scientific reports

Check for updates

OPEN

Topological data analysis of pattern formation of human induced pluripotent stem cell colonies

Iryna Hartsock[1,8], Eunbi Park[2,8], Jack Toppen[2,3], Peter Bubenik[4], Elena S. Dimitrova[5], Melissa L. Kemp[6] & Daniel A. Cruz[7]✉

Understanding the multicellular organization of stem cells is vital for determining the mechanisms that coordinate cell fate decision-making during differentiation; these mechanisms range from neighbor-to-neighbor communication to tissue-level biochemical gradients. Current methods for quantifying multicellular patterning tend to capture the spatial properties of cell colonies at a fixed scale and typically rely on human annotation. We present a computational pipeline that utilizes topological data analysis to generate quantitative, multiscale descriptors which capture the shape of data extracted from 2D multichannel microscopy images. By applying our pipeline to certain stem cell colonies, we detected subtle differences in patterning that reflect distinct spatial organization associated with loss of pluripotency. These results yield insight into putative directed cellular organization and morphogen-mediated, neighbor-to-neighbor signaling. Because of its broad applicability to immunofluorescence microscopy images, our pipeline is well-positioned to serve as a general-purpose tool for the quantitative study of multicellular pattern formation.

Keywords Pluripotent stem cells, Microscopy images, Cell pattern formation, Topological data analysis, Persistence landscapes

Fluorescence microscopy is one of the most important and widely used tools for studying cell physiology[1,2], from intracellular interactions to multicellular organization and beyond. The processes affecting multicellular pattern formation are typically studied with multichannel microscopy, with the detection of spatial features and dynamics often relying upon the visual inspection of several images. Quantitative methods have been developed for recognizing phenotypes at designated spatial scales, and improvements in individual cell identification and segmentation within images of tissues have enabled the ability to characterize patterning as a collective property of multiple cells. For example, deep learning methods have been used for classification to compare and track the spatiotemporal characteristics of each cell[3,4] for the definition of tissue features[5].

There are several statistical approaches which have been traditionally applied to the context of analyzing spatial, multicellular data[6–12]. However, these approaches are limited in that they tend to only capture structural features at a fixed scale. By not considering how spatial features evolve across scales, such methods may fail to incorporate the effects that different degrees of noise could have on a multicellular network[13]. A different approach to quantifying multicellular organization uses temporal-spatial signal logic to comparatively score the complexity of an image to a target pattern[14]. While such a data-driven tool is effective in computing the difference between patterns, its outputs (e.g. latent variables, similarity scores) do not necessarily provide biologically interpretable information on intercellular interactions.

Graph-based methods have also been applied to derive quantitative descriptors of a multicellular construct such as an aggregate or colony[15–18]. With this strategy, each pattern is considered as an ensemble of network connections between neighboring cells. Neighborhood network features (e.g. path lengths or a number of like-cell clusters) are extracted from segmented, digitized microscopy images for statistical comparison across conditions[18] or are subsequently evaluated with dimension reduction techniques to classify or compare spatiotemporal patterns with other images[16,17]. However, these tools require some amount of human annotation

[1]Department of Machine Learning, H. Lee Moffitt Cancer Center & Research Institute, Tampa 33612, US. [2]School of Biological Sciences, Georgia Institute of Technology, Atlanta 30332, US. [3]Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge 02139, US. [4]Department of Mathematics, University of Florida, Gainesville 32611, US. [5]Mathematics Department, California Polytechnic State University, San Luis Obispo 93407, US. [6]The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology & Emory University, Atlanta 30332, US. [7]Department of Medicine, University of Florida, Gainesville 32611, US. [8]Iryna Hartsock and Eunbi Park have contributed equally to this work. ✉email: daniel.cruz@medicine.ufl.edu

or a selection of network metrics *a priori* to learn the data or define the reduced dimension metrics. Similar approaches which consider pairwise connections and distances between subcellular structures may require less supervision; these techniques use distance distributions to measure the correlation between each cell in a microscopy image[19–21]. Though they quantify some morphological details, these methods share the same limitation as several of the ones mentioned previously in that they can only capture structural features at a specific scale.

As an alternative to the methods mentioned above, we propose to use topological data analysis (TDA) to examine pattern formation. TDA is a fast-growing field which provides methods for summarizing the shape of complex data[22] and has found useful applications in several fields, including cosmology[23], material science[24], neuroscience[25], anomaly detection[26], and *C. elegans* behavior[27]. Persistence homology, the central tool of TDA, tracks the appearance and disappearance of structural features (e.g. bounded empty regions) of data across different scales[28]. A persistence diagram is the output of persistence homology, and it is stable under various data perturbations[29]. One can map persistence diagrams to persistence landscapes[30] to bridge TDA with other data analysis techniques. Persistence landscapes are shape descriptors that live in a vector space and can be used as inputs to machine learning algorithms. Persistence landscapes have unique averages and satisfy the Strong Law of Large Numbers and the Central Limit Theorem, which makes them suitable for statistical inference.

