Computational topology of *Drosophila melanogaster* embryos: Emergent anatomy from the spatiotemporal distributions of hemocytes

Xinrui Hu [1], Nickolay Korabel [2], Tom H. Millard [3,*] and Thomas Andrew Waigh [1,†]

[1] *Biological Physics Group, School of Physics and Astronomy, The University of Manchester, Oxford Rd., Manchester M13 9PL, United Kingdom*

[2] *Department of Mathematics, The University of Manchester, Manchester M13 9PL, United Kingdom*

[3] *Division of Developmental Biology and Medicine, Faculty of Biology, Medicine and Health, Oxford Rd., Manchester M13 9PL, United Kingdom*

(Received 25 June 2025; accepted 5 September 2025; published 16 October 2025)

Hemocytes migrate within *Drosophila melanogaster* embryos during their morphogenesis. The spatial distribution of the hemocytes reflects the emergent anatomy of the embryos. It poses a significant challenge to statistical characterization using conventional methods, such as mean-square displacements of tracks and Ripley's L function. This challenge arises from the complex heterogeneity of hemocyte dynamics and the variability of the organism's anatomy due to inherent nonlinear stochastic processes. We apply topological data analysis (TDA) to three different genotypes of *Drosophila* embryos: wild type, $LanB1^{Df}$ (extracellular matrix mutants), and $SCAR^{\Delta 37}$ (motility mutants). Computationally efficient methods are developed to characterize the embryos' topologies. Based on the Wasserstein distances, the *LanB1* mutant is topologically distinct from the wild type and *SCAR* mutant, which are indistinguishable. Furthermore, an increase in the Betti-1 number (the number of holes in the hemocyte configurations) is observed across all samples with age, which is consistent with prior findings on cellular tiling behavior due to contact inhibition of locomotion. We also apply TDA methods to the hemocyte tracks within the embryos, allowing the wild type and *SCAR* mutant to be distinguished. Computational topology thus provides a robust method to quantify the emergent anatomy of embryos that is well suited to the complex heterogeneous datasets observed with hemocyte dynamics.

DOI: 10.1103/mc9z-kw1q

I. INTRODUCTION

Multicellular organisms grow from single cells via a process of morphogenesis. Morphogenesis is extremely complicated and is guided by a combination of both physical and biological cues [1]. A challenge in biological physics is thus to develop robust mathematical tools to quantify embryogenesis in highly variable cellular systems that are actively remodeled over time. We introduce techniques developed from computational topology that allow the anatomy of genetically modified *Drosophila* embryos to be quantified in a robust manner based on both the static spatial distribution of hemocytes within the embryo and the dynamic tracks of single hemocytes. This is shown to be superior to classical methods of spatial statistics, such as the Ripley L function [2], and provides complementary information to methods to describe motility, such as mean-square displacements (MSDs) and neural networks for anomalous transport [3–5].

During morphogenesis in flies, hemocytes move around the developing embryo, playing roles in immune responses, the synthesis of the extracellular matrix, and wound healing. The hemocytes migrate from their point of creation near the head of the embryo and spread out during the developmental stages of embryogenesis. Their migratory routes follow the anatomy of the developing organism, highlighting its open circulatory system [6–8].

In order to understand developmental biology, we investigated the distribution of hemocytes in the embryos of *Drosophila melanogaster*, a widely used model system [6] (Fig. 1). Hemocytes are the insect analogs of human white blood cells, which play a crucial role in tissue and organ formation during embryonic development [9]. *Drosophila melanogaster* serves as an ideal model system due to its rapid development, ease of maintenance, and robustness for experimental manipulation. To fulfill their complex role in the embryo, hemocytes must follow well-coordinated patterns of movement to ensure that they are evenly distributed and quickly reach the required locations, e.g., during wound healing to combat infections [10].

During the transport of hemocytes, the extracellular matrix (ECM) plays a crucial role [11]. The ECM is a dynamic biopolymer network composed of various structural components, including proteoglycans, glycoproteins, and fibrous proteins [12]. It provides essential support for cell shape, facilitates cell migration, provides a labyrinth through which the cells move, and contributes to wound healing [13]. Despite extensive experimental studies of the ECM, most research has focused on its functional properties rather than its

*Contact author: tom.millard@manchester.ac.uk

†Contact author: t.a.waigh@manchester.ac.uk

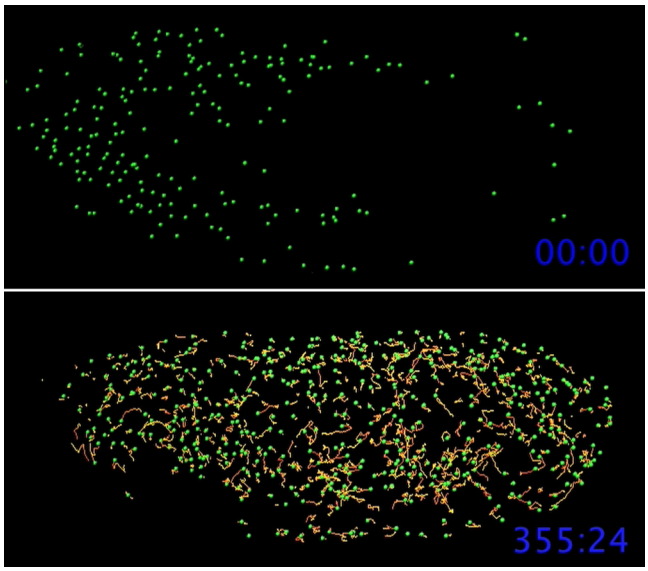

FIG. 1. Images of the segmented hemocytes at the start and the end of the experiment created using selective plane illumination microscopy. The times are shown at the bottom right of the figures (units of seconds). The green dots represent the hemocytes and the brown lines indicate the tracks of individual cells over the previous ten frames (180 s).

topological structure. In this study we aim to apply concepts from topological data analysis (TDA) to explore the topology of the ECM, a key component of the anatomy of *Drosophila*. Topological data analysis has been shown to outperform classical spatial statistical methods such as Ripley's *L*-function [2,14]. It provides complementary information to mean-square displacement and offers a wealth of topological information about the underlying anatomy.

