filamentous networks, analyze medical images, and detect anomalies within large datasets.[1–5] While TDA remains a developing field, it has shown promising results and is increasingly integrated with other data analysis methodologies. In this paper, we present an overview of the core TDA concepts essential for understanding our specific application: spatial detection of DNA loops in living cells. To evaluate the persistence of DNA loops over a range of scales, we use *persistent homology* (see Key Concept in the Explanatory Box). For a more in-depth treatment of the subject, readers are referred to Refs. 1–3 and 6.

When applied to spatial data, persistent homology characterizes spatial features of data based on proximity and positions of data points. The more persistent a particular feature is over a range of proximities (distances), the more resolved and defined the feature is.

Persistent homology allows us to transform data from single particle tracking (collection of points in space and/or time) into topological data to provide insight into underlying spatial patterns. To achieve this, it employs three core concepts: simplicial complexes, filtrations, and topological summaries. A **simplicial complex** is a geometric structure representing data as a collection of points, lines, triangles, and higher-dimensional shapes (see Key Concept in the Explanatory Box). By constructing a sequence of simplicial complexes at increasing levels of detail, **filtration** allows for the study of how topological features evolve. **Persistence diagrams** visualize the birth and death of these features across the filtration, providing insights into the data's underlying shape.

The filtration process used in the methods in this paper is the Vietoris–Rips filtration. Panel (a) of Fig. 1 illustrates this filtration process applied to a point cloud. The method involves surrounding each point with a circle and systematically enlarging its radius. Connections are established based on the intersection of these enclosing circles. This expansion transforms the data in the example from isolated points to connected lines, to a single connected component with a hole. In general, by tracking changes in homology groups throughout this process, the persistence of topological features, such as the birth and death of holes, can be quantified.

Figure 2 demonstrates TDA's ability to identify and differentiate loop structures of varying sizes within a point cloud. By tracking topological changes through a filtration process and analyzing the resulting persistence diagram, the figure reveals the presence of both a larger and a smaller loop, quantified by their respective persistence values.

Figures 1 and 2 showcase the potential of TDA in identifying loop structures within static datasets. However, when applying this methodology to single-particle tracking experiments, a critical limitation arises. Unlike static datasets, which allow for the observation of *multiple* particles simultaneously, single-particle experiments generate time-series data that capture the trajectory of an *individual* particle over time. This presents a challenge in determining whether the particle is part of a loop as it evolves through space and time. To overcome this limitation, we propose a novel approach that combines the simulation dynamics of chromatin chains with TDA-based methods to detect the existence of loops.

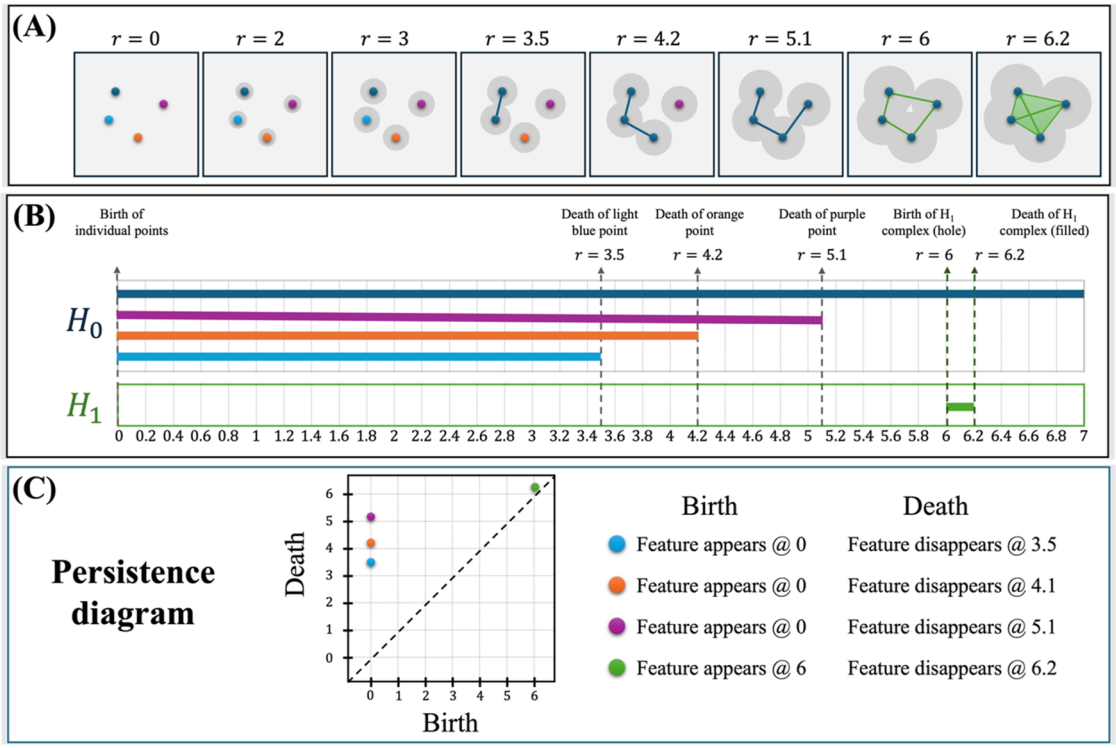


FIG. 1. (a) Initially, each data point is considered an independent connected component (H_0). The filtration process involves expanding circles around each point with an increasing radius. When two circles intersect, the corresponding points become connected, forming a new component (H_0), for example at $r = 3.5$. This process continues, with components merging as their enclosing circles overlap. The emergence of a closed curve formed by overlapping circles signifies the birth of a one-dimensional hole (H_1). As these circles further expand and overlap, the hole is filled, marking its death. For instance, in the figure, a hole is born at radius $r = 6$ and dies at $r = 6.2$. (b) Graphical representation of the birth and death of H_0 and H_1 features in panel (a). (c) A persistence diagram plots points corresponding to the birth, on the x-axis, and the death, on the y-axis, of each feature within the dataset.