Previous work established the feasibility of studying emergent spatial properties in developmental biology by combining TDA with a clustering method to study zebrafish patterns across agent-based model simulations[31]. Since then, approaches employing TDA have been applied to profile spatial configurations of epithelial cells[32] and to classify differences in bone microstructure[33]. Landscapes derived from the multiparameter persistence homology have been utilized to identify spatial patterns of both imaged and simulated immune cells[34], highlighting the translatable power of TDA-based approaches. The four instances above show how TDA can give insight into understanding multicellular organization. Recent works on collective cell motion[35,36] and on efficient and automated multicellular pattern identification[37] hold promise for broad applicability as well. In Edwards *et al* 2021[38], the authors incorporated persistence landscapes into a microscopy image analysis pipeline (TDAExplore) for detecting changes in the architecture of actin cytoskeleton. While this tool is accessible and broadly applicable, its output is produced on a per-channel basis; thus, it would need to be modified in order to compute topological features for a cell type which is identified using multiple channels. Newly published work on glial scar formation shows the advantages of pairing TDA with more traditional methods for spatial data analysis[39]; however, the pipeline associated with this work is tailored to its biological context and relies on the assumption that each cell type is associated with a single biomarker.

Here, we develop a computational pipeline for quantifying multicellular patterns observed in 2D multichannel microscopy images using TDA. Three sequential modules form a complete pipeline which automates cell segmentation, cell type identification, and the generation of multiscale, topological descriptors (persistence landscapes) for a given microscopy image set. The identification module allows a cell type to be determined using signals from multiple channels. The TDA module outputs persistence landscapes of an image set which allows them to be combined with statistical and machine learning tools; the module also generates cycle representatives of detected structural features for each input image. The end result is a modular, general-purpose pipeline aimed at broadening the access to TDA in the context of microscopy image analysis. The TDA module may be used instead of or in addition to more traditional, statistical methods for analyzing spatial data to gain further insight into multicellular organization.

To explore the usefulness of the above multiscale descriptors, we apply our pipeline to study stem cell colonies actively undergoing differentiation. In particular, we study loss of pluripotency in the context of human induced pluripotent stem cells (hiPSCs), which are reprogrammed from somatic cells and have the capacity to differentiate into all primary germ layers represented in embryos[40]. Because of their comparable characteristics to human embryonic stem cells[41], hiPSC cultures have become powerful, patient-specific *in vitro* test beds for investigating the early stages of human embryonic development[42]. During differentiation, a population of hiPSCs undergoes multicellular self-organization which depends on intrinsic, autonomous properties of cells and intercellular communication involving molecules known as morphogens[43–46]. To facilitate our study of hiPSC differentiation, we employ a cell line for which differentiation can be induced synthetically[47]. Prior imaging observations on this cell line imply that cell fate acquisition and localization are influenced by neighboring cells[47,48].

We find that varying the amount of chemical induction reveals trends in pattern formation across loss of pluripotency by applying our pipeline. We also show that studying spatial information enhances our ability to detect and examine these trends quantitatively. Finally, we analyze the patterning dissimilarities between all differentiated cells and those whose differentiation is synthetically induced. In doing so, we are able to draw a connection between local changes in the neighborhood composition of differentiated cells and system-level effects of synthetic induction. Thus, we provide evidence for how the interpretation of pattern formation can inform the discovery and study of the molecular mechanisms driving cellular organization and cell fate decision-making. By also comparing our use of TDA with two traditional, fixed-scale methods for spatial statistics, we not only validate our findings but also provide examples where TDA is able to yield additional insight in uncovering subtle structural details and establishing relationships in cellular dynamics across scales.

Results

Pipeline for microscopy image analysis

Our main methodological contribution is a computational pipeline for 2D microscopy image analysis that produces topological, multiscale descriptors of multicellular organization. Our goal in developing this computational framework is to open new avenues for the spatial cell analysis by making the tools from TDA accessible to a broader audience. To this end, we organized the pipeline into three sequential modules: segmentation, cell type identification, and TDA. Each module is automated, can be applied to data from entire image sets, and relies on

a few user-set parameters; see Fig. 1. First, the segmentation module obtains cell-specific locations and signal intensities of n biomarkers (one signal per channel) from each immunofluorescence microscopy image in the input set (Fig. 1a). Then, the cell type identification module categorizes each cell based on their signal intensities into one of 2^n cell types based on a user-selected percentile threshold (Fig. 1b). Finally, the TDA module derives topological descriptors of patterning for a given combination of the 2^n cell types (Fig. 1c–h).