The mathematical discipline of topology involves the study of spatial relationships. In recent years, there have been increasing applications of topological concepts to the analysis of large and complex datasets, a methodology known as topological data analysis. Topological data analysis has gained significant traction in biological research, for various applications, such as cluster detection [15], the analysis of survival times among different *E. coli* strains [16], characterization of cell-cell adhesion [17], and breast cancer analysis [18].

Topological data analysis in biological research [19] is usually used to investigate the geometric or topological structure of an assumed underlying manifold in the data, such as horizontal gene transfer in evolutionary trees. Topological data analysis has identified elliptical structures in *Drosophila* epithelial cells [20]. However, our work differs from previous approaches in that we apply TDA without assuming that such a manifold exists. This shift is significant because the absence of a manifold structure can cause strong fluctuations in persistence diagrams, raising concerns about their interpretability and stability. Nevertheless, we argue that these challenges can be effectively overcome using statistical techniques and convergence theorems.

Extensive research has been conducted on collective cell movement [21]. However, conventional methods of analysis are inefficient in quantifying the spatial distributions of hemocytes. Traditional statistical approaches typically analyze MSD of tracks across the entire ensemble [22], allowing for the classification of cell migration as subdiffusive or superdiffusive. Additionally, velocity autocorrelation functions have been employed to detect directional diffusion in embryos [23], while first-passage probability has been widely used to understand reaction rates [24]. More recently, neural networks have shown great promise for the analysis of heterogeneous anomalous transport common in molecular and cellular biology [4,5].

Despite these alternate methods, modeling hemocyte migration presents unique challenges compared to other forms of biological motility. While cell-cell interactions, such as contact inhibition of locomotion play a role, hemocyte migration is primarily regulated by cell-intrinsic factors rather than emergent collective dynamics. As a result, treating the system as an ensemble in nonequilibrium thermodynamics may be insufficient, as the governing rules can change dramatically during morphogenesis.

Topological data analysis has not been extensively applied to the analysis of particle tracks in biology [25,26]. The techniques we develop could thus be applied to a wide range of particle tracking experiments in biophysics, soft matter, ecology, and beyond.

II. METHODS

In our experiments, we used selective plane illumination microscopy (SPIM) to track hemocyte movement in *Drosophila* embryos in three dimensions [27]. To evaluate the effectiveness of our method to analyze the ECM, we examined three different genotypes of embryo: wild type, $SCAR^{\Delta 37}$, and $LanB1^{Df}$. The $LanB1^{Df}$ mutants lack the *LanB*1 gene, impairing production of laminin, a key component of the ECM [28], i.e., the underlying labyrinths the hemocytes move through. In contrast, $SCAR^{\Delta 37}$ mutants lack the *SCAR* gene, which promotes actin filament assembly at the cell front, and therefore its absence impairs their active transport [29]. By comparing these embryos, we aimed to determine whether the changes in hemocyte behavior result from intrinsic cellular movement (*SCAR*) or external environmental factors (*LanB1*).

Histones in all of the hemocytes were fluorescently tagged to a red fluorescent protein, and the SPIM experiments allowed the positions of the nuclei in every hemocyte cell in half of an embryo to be tracked. Hemocytes were examined during stages 13 and 14 of embryogenesis. Selective plane illumination microscopy movies consisted of 400 frames, with each frame taking 18 s to collect.

We performed the TDA analysis using python, employing the persim package from the scikit-learn open-source machine learning library to compute persistence diagrams [30]. We analyzed the differences between the mutant embryos using the Wasserstein distance metric, rather than the bottleneck distance metric, which is commonly used to measure similarity between persistence diagrams. This choice is motivated by computational efficiency, as bottleneck distance calculations require not only pairing points in persistence diagrams but also computing their distances to the diagonal, making the process at least four times slower [31]. Additionally, the Hungarian algorithm used for Wasserstein distance computation is well optimized for our application.

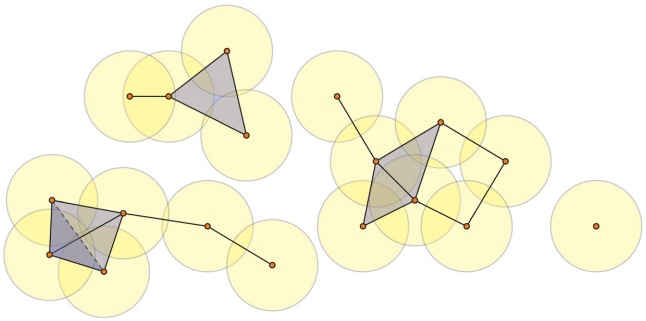

FIG. 2. Schematic diagram for the construction of a Vietoris-Rips complex. The red dots are the point cloud and the yellow region is the ϵ ball of each point (ϵ is the ball radius). The homology of the connected components is the persistent homology we are focusing on, i.e., the three-hole regions shaded in gray.