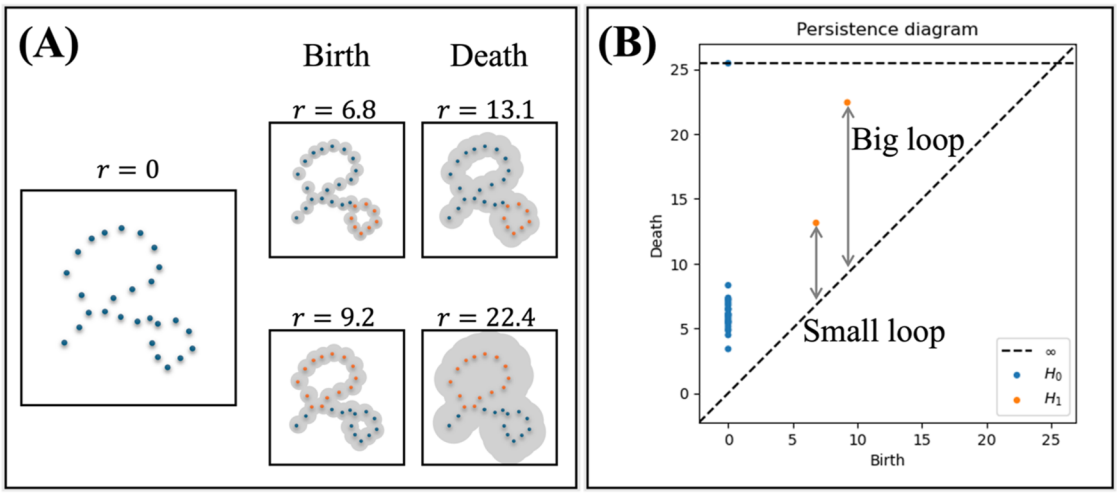


FIG. 2. Identification of loops within a point cloud using persistent homology. (a) Filtration process to determine the birth and death of each of the two loops. (b) Resulting persistence diagram with the birth and death of topological features (H_0: connected components and H_1: loops). Arrows indicate the persistence of each feature.

Key Concepts

Geometry and Topology

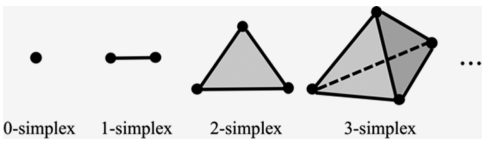
Topology studies how shapes can be completely deformed without losing their core properties. For example, a coffee mug and a donut are considered topologically equivalent because they can be transformed into each other through continuous deformations.

Homology is a topological tool for classifying spaces based on their shape by identifying and counting holes. Spaces with identical hole structures, called homology groups, are considered topologically equivalent.

Homology groups capture information about holes. They classify these holes by dimension: H_0 represents the number of connected components, H_1 quantifies one-dimensional holes (such as circles), H_2 signifies two-dimensional holes (such as voids), and so on for higher dimensions.

Simplex is a basic shape in topology. It can be a point, line, triangle, or higher-dimensional equivalent, defined by its corners (vertices).

A **simplicial complex** is a collection of simplexes.



Persistent homology is a method for studying the topological features of data that persist over a range of scales. It tracks the birth and death of topological features (such as holes) as the data is filtered.

Filtration is a nested sequence of simplicial complexes, essentially a gradual construction of a shape by incrementally adding simplexes. In persistent homology, filtrations are employed to extract topological signatures from data. Figure 2 illustrates a filtration process where the enclosing radius, determining the size of circles centered on data points, serves as the **filtration parameter**. As this radius increases, the point cloud evolves from individual points (H_0 homology group) to a structure of lines (still H_0, i.e., no holes) and finally to a connected polygon with a single one-dimensional hole (H_1 homology group at $r = 6$). By tracking changes in the homology, the persistence of topological features throughout the filtration can be quantified. Although the Vietoris–Rips filtration, which is used in this paper, is widely adopted, there are other filtration methods available. Two examples are the Čech filtration and the alpha-complex filtration. The choice of filtration depends on the specific application and the type of data being analyzed.[7]

A **persistence diagram** is a graphical representation of persistent homology data, showing the birth and death of topological features.

Polymer Physics

A **bead–spring model** treats chromatin as a WLC (worm-like chain), a statistical mechanical representation of long-chain polymer. The chain consists of a series of beads connected via Hookean springs. Hinge-like forces between each bead are parameterized to give the strands the same bending rigidity of DNA, that is, a persistence length of 50 nm.[8]

Persistence length is a statistical mechanical measure of the stiffness of a polymer chain. It is the length scale over which the polymer's orientation is correlated, defined by the decay of the autocorrelation function $\cos(\theta) = \exp[-L/L_p]$ of the tangent vector along the chain. Here, θ is the angle between two points separated by a distance L, and the chain's persistence length is L_p.

Tether resistance: End beads of the chromatin chain are defined as tethers that constrain (provide resistance to) the movement of the entire chain via a viscous drag. In simulation, forces

acting on tethered beads are scaled down by dividing by their tether resistance (unitless quantity).

A condensin spring is a Hookean spring that cross-links two beads on the chain when they are within a prescribed distance. This spring represents an action of Structural maintenance of chromosomes (SMC) protein forming a chromatin loop.[9]

ChromoShake is a simulation package written in C++[10] and updated to run with ImageTank,[9,11,12] a GUI package from Visual Data Tools, Inc. All simulated loop data herein is found in Kolbin *et al.*[9]

RESULTS

Topological data analysis can detect loops with positional data of an entire chain

To assess the generalizability of TDA functions for loop identification beyond the idealized scenario depicted in Fig. 2, we use a bead–spring polymer model.[9] By using simulated data, we can (1) identify bona-fide loops, (2) modify loop size and duration, and (3) by comparing the TDA-identified loops with the known loop position from the model, evaluate the accuracy and robustness of the TDA methodology. For this study, we have used a chain of 101 beads, altering the chain properties as described in the Key