The three modules in our pipeline can be used in sequence to obtain topological descriptors from microscopy image sets without requiring users to perform additional computations like preprocessing their imaging data. Nonetheless, we have designed this computational framework to be modular so that users may swap one module with a similar tool of their choice as long as they respect the data formatting constraints of the next module in the sequence. The reason for this design is that users may prefer to employ tools which are more appropriate to their context than the general-purpose modules which we have written. In the case of the segmentation, we chose a histogram-thresholding based approach[49] because it is one of the most straightforward and widely used methods

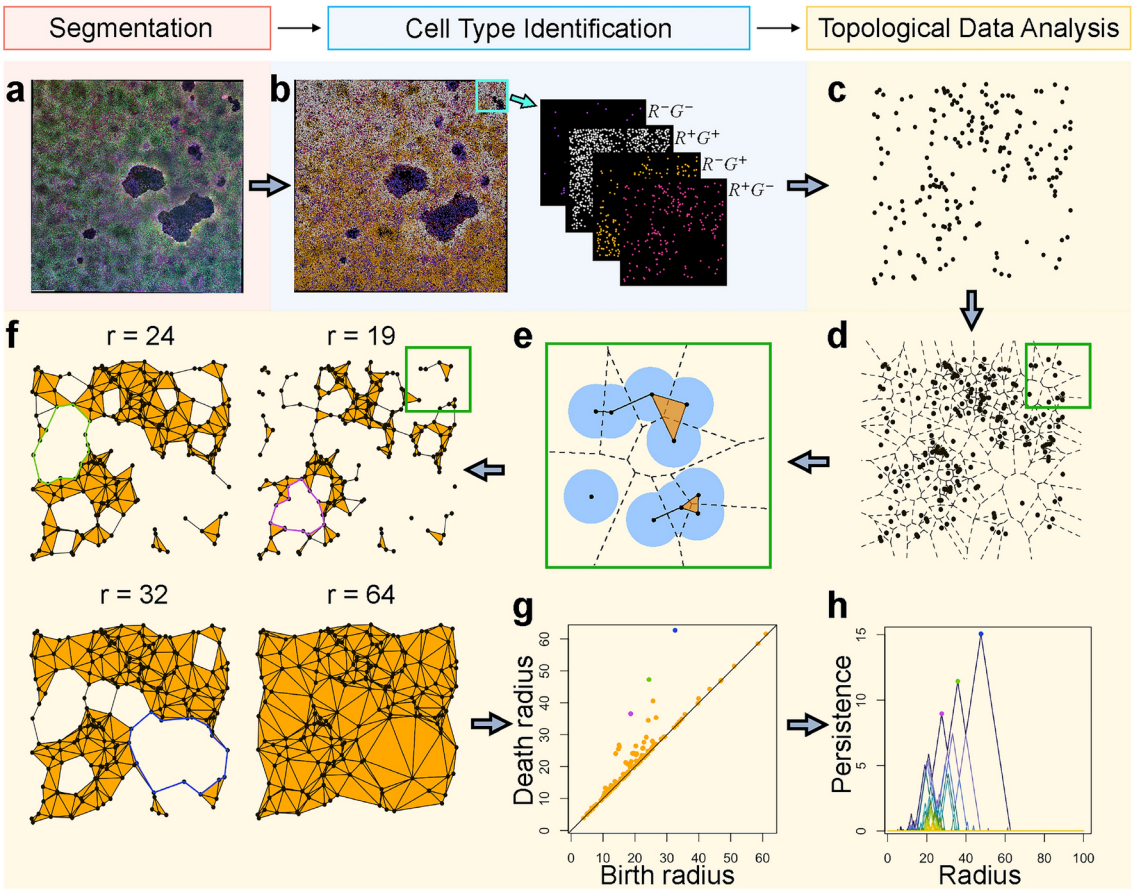

Fig. 1. Our pipeline extracts topological descriptors from microscopy images of multicellular colonies. (**a**) An input microscopy image of cells with a nucleus signal in blue and two other signals: red (R) and green (G) (scale bar, $440\mu m$). This image is processed through the segmentation module of our pipeline to identify cell locations and associate a signal intensity to each cell. (**b**) A discretized version of the microscopy image in which cells are represented as points in the Euclidean plane and categorized into one of four cell types based on signal intensities by the cell type identification module. For the upper right patch, points in each cell type are shown. In general, there will be 2^n cell types identified for n signals based on signal intensity. (**c**) The points of a single cell type ($\mathrm{R}^+\mathrm{G}^-$) in the patch. (**d**) The corresponding Voronoi diagram, a partition of the patch into regions enclosing a portion of the plane that is closest to each point. (**e**) A step in the Delaunay filtration for the top right part of the Voronoi diagram, built by connecting neighboring points with line segments and triangles using a proximity radius r. (**f**) Four stages in the Delaunay filtration of the patch, an increasing sequence of simplicial complexes, with the three most persistent enclosed empty regions shown at the radius r (in pixels) at which they appear, colored in blue, green, and purple, where we have listed the colors in order of decreasing persistence. (**g**) The resulting persistence diagram, a topological summary that encodes the radii at which holes appear and disappear, with points corresponding to the most persistent enclosed empty regions highlighted in the same blue, green, and purple colors. (**h**) The corresponding persistence landscape, a decreasing sequence of piecewise-linear functions, can be combined with statistical and machine learning tools. This is the primary output of the TDA module in our pipeline. For simplicity, the $(\,\mathrm{birth}\,+\,\mathrm{death}\,)/2$ axis is labeled 'Radius' and the $(\,\mathrm{death}\,-\,\mathrm{birth}\,)/2$ axis is labeled 'Persistence'. Note that the blue, green, and purple points in the persistence diagram map to points of the same color in the persistence landscape.