Nevertheless, computing the distance between all persistent diagram pairs remains computationally intensive. To mitigate this, we used the joblib package to parallelize the computation across all eight cores of an Intel i7 processor. Although access to high-performance computing resources could accelerate the process further, this is not strictly necessary for this study. Our python code is freely available on GitHub [32].

III. THEORY

The primary tool in TDA is homology, specifically persistent homology [33]. Homology quantifies the number of n-dimensional holes in a given topological space, allowing for the classification of topologically distinct spaces using algebraic invariants, such as the Betti numbers and the Euler characteristics.

The basic building blocks of homology are simplices, which are the simplest geometric objects in each dimension: points (0-simplices), line segments (1-simplices), triangles (2-simplices), tetrahedra (3-simplices), and so on. By gluing simplices together while ensuring that shared boundaries remain subsimplices, we construct simplicial complexes. In general, a Vietoris-Rips (VR) complex can be constructed around any given point cloud; in our case it was the spatial distribution of the hemocytes. To transform a point cloud into a simplicial complex, we connect points within a given proximity parameter ϵ, which progressively increases from 1 to ∞. The resulting VR complex is illustrated in Fig. 2.

The number of topological features (holes) in the VR complex is represented by persistent diagrams, which track their birth (the value of ϵ at which a hole appears) and death (the value of ϵ at which the hole disappears). These diagrams offer a concise summary of the topological structure of the data. The same information can be presented in the form of barcodes, which are defined as follows.

Notation (from [34]). Let $\mathcal{B}$ denote the set of persistence barcodes. Given a persistence barcode $A \in \mathcal{B}$, its n_a intervals will be denoted by $[x_i^a, y_i^a)$ for $1 \leqslant i \leqslant n_a$. In addition, the length of $[x_i^a, y_i^a)$ will be denoted by ℓ_i^a, that is, $\ell_i^a = y_i^a - x_i^a$. Finally, L_a will denote the sum $\sum_{i=1}^{n_a} \ell_i^a$. Moreover, given two persistence barcodes A and B, the $\max\{n_a, n_b\}$ is denoted by $n_{\max}$ and the $\max\{L_a, L_b\}$ by $L_{\max}$.

The primary statistical measure applied to the *Drosophila* embryo data is the Wasserstein distance [35]. In an active, noisy biological environment, methods must be robust against random fluctuations introduced by the stochastic motility of the hemocytes, and we believe this is the case for the Wasserstein distances.

For the Wasserstein distance, the following is the main theorem.

Theorem 1 ([35]). Consider two finite metric spaces (X, d_X) and (Y, d_Y). Let A and B be the two persistence barcodes obtained from their Vietoris-Rips complexes $\mathrm{Rips}(X,t)\,|_{t\in\mathbb{R}}$ and $\mathrm{Rips}(Y,t)\,|_{t\in\mathbb{R}}$, respectively. Then

$$d_\infty(A, B) \leqslant d_{\mathrm{GH}}(X, Y),$$

where d_{GH} is the Gromov-Hausdorff distance, which is calculated using

$$d_{\mathrm{GH}}(A, B) = \inf_{f_A, f_B} d_{\mathrm{H}}(f_A \to X(A), f_B \to X(B)), \qquad (1)$$

i.e., the minimum Hausdorff distance of two metric spaces among all embeddings. The above result suggests that when the physical distance between two point clouds is small, they are also topologically close in the Wasserstein sense. When representing a persistence diagram in barcode form, each bar corresponds to a topological feature, where the starting point of a bar represents its birth time and the ending point signifies its death time. By converting the persistence diagram into a barcode representation, we obtain a structured framework for analyzing topological persistence.

IV. RESULTS

A. Ripley's L function is insufficient to distinguish all different mutants

The topology of data is directly correlated with the spatial distribution of the point cloud [36]. Therefore, before implementing TDA methods, we performed a preliminary analysis using Ripley's L function to identify potential clustering patterns in embryos with mutant hemocytes to assess whether this approach could distinguish between different mutations.

The Ripley L function was computed using only the x and y coordinates, and the z coordinate was excluded. The z axis spanned only approximately 50 μm compared to approximately 400 μm in the x-y plane due to the functioning of the selective-plane illumination microscope (this is common to the majority of optical microscopes and SPIM provides state-of-the-art imaging with embryos), leading to a negligible contribution of z in the pairwise distances. So excluding the z coordinate does not affect our conclusions and avoids errors caused by its lower imaging resolution in SPIM. Hemocyte distributions were examined at $t = 120$ min from the beginning of development, a time point that generally marks the full maturation of the embryo. The Ripley L function was calculated by iterating over the distance matrix and summing all pairwise distances below the distance parameter r. To ensure statistical reliability, we repeated this test across three independent datasets for each mutant group and incorporated error bars in our results.