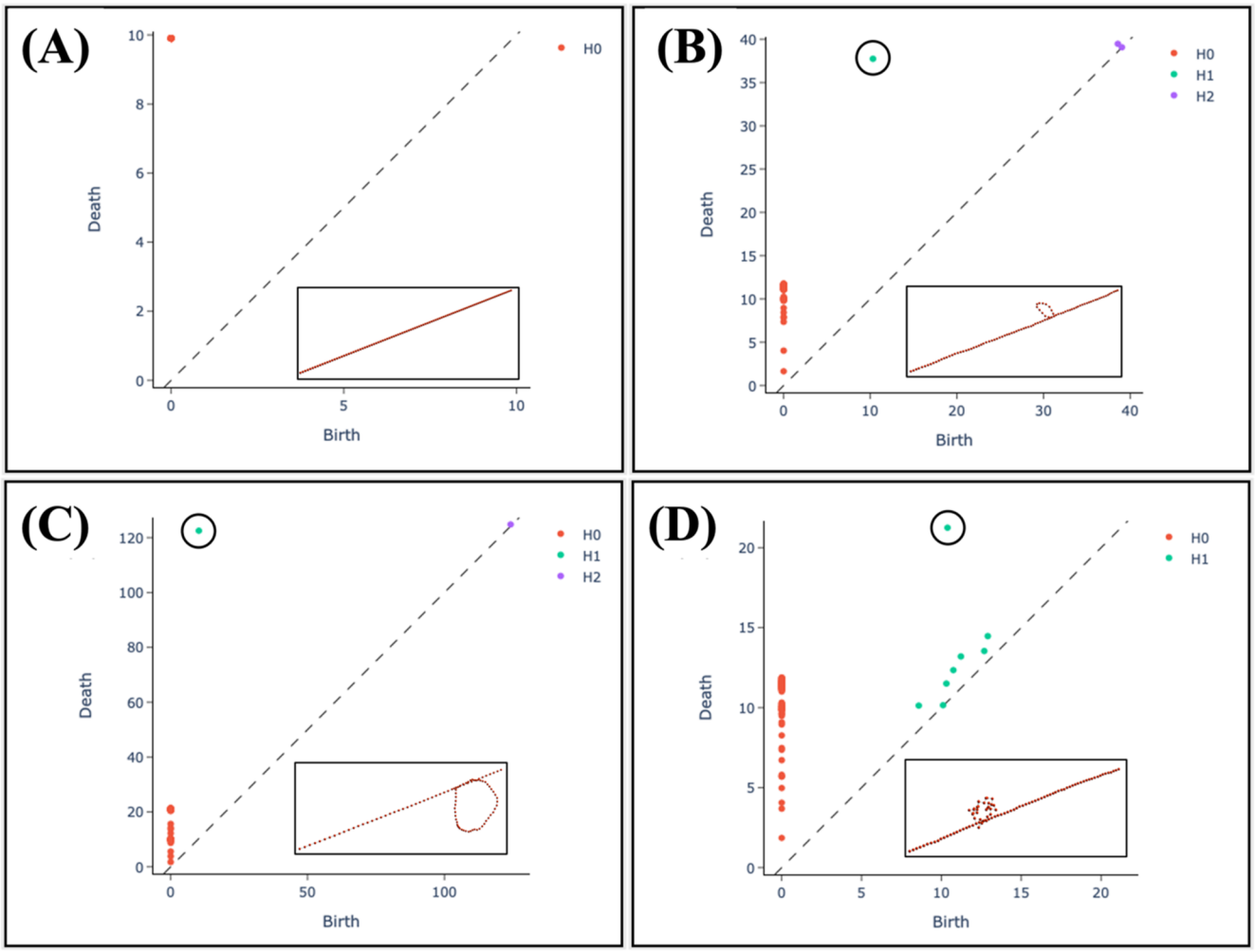


FIG. 3. Examples of modeled loops and associated persistence diagrams. (a) Representative of the initial configuration of the chain before simulation. (b) and (c) Representative of small and large loops in the chain, respectively. (d) Representative of a floppy loop within the chain.

Concept (Polymer physics). Loop size and duration can be controlled by changing the tethering resistance of the chain (e.g., increasing the weight of the ends), changing the spring strength (e.g., changing the force connecting beads that are cross-linked), and changing the persistence length (e.g., changing the degree of stiffness). For a more comprehensive understanding of the effects of these parameters, the reader is referred to Kolbin *et al.*[9] The initial configuration is a linear, tethered chain, where both ends remain fixed throughout the simulation. In addition to the beads representing the main chain, the simulation incorporates a cross-linker spring that connects two non-adjacent beads. The cross-linker extrudes loops as it traverses the chain.[9]

To evaluate the accuracy and robustness of the TDA procedure, we construct a point cloud from the positions of all beads from a simulation at a timepoint when loops are present. Figure 3 shows four persistence diagrams with detected features (loops), highlighting the largest H_1 features in each diagram. Figure 3(a) shows a linear chain with no loops, as confirmed by the absence of any H_1 (green) points in its corresponding persistence diagram. Conversely, introducing a loop, as depicted in Fig. 3(b), results in a distinct H_1 point offset from the diagonal, indicating the presence of a topological hole associated with the loop. Figure 3(c) further demonstrates this relationship for a larger loop, characterized by a point located farther from the main diagonal. Finally, Fig. 3(d) depicts a floppy loop configuration, which is evident from the presence of more than one H_1 point. The points closer to the diagonal are referred to as "noise," since they correspond to topological features with negligible persistence. However, despite this increased noise, TDA successfully identifies one primary loop structure.

While TDA effectively distinguishes between structures with and without loops, it currently lacks the capability to directly quantify loop dimensions. These dimensions are influenced by parameters such as chain stiffness and cross-linker spring strength, which dictate the size and persistence of the loops. To further quantify loop size, we introduce the concept of **amplitude or lifespan**, a metric that measures the persistence of a topological feature within a dataset. The lifespan is calculated as the difference between a feature's birth and death, and it provides a numerical value for loop size. Analysis of the primary H_1 points in Figs. 3(b)–3(d) reveals single-point lifespans of 27.43, 112.2, and 10.87, respectively, suggesting a correlation between lifespan and loop size.

Single-particle tracking offers an experimentally accessible means of analysis

While having the positional data of all beads of a chain across time in the simulation provides a robust means of tracking loops, this is not a practical approach for tracking loops in living cells. Single-particle tracking, either through sparse labeling of chromatin proteins such as histones or the introduction of integrated DNA-binding sites and visualization with a fluorescent binding protein (FROS-fluorescence repressor-operator system), allows analysis of a single segment of the DNA chain over time. To simulate experimental conditions, we computationally labeled a single bead roughly in