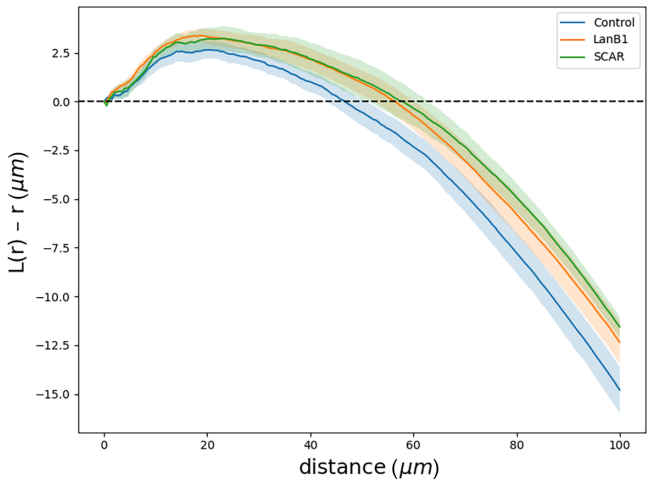

FIG. 3. Comparison of Ripley's L function $L(r) - r$ of wild type and *LanB1* and *SCAR* mutants as a function of the distance between the mutants r at 120 min (400 frames) after the start of the experiment. The shaded region shows the standard deviation for all three repeats.

The findings (Fig. 3) indicate that all Ripley's L functions are initially super-Poissonian, but are sub-Poissonian after $r \simeq 60$ μm. At shorter times, Ripley's L functions are indistinguishable among mutant groups; the differences are not statistically significant, with a p-value threshold of 0.05.

This comparison suggests that hemocyte distributions exhibit a degree of spatial clustering, which is influenced by genetic mutations. The finding can also be observed using Ripley's K function without normalization (see Fig. S2 in the Supplemental Material [37]). Specifically, while the *SCAR* and *LanB1* mutants can be distinguished from the wild-type embryos based on their clustering patterns (with p values of 0.89 and 0.24, respectively), the two mutants cannot be reliably distinguished from each other (with a p value of 0.028). The drop of Ripley's L function at large distances (greater than 60 μm) could be interpreted as due to boundary constraints. Additionally, video analysis (see Fig. S3 in the Supplemental Material [37]) indicates that while hemocytes generally maintain an even spatial distribution, they tend to cluster around nerve cord and head mesoderm regions.

B. Betti-1 number of circulating hemocytes in embryos: Static analysis

The Betti-1 number represents the number of one-dimensional (1D) holes in the VR complex of a point cloud and serves as an essential measure of the complexity of the underlying spatial distribution. To investigate the impact of morphogenesis on the ECM, it is instructive to determine whether the topology of the hemocyte distribution evolves over time. Accordingly, we performed topological data analysis on datasets corresponding to our *LanB1*, *SCAR*, and control embryos. Each experiment was repeated three times, and we can plot the averaged Betti-1 curves, justified by the additive property of Betti numbers. As before, due to heterogeneity in the z direction, two-dimensional holes are considered negligible. Additionally, we calculated the number of significant holes (holes with persistence greater than

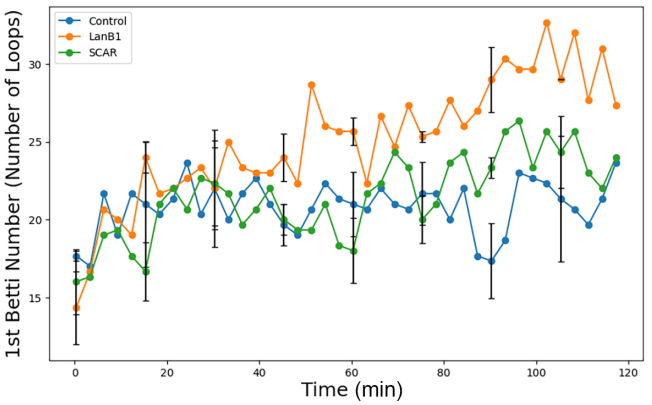

FIG. 4. Betti-1 number as a function of time for three varieties of *Drosophila melanogaster* embryo. The blue, yellow, and green lines represent the Betti-1 number of the control (wild type), *LanB1*, and *SCAR* mutants. The error bars represent the standard deviation for all three repeats and are drawn for every 50 frames.

5 μm) to identify potential differences. The number 5 μm is chosen to exclude 95% of holes caused by noise. Changing the threshold shows that the trend does not depend on the exact choice of the threshold value (see Fig. S4 in the Supplemental Material [37]).

The resulting data are shown in Fig. 4. During embryo development, due to hemocyte circulation in the embryos (the hemocytes explore a larger volume of the embryo as time progresses), the Betti-1 number increases across all three mutant groups, with the most notable increase observed in the *LanB1* mutant, reaching a value of approximately 30. This increase reflects a significant rise in complexity within the hemocyte distribution and may provide supporting evidence for cellular tiling behavior previously described in the literature [38]. A similar trend is also observed when focusing only on holes with sizes above a certain threshold.

This topological behavior can be explained by the circulation phase of hemocytes in *Drosophila* embryos. Initially, hemocytes are produced in the head mesoderm of the embryo [10]. As these cells circulate throughout the embryo, they rapidly occupy previously unpopulated regions, leading to the fragmentation of larger voids into smaller ones. Once hemocytes are evenly distributed across the entire embryo, they undergo trapped motion, restricted to coraled motions within their respective spatial domains.

Furthermore, due to the interaction between hemocytes and the ECM, the ECM plays a substantial role in increasing the topological complexity of the system. The structural modifications induced by ECM interactions may contribute to the observed changes in the hemocyte distribution, leading to the emergence of more sophisticated topological features.

C. Detection of persistent holes in the tracks of single hemocytes: Dynamic analysis

To detect cyclical or looping behaviors of hemocytes and uncover hidden barriers within the embryo, we apply TDA methods to tracks of single hemocytes in the embryo. A representative VR complex reconstruction is shown in Fig. 5.

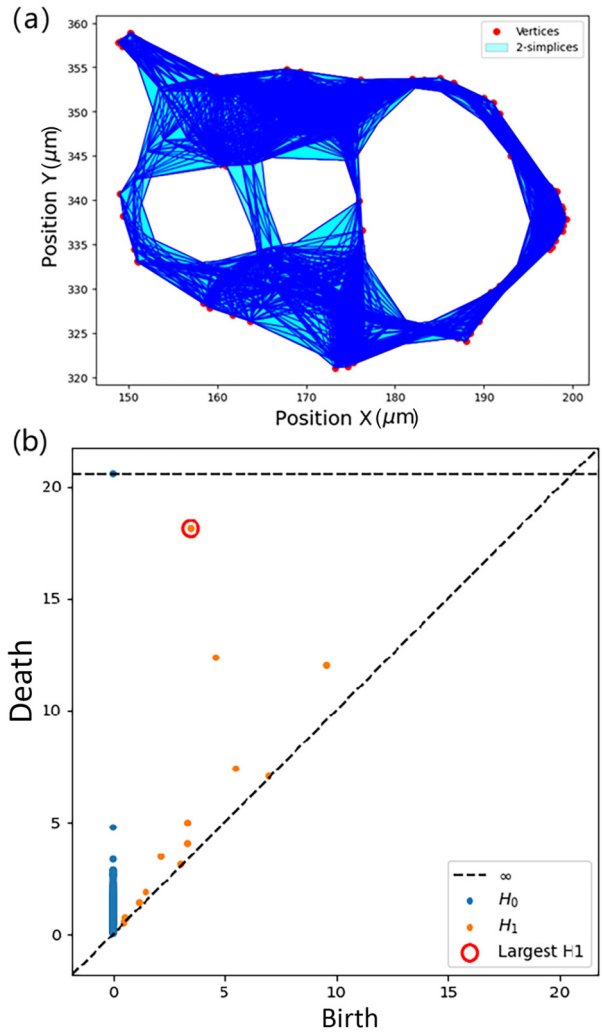

FIG. 5. (a) The VR complex of the track of a single hemocyte. Red points are the hemocytes and the blue region is the reconstructed complex. Three consecutive holes are identified in this case. (b) Corresponding persistent diagram. Blue and yellow points represent 0D and 1D holes, respectively.

From visual inspection, we can identify the presence of holes, although their number is not stable and may vary with persistence. Due to this variability, only the most prominent (i.e., largest) hole is considered in our analysis. This is justified by the observation that each hemocyte loops around the embryo at most once. Therefore, our method effectively captures the essential topological feature, i.e., the existence of a significant loop, without being affected by minor topological noise. The length of each hemocyte track used for the analysis is similar, hence not affecting the Betti-1 number distribution (see Table S1 in the Supplemental Material [37]).

D. Statistical measures of holes in a single track

To identify the number of tracks with significant loops, we plot the distribution of the largest holes for the wild-type embryo. The result is as follows.

First, we calculate the number of trajectories containing at least one topological hole. Of the 2017 trajectories analyzed, 1674 (approximately 83%) exhibit at least one hole. To distinguish meaningful topological features from noise, we identify nontrivial holes, which are defined as those with persistence greater than 1; only 15% of the detected holes fall into this category, indicating that the majority are likely to be noise within the trajectory data.

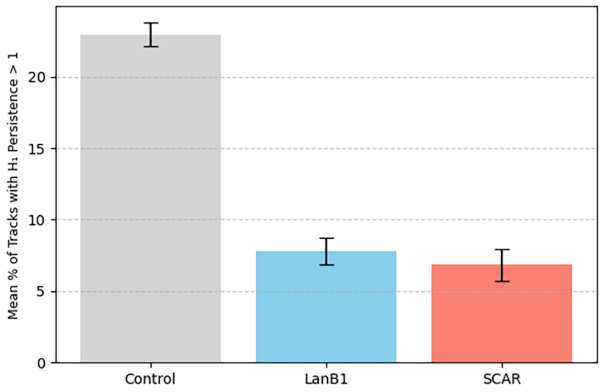

FIG. 6. Percentage of the largest holes within the tracks of single cells of different mutants. The horizontal axis is the names of the datasets and the vertical axis is the percentage of tracks that contain at least one significant hole.

We apply the same analysis to different genetic variants (Fig. 6). Wild-type datasets consistently show a higher proportion of trajectories containing at least one hole than the mutant strains, e.g., 15% in wild type versus approximately 5% in *LanB1* and *SCAR* mutants. This discrepancy can be attributed to the motility-restricted nature of these mutants. Reduced motility decreases the likelihood of cells looping around the embryo, thereby reducing the probability of topological holes forming in the corresponding trajectories.

Next we examine the distribution of hole persistence by plotting the survival function (Fig. 7), defined as the probability that the largest hole in a trajectory persists beyond a given time, averaged across all hemocyte tracks. Specifically, we plot the persistence values against the logarithm of the survival function and fit the resulting curve using a double-exponential function. The survival function shows exponential decay overall, with two distinct phases: rapid decay corresponding to short-lived holes (interpreted as noise) and slower decay for longer-lived holes (potentially meaningful structures). This pattern indicates that both noise-induced and biologically relevant holes follow exponential persistence distributions, although with different decay rates.

E. Persistence diagram

The persistence diagram is a part of the standard protocol in topological data analysis for visualizing topological features and their significance. Each point on the persistence diagram represents a 1D hole, where the x axis denotes the birth time, i.e., the proximity parameter ϵ at which the hole appears, and the y axis represents the death time, i.e., the ϵ value at which the hole disappears. Consequently, no holes exist below the diagonal, as they would correspond to features that vanish instantaneously or disappear before they are born.