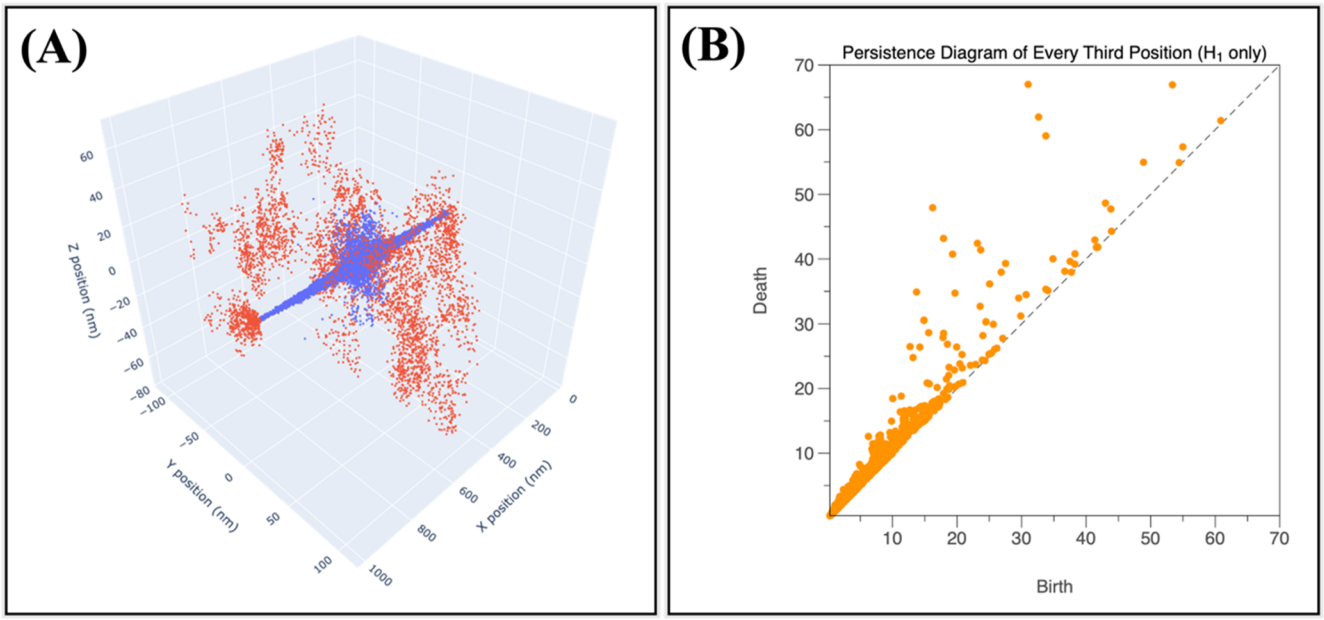


FIG. 4. Example analysis of simulated single-particle tracking. (a) Each point within the point cloud represents a position occupied by the bead at a specific timepoint during the simulation. The single bead is represented by a red point when it is part of a loop and a blue point when it is not. (b) Persistence diagram associated with the point cloud in (a), showing only the persistence of H_1 features.

the middle of the chain (position 50 out of 101 beads) and tracked its position over time, generating a point cloud representing 17 500 time steps as shown in Fig. 4(a).

For this 3D point cloud, we compute its corresponding persistence diagram using every third time step, which is depicted in Fig. 4(b). When a loop encompasses the beads, they are displaced from the chain's main axis, which can be associated with more displacement, which in turn leads to an increased probability of forming topological holes in the spatial data. This is captured in the persistence diagram by points away from the main diagonal. However, given the complexity of the point cloud, it is challenging to directly correlate a given point in the persistence diagram to fluctuations in the linear chain, short-lived loop-like structures, or bona-fide persistent loops over the entire time series. To address this, we employ a sliding window approach to generate persistence diagrams at discrete time intervals throughout the simulation, as described in the next section.

Sliding windows and lifespan allow for finding loops through time

To accurately capture the dynamic nature of individual particle trajectories, we introduce TopoLoop, a pipeline that employs a sliding window technique. This approach generates a temporal lifespan plot by calculating persistence diagrams at discrete time intervals, as illustrated in Fig. 5.

To analyze our polymer simulation data, we divided the time series into overlapping windows, defined by window size and step size. Window size determines the length of each analysis window, while step size specifies the amount by which the window is shifted for subsequent analyses. For instance, using a window size of 100 and a step size of 20, we create a series of overlapping windows: [0, 100], [20, 120], [40, 140], and so on. Each window was then processed to generate a persistence diagram. The lifespan of a persistence diagram is computed by taking the norm of the vector representing the lifespans of each H_1 point within the diagram. For this, we use the "amplitude" function from the Giotto-tda package with a bottleneck metric.[13,14] These calculations use the bottleneck (L_∞) norm of the distance between the given persistence diagram and the trivial diagonal diagram, which consists solely of points on the diagonal line where $r_{\text{birth}} = r_{\text{death}}$. In Giotto-tda, this option can be selected by setting *metric* = "*bottleneck*" and *order* = "*None*." For other options, see the Giotto-tda documentation.[14]

To visualize the temporal evolution of loop formation, we then associated each calculated lifespan with the midpoint of its corresponding time window. For example, if the lifespan of a window spanning timepoints 0–100 is 3.47, we generate the ordered pair (50, 3.47). By connecting these points throughout the entire time series, we generate the temporal lifespan plot shown in the bottom left of Fig. 5.

In order to determine whether a bead is within a loop, we consider a **persistence threshold**. Timepoints with lifespans exceeding this threshold are indicative of loop presence. The value of the threshold plays a significant role in TopoLoop's ability to identify loops. As shown in the bottom left plot of Fig. 5, adjusting the dashed line (threshold) will directly influence the number of peaks