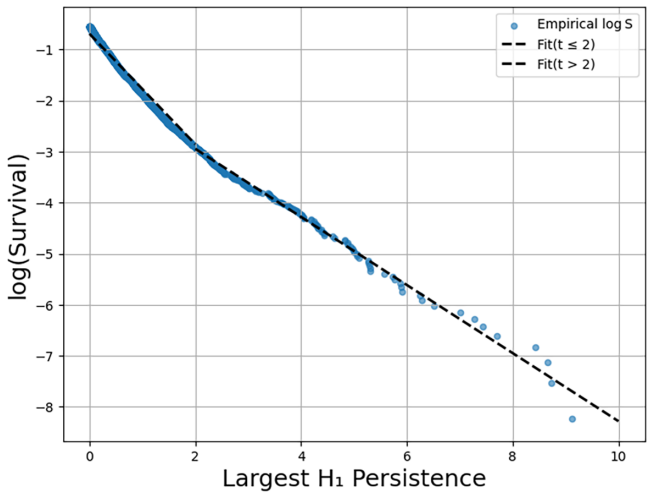

FIG. 7. Fit of the survival function of the largest holes over all trajectories. The x axis is the largest 1D hole in each trajectory and the y axis is the logarithm of the survival rate. The black lines are the fits for Betti number $0 < t < 2$ and $2 < t < 10$ regions, respectively. The fit lines are $y = -1.08t - 0.70$ $(0 < t < 2)$ and $y = -0.67t - 1.61$ $(2 \leqslant t < 10)$, respectively, i.e., two exponential functions on a linear-linear plot.

We generated a sample persistence diagram using the risper package, and the result is shown in Fig. 8. Zero-dimensional features (connected components), 1D features (loops), and 2D features (voids) are displayed. The hemocyte distribution

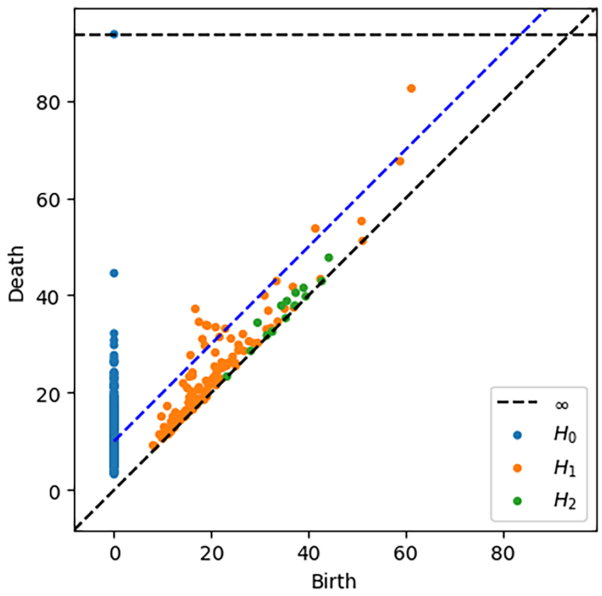

FIG. 8. Persistence diagram of a wild-type embryo. Blue dots indicate connected components in the VR complex, while orange dots represent 1D loops and green dots represent 2D holes. The diagonal black dashed line corresponds to trivial holes, whereas the horizontal black dashed line represents the largest connected component. Additionally, the blue dashed line denotes the error bar, approximated by the equation $y = x + 10.7$, which serves as a threshold for distinguishing significant topological features from noise.

diagram reveals that most holes from the static embryos have a short lifespan, defined as the difference between their death and birth times. To ensure that we are analyzing only significant topological features, we focused exclusively on persistent holes in our diagrams.

Error margin analysis in TDA is particularly challenging because topological similarity cannot be directly quantified using Euclidean distance. Instead, we employ the Wasserstein distance to measure structural differences between persistence diagrams. To estimate uncertainty, we apply a bootstrap method, randomly selecting 120 subsamples from the original dataset and generating persistence subdiagrams for each subset. We then compute the Wasserstein distance matrix between subdiagrams, as detailed in the Supplemental Material, Fig. S5 [37].

The results indicate that all holes with a size less than 10.7 μm are likely to be noise rather than physically meaningful topological features. The sample is "flattened" due to the inhomogeneity of the z direction, meaning that all 2D voids are insignificant.

F. Comparison of embryo topologies via Wasserstein distances

Instead of the conventional application of topological analysis, which primarily focuses on detecting holes from random samples on a manifold, we extend our analysis to 3D random distributions, specifically examining the distribution of persistent diagrams and barcode spaces.

From the clustering figures (Fig. 9), we observe that the control and *SCAR* groups are indistinguishable using the TDA method, with a cross-Wasserstein distance in the range of approximately 1000–2500 μm. In contrast, *LanB1* mutants exhibit significantly higher similarity among themselves, with a Wasserstein distance of less than 500 μm, allowing them to be clearly separated from other mutants.

LanB1 mutants are not easily distinguishable using traditional statistical methods. For instance, while the MSD and velocity autocorrelation function exhibit minor differences, these variations are generally not statistically significant ($p > 0.05$) [3]. However, in our experiments, when averaging across all time steps, the *LanB1* mutants display a substantial divergence from the other types of embryos.