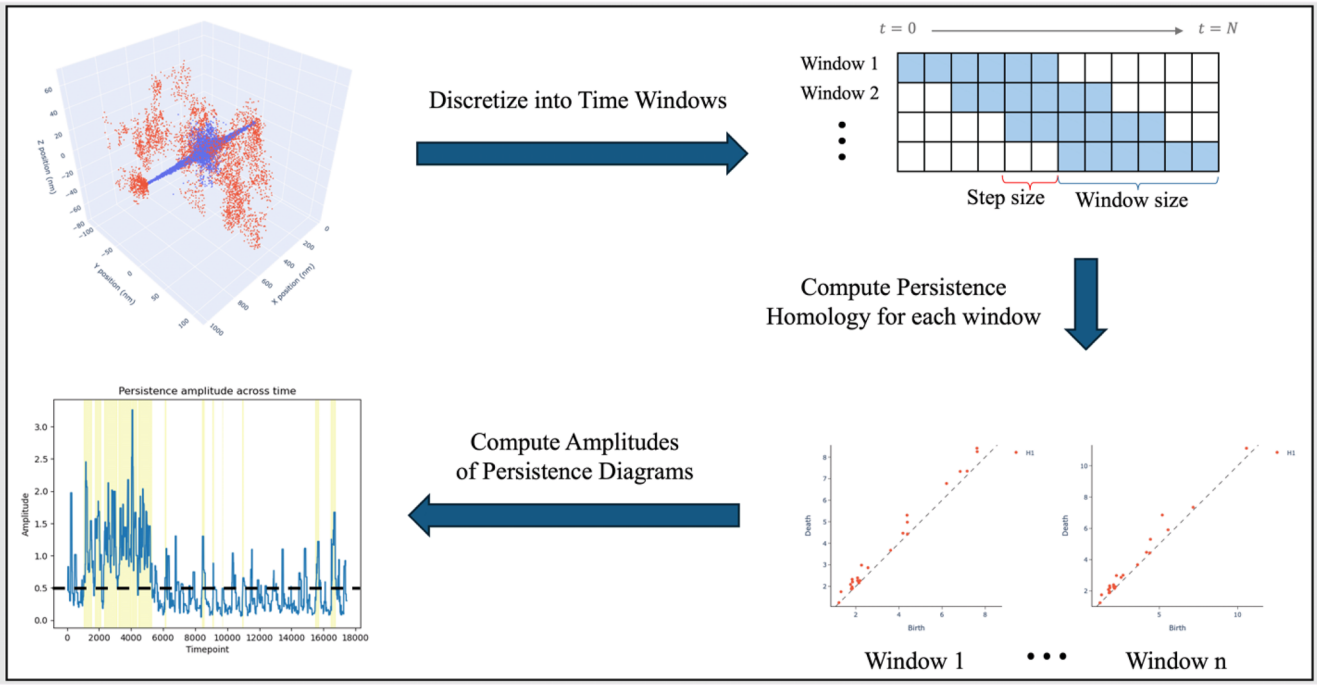


FIG. 5. Overview of TopoLoop pipeline. TopoLoop takes an input of position-time cloud from single-particle tracking data and predicts the times that a particle is in a chromatin loop via the sliding window, persistent homology, and lifespan. To evaluate the accuracy of our method, we indicate the times when the bead is in a loop based on our simulated data by highlighting them in yellow in the bottom-left plot, and we set a threshold of 0.5.

considered as loop presence. A higher threshold will result in fewer peaks being identified as loops, while a lower threshold will lead to more peaks being misclassified as loops. In addition, the chosen window and step size also impact the overall accuracy of loop detection.

To evaluate the accuracy of TopoLoop, we compare its predicted loop events to the actual loop occurrences, which are known from the simulation data and highlighted in yellow in the bottom left plot of Fig. 5. We calculate **accuracy** by determining the ratio of correct predictions to the total number of predictions. Predictions are considered to be the associated times of the time-lifespan ordered pairs where the lifespan is above the persistence threshold. Through systematic testing of various combinations of window size, step size, and persistence threshold, we have identified the optimal parameters that achieve an accuracy rate exceeding 70% across a representative set of chain configurations (Fig. S1 of the supplementary material). We note that these parameters are specific to the simulations, which strongly conform to what occurs *in vivo*. However, the temporal and spatial resolution of the simulations may not be representative of all microscopes, which can affect data collection and may require different parameters than the ones we outlined. These optimized parameters are used in all the analyses presented below.

Loop detection accuracy and chain parameters

The power of simulated datasets is the ability to define quantitative aspects of loop size and duration that can be reliably detected. As an initial foray into these parameters, we varied the three drivers of loop formation, namely, anchors or tethers that restrict chain motion (tethering resistance), the strength of the cross-linking spring (condensin moduli), and the stiffness of the chromatin chain (L_p). The stiffness of chromatin ranges experimentally from 5 to 220.[9,15] The strength of the condensin spring in Gigapascals (GPa) is based on Young's modulus of a coiled-coil protein (2 GPa Howard Mechanics of Motor Proteins and the Cytoskeleton) and fitting simulations of loop extrusion to experimental data.[9,15] Decreasing the strength of the spring in simulation is informed by the biological consequences of energy and temperature that soften these materials in live cells.[16] Tethering resistance was determined by the resistive force on a bead in a bead–spring model. The resistive force delays the time required for the chain to adopt a random coil. The forces were tuned to parameters that simulate *in vivo* relaxation time.[17] TopoLoop accuracy for variations of these three parameters is shown in Fig. 6.

Figure 6(a) shows that increasing tethering resistance significantly enhances overall TopoLoop's accuracy. While Fig. 6(b) shows that variations in condensin modulus can influence accuracy, the effect is less pronounced compared to tethering resistance. Notably, a significant increase in the number of accurate TopoLoop predictions is observed at moduli above 0.02. Finally, Fig. 6(c) shows that a smaller persistence length (L_p), indicative of a floppier loop, results in more accurate TopoLoop predictions compared to more stiff loops. We conclude that TopoLoop is most accurate when combinations of chain parameters include high tethering resistance, stiffer condensin spring, and small persistence length of the chain.

Our findings are consistent with the fundamental physical principles governing the system. A higher tethering resistance restricts overall chain movement, promoting more stable loop structures and increasing the likelihood of detection. Similarly, a higher condensin modulus brings the beads at the base of the loop closer together, facilitating loop identification. Finally, a smaller persistence length, indicative of a more flexible chain, allows for greater conformational freedom, potentially enhancing loop detection accuracy by enabling a given bead to explore a wider range of positions.

Projection to two dimensions maintains accuracy

Given the higher resolution of microscope single particle tracking when limiting collection to two dimensions, we aimed to test the effectiveness of TopoLoop with 2D data to mimic experimentally achievable data. To evaluate TopoLoop's performance with