V. DISCUSSION

Topological data analysis successfully distinguished between all three genotypes of *Drosophila* embryo. *LanB1* mutant embryos could be distinguished based on the topology of the static arrangements of the hemocytes (Fig. 4). *SCAR* mutant embryos could subsequently be distinguished based on the topology of the individual hemocyte tracks that were ensemble averaged (Fig. 6). The topological analysis provides a rich source of information, in addition to previous techniques, such as the classical tools for spatial statistics (e.g., Ripley's L function) and anomalous transport (e.g., mean-square displacements). Topological data analysis is well suited to the analysis of embryo anatomies that can be highly variable, but in which topological features are rigorously enforced for viable organisms, e.g., the development of the neural tube.

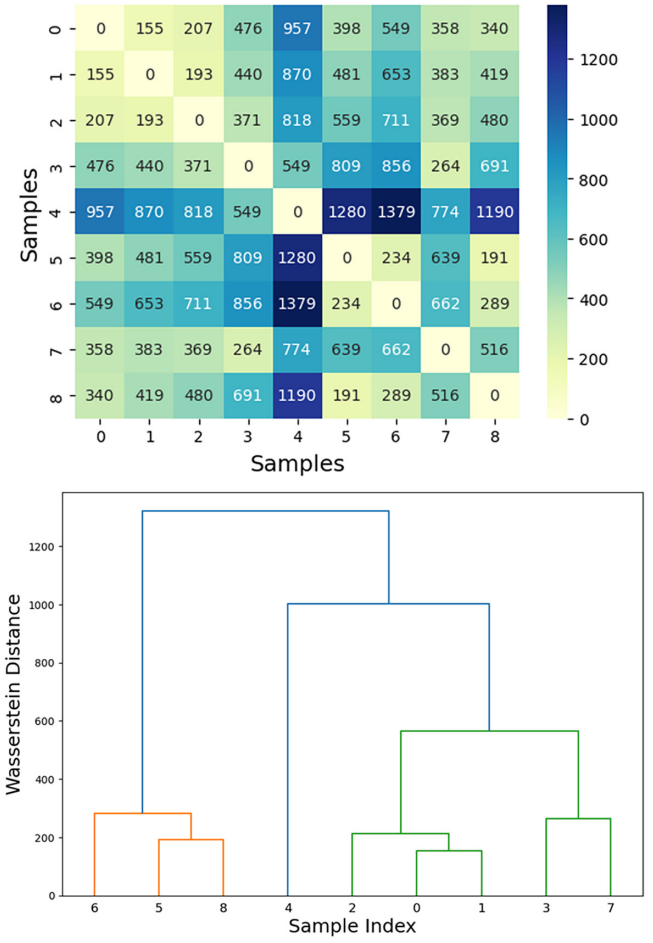

FIG. 9. Shown on top is the Wasserstein distance between different samples of hemocyte embryos. The color represents the relative distance, and the samples are given by 0-2 LanB1, 3-4 control, and 5-8 SCAR, respectively. Shown on the bottom is a hierarchical clustering diagram of all samples.

Topological data analysis could be used much more widely for the analysis of morphogenesis and the emergent anatomy of organisms. Furthermore, TDA with particle tracks has not been extensively studied in the biophysics literature, but could be very useful in future studies on the gait of motile organisms (e.g., how bacteria swim [39]), active motility (e.g., the cycling behavior of myosin head groups on actin filaments), flocking transitions in collective motility (e.g., murmurations of starlings), and the corralled motions of cargoes inside cells (e.g., vesicle motions around the endoplasmic reticulum). Both static and dynamic TDA could be used with machine learning to automate classification of active phases, e.g., stages of morphogenesis or flocking phase transitions of active particles.

VI. CONCLUSION

Computational topology provides a useful tool to describe *Drosophila* anatomy, e.g., to quantify key biological stages during development or to separate different species or quantify evolutionary relationships. While Ripley's L function provides the tool to distinguish wild-type embryos from mutants, it cannot distinguish between mutants. Topological data analysis from static hemocyte distributions of the whole embryo allows the ECM mutants (*LanB1*) to be distinguished from the control and the motility mutant (*SCAR*). Topological data analysis from dynamic hemocyte tracks allowed the motility mutant (*SCAR*) to be distinguished from the other two genotypes. Thus, a combination of static and dynamic TDA provides unique signatures for all three embryo types. Topological data analysis was applied to hemocyte tracks in *Drosophila melanogaster* embryos. It exhibited superior efficacy for embryo classification compared to conventional statistical methods, such as the mean-square displacement and the Ripley L function. The TDA method enables the classification of different mutants, thereby demonstrating a level of precision that surpasses that of previous methodologies [3]. The *LanB1* group is identified as the most topologically complex among all the mutants, owing to the variation in the ECM and thus the labyrinths in which the hemocyte cells circulate.

ACKNOWLEDGMENTS

N.K. was supported by EPSRC (Grant No. EP/V008641/1) and T.M. acknowledges financial support from the Wellcome Trust Grant No. 201958/Z/16/Z.

DATA AVAILABILITY

The data that support the findings of this article are openly available [32].

[1] G. Forgacs and S. A. Newman, *Biological Physics of the Developing Embryo* (Cambridge University Press, Cambridge, 2005).
[2] S. H. Holmes and W. Huber, *Modern Statistics for Modern Biology* (Cambridge University Press, Cambridge, 2018).
[3] N. Korabel, G. D. Clemente, D. Han, F. Feldman, T. H. Millard, and T. A. Waigh, Commun. Phys. **5**, 269 (2022).
[4] D. Han, N. Korabel, R. Chen, M. Johnston, A. Gavrilova, V. J. Allan, S. Fedotov, and T. A. Waigh, eLife **9**, e52224 (2020).
[5] T. A. Waigh and N. Korabel, Rep. Prog. Phys. **86**, 126601 (2023).
[6] M. H. Baron, *Developmental Hematopoiesis: Methods and Protocols* (Humana, Totowa, 2005).
[7] M. Lavine and M. Strand, Insect Biochem. Mol. Biol. **32**, 1295 (2002).
[8] R. Horwitz and D. Webb, Curr. Biol. **13**, R756 (2003).
[9] A. Holz, B. Bossinger, T. Strasser, W. Janning, and R. Klapper, Development **130**, 4955 (2003).
[10] W. Wood and A. Jacinto, Nat. Rev. Mol. Cell Biol. **8**, 542 (2007).
[11] C. J. Hughes and J. R. Jacobs, Vet. Sci. **4**, 24 (2017).
[12] L. I. Fessler, R. E. Nelson, and J. H. Fessler, Drosophila extracellular matrix, in *Methods in Enzymology*, edited by E.

Ruoslahti and E. Engvall (Elsevier, Amsterdam, 1994), Vol. 245, pp. 271–294.
[13] B. Stramer, W. Wood, M. J. Galko, M. J. Redd, A. Jacinto, S. M. Parkhurst, and P. Martin, J. Cell Biol. **168**, 567 (2005).
[14] A. Hohl, M. Zheng, W. Tang, E. Delmelle, and I. Casas, Spatiotemporal point pattern analysis using Ripley's K function, in *Geospatial Data Science Techniques and Applications* (CRC Press, Boca Raton, FL, 2017), pp. 155–176.
[15] X. Wang, F. Sohel, M. Bennamoun, Y. Guo, and H. Lei, Pattern Recog. **71**, 414 (2017).
[16] A. M. Ibekwe, J. Ma, D. E. Crowley, C.-H. Yang, A. M. Johnson, T. C. Petrossian, and P. Y. Lum, Front. Cell. Infect. Microbiol. **4**, 122 (2014).
[17] D. Bhaskar, W. Y. Zhang, A. Volkening, B. Sandstede, and I. Y. Wong, npj Syst. Biol. Appl. **9**, 43 (2023).
[18] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park, and J. Arsuaga, Topol. Appl. **157**, 157 (2010).
[19] E. J. Amézquita, M. Y. Quigley, T. Ophelders, E. Munch, and D. H. Chitwood, Dev. Dyn. **249**, 816 (2020).
[20] D. Bhaskar, W. Y. Zhang, and I. Y. Wong, Soft Matter **17**, 4653 (2021).
[21] N. Sepúlveda, L. Petitjean, O. Cochet, E. Grasland-Mongrain, P. Silberzan, and V. Hakim, PLoS Comput. Biol. **9**, e1002944 (2013).
[22] N. Gal, D. Lechtman-Goldstein, and D. Weihs, Rheol. Acta **52**, 425 (2013).
[23] A. Fuliński, J. Phys. A: Math. Theor. **50**, 054002 (2017).
[24] D. A. Kenwright, A. W. Harrison, T. A. Waigh, P. G. Woodman, and V. J. Allan, Phys. Rev. E **86**, 031910 (2012).
[25] L. L. Bonilla, A. Carpio, and C. Trenado, PLoS Comput. Biol. **16**, e1008407 (2020).
[26] K. C. Nguyen, C. D. Jameson, S. A. Baldwin, J. T. Nardini, R. C. Smith, J. M. Haugh, and K. B. Flores, Math. Biosci. **370**, 109158 (2024).
[27] J. Huisken and D. Y. R. Stainier, Development **136**, 1963 (2009).
[28] B. J. Sáanchez-Sánchez, J. M. Urbano, K. Comber, A. Dragu, W. Wood, B. Stramer, and M. D. Martín-Bermudo, Cell Rep. **21**, 1461 (2017).
[29] I. Evans, P. Ghai, V. Urbančič, K. L. Tan, and W. Wood, Cell Death Diff. **20**, 709 (2013).
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011).
[31] M. Catanzaro, C. Tralie, and N. Saul, Persim—A scikit-tda project (Zenodo, Geneva, 2024), http://doi.org/10.5281/zenodo.2533369.
[32] X. Hu (GitHub, San Francisco, 2025), https://github.com/Ramarujun/Topological-Data-Analysis-of-Hemocytes.
[33] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, EPJ Data Sci. **6**, 17 (2017).
[34] R. Ghrist, Bull. Am. Math. Soc. **45**, 61 (2008).
[35] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko, Found. Comput. Math. **10**, 127 (2010).
[36] J. R. Davis, C. Y. Huang, J. Zanet, S. Harrison, E. Rosten, S. Cox, D. Y. Soong, G. A. Dunn, and B. M. Stramer, Development **139**, 4555 (2012).
[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/mc9z-kw1q for the distribution of track lengths for control, LanB1, and SCAR samples; analysis of Ripley's K functions to complement the Ripley L results in the main text; snapshots of hemocyte migration in *Drosophila* embryos; the effect of different persistence thresholds on Betti-1 calculations; and calculation of error bars for persistence diagrams.
[38] J. R. Davis, A. Luchici, F. Mosis, J. Thackery, J. A. Salazar, Y. Mao, G. A. Dunn, T. Betz, M. Miodownik, and B. M. Stramer, Cell **161**, 361 (2015).
[39] T. A. Waigh, *The Physics of Bacteria* (Cambridge University Press, Cambridge, 2024).