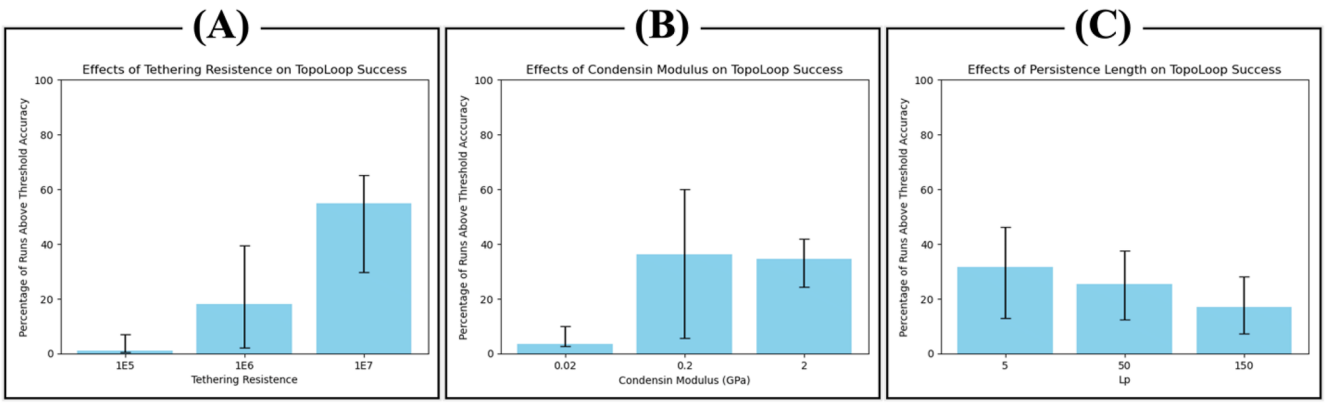


FIG. 6. Each panel shows the percentage of runs that achieve an accuracy of 70% or higher varies for different combinations of the corresponding chain parameter: (a) tethering resistance, (b) condensin modulus, and (c) persistence length. The error bars represent the range of accuracy, with the upper bound corresponding to 50% accuracy and the lower bound to 90% accuracy. A total of 1620 runs were analyzed (20 replicates of 27 parameter combinations across beads 10, 25, and 50), with each bar in the graph representing 540 runs (nine specific parameter combinations for each bar).

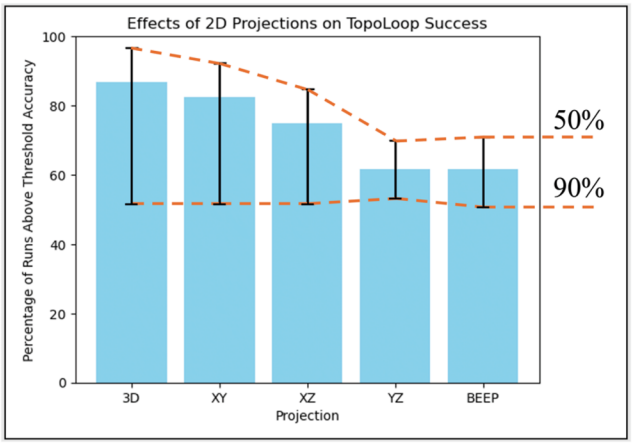


FIG. 7. Projection analysis of TopoLoop. We measure TopoLoop's success when projecting 3D coordinates to 2D using simple projections and the BEEP projection. This was used to imitate the 2D data that can be created from single particle tracking in a microscope. We analyze runs with tethering resistance = 1×10^7, $L_p = 5$, and condensin spring strength = 0.2 or 2 GPa (total of 120 runs: 20 replicates of two parameter combinations for beads 10, 25, and 50). Bars are indicative of a threshold accuracy set to 70%, and the lower and upper errors are set to 90% and 50%, respectively.

2D data, we conducted simulations using two projection methods: simple projection onto the (x, y), (x, z), or (y, z) plane, and a BEEP (Bead-End-End Plane) projection onto a plane defined by the spatial coordinates of the bead and both chain ends at every timepoint. We use the following chain parameters to define our dataset and measure the average accuracy for each projection: tethering resistance = 1×10^7, condensin modulus = 0.2 or 2 GPa, and $L_p = 5$. These results are shown in Fig. 7.

Figure 7 shows that projecting down to XY and XZ planes does not significantly reduce the number of runs that exceed the accuracy threshold (70%) while projecting onto YZ or BEEP planes leads to a notable decrease. We attribute the reduction in accuracy to the usage of high tethering resistance and note that with this parameter, a central axis is more prominently formed on the X-axis, which is accounted for in the XYZ, XY, and XZ projections but not directly in the YZ and BEEP projections. However, despite this lower overall success rate for YZ and BEEP projections, when they do identify a loop, the accuracy is still high, as demonstrated by the relatively similar levels of 90% accuracy across the different projections. This suggests that while these projection methods might be less sensitive in terms of identifying loops overall, they exhibit high precision when a loop is correctly detected.

Potential of TDA for *in vivo* experiments

Our current method shows promise in analyzing timelapse data, but it requires a substantial number of timepoints. However, obtaining these timepoints can be challenging due to limitations in optical imaging, such as photobleaching. In order to demonstrate the potential of TDA for analyzing experimental data without a time element, we decided to investigate a different dataset.

In this study, we placed a LacO (*E. coli* lac operator array) at 1.8 kb from the centromere in chromosome XV (CEN15) in a strain of budding yeast that only had two chromosomes, which we referred to as the WT. The LacO array was visualized with the lac repressor (LacI) fused to GFP (green fluorescent protein). We chose this

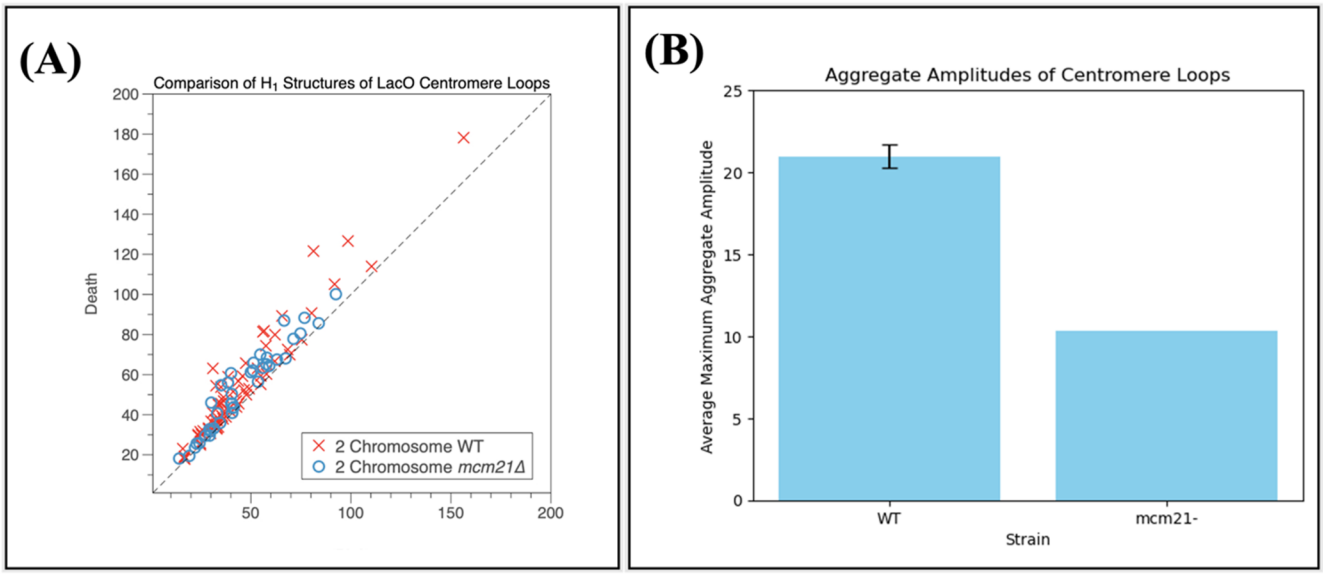


FIG. 8. Persistence diagram and lifespans of the centromere loop. (a) (x, y) positions of the LacO in the loop were determined as outlined, and analysis via TDA was performed to produce H_1 persistence diagrams. (b) The lifespan was calculated for both WT and mcm21Δ using the outlined sampling method. Error bars on WT are indicative of a 95% confidence interval. No error bars were generated for mcm21 since the entire dataset was analyzed simultaneously rather than sampled numerous times.

genomic location because it is known to form long-lasting and stable loops during metaphase, which is the same phase of the cell cycle where our images were captured. In addition, we marked the spindle pole body with RFP and used it as a reference point for tracking the position of the LacO marker.

By tracking the (x, y) positions of each LacO marker in different cells, we were able to compile a two-dimensional point cloud for subsequent TDA analysis. To further explore the role of loop formation at this locus, we created another strain with a deleted MCM21 gene (mcm21Δ). This gene encodes a protein that recruits cohesin to the centromere, which in turn promotes loop formation at this locus.

We calculated the lifespan of the full dataset (197 observations) for the mcm21 delete strain. To ensure a consistent number of observations, we randomly sampled 197 observations from the full dataset of WT (341 observations) and repeated this process 100 times. Figure 8 compares the lifespans of the mcm21Δ to the average lifespan of WT. Notably, the WT strain exhibited an approximately twofold increase in lifespan compared to the mcm21Δ strain. We also note that the sampling method of the WT results in consistency with a TDA of the full dataset (lifespan = 20.2). This difference suggests the presence of smaller, more unstable loops, consistent with previous findings in mcm21Δ.

To further confirm that the difference in lifespan, as determined by TDA, is indicative of weakened spring complexes forming the loops, we examined the maximum lifespan achieved for condensin moduli of 0.02, 0.2, and 2 GPa across various combinations of tethering resistance and L_p (Fig. 2 of the supplementary material). Our findings revealed a significant reduction in lifespan at 0.02 GPa compared to 0.2 and 2 GPa, typically by a factor of 2–3, across most tethering resistance and L_p combinations except for 1×10^5 tethering resistance. This observation supports our experimental results and confirms that the difference in lifespan may be due to a reduction in spring strength rather than random fluctuations.

CONCLUSIONS

Chromatin loops are a pervasive feature of higher order chromosome organization found throughout phylogeny, including pro- and eukaryotes. Loop detection is largely limited to inferences from fixed cell analysis through techniques that allow quantitation of proximal but non-contiguous DNAs. The ability to detect loops in living cells on a genome-wide scale would markedly advance the study of loop formation, persistence, and destruction. Topological Data Analysis (TDA) is a relatively recent approach (circa 1990's) to analyze the shape of data. It is based on algebraic topology and posits that persistent shapes in a dataset are diagnostic of critical relationships. We reasoned that TDA might provide a means to identify loops from a set of single-particle trajectories. In order to rigorously test our hypothesis, we demonstrate the power of TDA to identify loops from simulated datasets consisting of tens of thousands of data points of a fluctuating bead–spring chain.

Loops are defined as regions of DNA spatially segregated from the chromosome through the action of cross-linking proteins that bring non-contiguous regions together. DNA in a loop might be a region of the chromosome that needs to be transcribed, repaired, or recombined. It is a way for the cell to control DNA regions on a scale larger than a gene. Control can be through over- or underwinding the DNA or through gene clustering, where transient or long-lived networks can be assembled and disassembled. The life cycle of these loops is, therefore, critical for understanding dynamic features of genome organization and emergent structures that ensue from the sum of short-lived spatially regulated interactions.

In TDA parlance, a loop is a topological feature defined by a set of data points enclosing a hole within a point cloud. The persistence of a loop, a measure of its size and stiffness, is determined by its ability to maintain its topological identity as the radius used to define points within the point cloud increases. Persistent homology is the computational tool used to quantify this persistence, and persistence diagrams (birth vs death plots) are used to visualize it.

Examination of a loop along a chain of fixed length is readily apparent as an off-diagonal position in a persistence diagram (Fig. 3). In cells, there are three major parameters that dictate the behavior of loops. These are the stiffness of the chain (L_p), tethers throughout the genome (e.g., anchors), and the strength of the cross-linking spring (SMC spring constant). Our simulations enabled us to assess the efficacy of TDA in loop detection by varying the values of these parameters. As shown in Fig. 6, the strength of anchors is the single most significant driver of accuracy in loop prediction [Fig. 6(a)]. Over a range of simulated runs, we can attain up to 90% accuracy in loop detection when the strength of the anchor is high. In addition, the strength of the cross-linking spring has a major influence on loop detection [Fig. 6(b)]. With very weak springs (0.02 GPa), loops are very small and short-lived and are not reliably detectable. Increasing the spring strength increases the ability to accurately detect the loops. Chain stiffness (increasing L_p), however, was not a strong driver of loop detection. Over a range of chain stiffness, the accuracy of loop detection was comparable [Fig. 6(c)].

We demonstrated that TopoLoop can effectively identify loops in our bead–spring model. To determine if TDA could also identify genuine loops *in vivo*, we examined a region of pericentromere DNA known to contain a high density of loops. As shown in Fig. 8, we observed persistent shapes (births > deaths). The persistent lifespan for this dataset (n = 341 observations) was ~20. To confirm the identity of loops, the same analysis was performed in cells lacking a key protein that recruits the cross-linker cohesin (mcm21Δ). In the dataset from the mutants (n = 197 observations), the persistent lifespan decreased by a factor of 2 (from 20 to 10). Thus, in an experimental system, TopoLoop provides an unprecedented means for loop identification in living cells.

TopoLoop offers a significant advancement in our ability to analyze chromatin motion. While mean square displacement (MSD) provides insights into the nature and range of motion, TopoLoop introduces a novel approach by providing geometric information and the capability to distinguish movement within a loop. TopoLoop can be applied to various datasets, including fluorescent repressor-operator spots (FROS) and spots from single-particle tracking of DNA-binding proteins. Similar to MSD analysis, the constraints on TopoLoop include spot intensity, susceptibility to photobleaching, and tracking accuracy. The method is applicable to both single time points or fixed cells, as well as time-lapse series from living cells. Given the recent findings highlighting the role of loops as a major organizing principle of chromosomes for gene expression, chromatin compaction, and centromere function, methodologies such as TopoLoop become increasingly valuable for examining these loops at the single-cell level.

MATERIALS AND METHODS

Simulation of chromatin chains and loops

ChromoShake is a three-dimensional chromatin simulator designed to find the thermodynamically favored states for assigned geometries.[10] The simulator represents a one-micron segment of chromatin as a linear chain comprised of 101 beads and connected by Hookean springs. The first and last beads of the chains act as tethers and constrain chain motions to a degree determined by tethering resistance (see Key Concepts). Additional springs representing condensins are stochastic, and dynamic cross-linkers provide the means to study the physical properties of chromatin that influence the size, duration, and translocation of loops along the DNA.[9] A single cross-linking spring is seeded randomly on the bead–spring chain and is active for a duration representative of the lifetime of condensin on the DNA. The spring traverses the chain by stepping in a random direction. It can crosslink distant parts of the chain in a single step, intermittently stall, or unbind. The cross-linking spring is responsive to chromatin tension and extends to its maximum length. As shown previously, the drivers of loop behavior are the tethering resistance of the end beads, the persistence length of the DNA, and the strength of the cross-linking spring.[9] *In vivo*, the drivers are subject to modifications and cellular processes and, thus, depict physiologically relevant states. Our dataset obtained in Ref. 9 considers 27 combinations of three parameters (three values for tethering resistance, three values for condensin spring strength, and three values for DNA persistence length). Each combination is run in 20 replicates for 540 total simulations. For TDA analysis, either a snapshot of x, y, and z positions of all beads at a single timepoint or a position of a single bead across time scales is considered.

SUPPLEMENTARY MATERIAL

See the supplementary material for Fig. S1 shows parameter optimization for the sliding window technique. Figure S2 illustrates the impact of chain parameters on loop detection. Video 1 demonstrates the sliding window approach on a simulated bead–spring chain.

ACKNOWLEDGMENTS

We acknowledge the support from NIH R01 Grant No. GM32238 to K.B. and Grant No. NSF-DMS-1751339 to P.V.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Aryan Kokkanti: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Andrew Atanasiu**: Software (equal); Visualization (equal); Writing – original draft (equal). **Daniel Kolbin**: Data curation (equal); Visualization (equal); Writing – original draft (equal). **David Adalsteinsson**: Software (equal). **Kerry Bloom**: Funding acquisition (equal); Investigation (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Paula A. Vasquez**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] E. J. Amézquita, M. Y. Quigley, T. Ophelders, E. Munch, and D. H. Chitwood, "The shape of things to come: Topological data analysis and biology, from molecules to organisms," Dev. Dyn. **249**, 816–833 (2020).
[2] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," Front. Artif. Intell. **4**, 667963 (2021).
[3] Y. Skaf and R. Laubenbacher, "Topological data analysis in biomedicine: A review," J. Biomed. Inf. **130**, 104082 (2022).
[4] A. Bukkuri, N. Andor, and I. K. Darcy, "Applications of topological data analysis in oncology," Front. Artif. Intell. **4**, 659037 (2021).
[5] M.-V. Ciocanel, R. Juenemann, A. T. Dawes, and S. A. McKinley, "Topological data analysis approaches to uncovering the timing of ring structure onset in filamentous networks," Bull. Math. Biol. **83**, 21 (2021).
[6] L. Wasserman, "Topological data analysis," Annu. Rev. Stat. Appl. **5**, 501–532 (2018).
[7] A. Zomorodian and G. Carlsson, "Computing persistent homology," in *Proceedings of the 20th Annual Symposium on Computational Geometry (ACM)* (Springer Nature, 2004), pp. 347–356.
[8] K. S. Bloom, "Beyond the code: The mechanical properties of DNA as they relate to mitosis," Chromosoma **117**, 103–110 (2008).
[9] D. Kolbin, B. L. Walker, C. Hult, J. D. Stanton, D. Adalsteinsson, M. G. Forest, and K. Bloom, "Polymer modeling reveals interplay between physical properties of chromosomal DNA and the size and distribution of condensin-based chromatin loops," Genes **14**, 2193 (2023).
[10] J. Lawrimore, J. K. Aicher, P. Hahn, A. Fulp, B. Kompa, L. Vicci, M. Falvo, R. M. Taylor, and K. Bloom, "ChromoShake: A chromosome dynamics simulator reveals that chromatin loops stiffen centromeric chromatin," Mol. Biol. Cell **27**, 153–166 (2016).
[11] Y. He, J. Lawrimore, D. Cook, E. E. Van Gorder, S. C. De Larimat, D. Adalsteinsson, M. G. Forest, and K. Bloom, "Statistical mechanics of chromosomes: *In vivo* and *in silico* approaches reveal high-level organization and structure arise exclusively through mechanical feedback between loop extruders and chromatin substrate properties," Nucleic Acids Res. **48**, 11284–11303 (2020).
[12] Y. He, D. Adalsteinsson, B. Walker, J. Lawrimore, M. G. Forest, and K. Bloom, "Simulating dynamic chromosome compaction: Methods for bridging in silico to in vivo," Methods Mol. Biol. **2415**, 211–220 (2022).
[13] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. M. Medina-Mardones, A. Dassatti, and K. Hess, "giotto-tda: A topological data analysis toolkit for machine learning and data exploration," J. Mach. Learn. Res. **22**(39), 1–6 (2021).
[14] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. M. Medina-Mardones, A. Dassatti, and K. Hess (2021). "giotto-tda: A high-performance topological machine learning toolbox," giotto_tda, gitHub. https://github.com/giotto-ai/giotto-tda
[15] J. Lawrimore, B. Friedman, A. Doshi, and K. Bloom, "RotoStep: A chromosome dynamics simulator reveals mechanisms of loop extrusion," Cold Spring Harbor Symp. Quant. Biol. **82**, 101–109 (2017).
[16] J. Yan, T. J. Maresca, D. Skoko, C. D. Adams, B. Xiao, M. O. Christensen, R. Heald, and J. F. Marko, "Micromanipulation studies of chromatin fibers in *Xenopus* egg extracts reveal ATP-dependent chromatin assembly dynamics," Mol. Biol. Cell **18**, 464–474 (2007).
[17] K. Bloom and D. Kolbin, "Mechanisms of DNA mobilization and sequestration," Genes **13**, 352 (2022).