for image segmentation[50]; however, there are several other segmentation tools available[3,5,50–56]. The approach we take to identify cell types is only based on signal intensities; see Methods for details. As such, it is broadly applicable because it does not require training nor rely on biological assumptions associated to a particular data set. However, users may opt to identify cell types differently based on additional data or insights[3,4]; there are even segmentation-free approaches to cell type identification[57] that could be adapted to work with our pipeline. Lastly, there are several statistical approaches which have been traditionally applied to the context of analyzing spatial data, including Ripley's K-function, F-function, cell point density, and (Voronoi) nearest neighbors[6–12]; these approaches may be used in addition to our TDA module for comparison and/or validation. In general, we recommend that users apply TDA to analyze multicellular behavior in their microscopy images when this behavior involves interactions across multiple scales and/or results in complex, heterogeneous organization (e.g. Turing-like patterning)[31,37,58].

We now give a brief summary of persistence homology[28] which is employed in the last module of our pipeline and is further detailed in the Methods. Let P be a collection of points on the plane (Fig. 1c). The *Voronoi diagram* of P (Fig. 1d) partitions the plane according to which point in P is nearest. Subject to the constraints of the Voronoi diagram we grow balls of radius r around every point in P, and introduce edges and triangles connecting pairs and triples of points whose balls intersect (Fig. 1e). By allowing r to grow, we get an increasing sequence of structures called the *Delaunay filtration* (Fig. 1f)[59]. Homology detects sequences of edges arranged in a cycle (see the colored cycles in Fig. 1f). Persistence homology records the "births" and "deaths" of empty regions surrounded by cycles as the scale parameter r increases. These $(\,\mathrm{birth}\,,\,\mathrm{death}\,)$-pairs are plotted in the *persistence diagram* (Fig. 1g). In order to apply statistics and machine learning, we convert the persistence diagram to a *persistence landscape* (PL)[30,60] (Fig. 1h). The vectorized PLs (PL vectors) serve as input for downstream tasks. We apply our pipeline to a small patch of hiPSC colonies in Fig. 1c–h. We have highlighted the three most persistent points in the persistence diagram (Fig. 1g), their corresponding representative cycles (Fig. 1f), and the corresponding peaks in the PL (Fig. 1h).

The position and height of a PL provide key insights into the structure of the underlying data. If a PL is shifted to the left, it indicates that the topological features appear and disappear at smaller scales. Conversely, a PL shifted to the right reflects that topological features occur and resolve at larger scales. The height of the PL is also important: a taller PL suggests that the topological features are significant and exhibit greater persistence, whereas a shorter PL indicates weaker, less pronounced structural features. See Supplementary Fig. S1 for a simple example of how PLs can be used to distinguish noise from prominent topological features across scales.

Microscopy imaging of human induced pluripotent stem cells

We apply our pipeline to study differentiation and pattern formation in a genetically engineered hiPSC line capable of overexpressing transcription factor GATA6, a marker for differentiation potential[61]. We focus on the initial loss of pluripotency and thus consider protein markers GATA6 (differentiation) and NANOG (pluripotency). This cell line has been transduced with a gene construct (Fig. 2a) to express HA epitope-tagged GATA6 (GATA6-HA) transgenes by Doxycycline (Dox) induction, which results in rapid colony patterning between pluripotent GATA6- cells and GATA6+ cells[47]. The extent of the GATA6-HA activation can be regulated by the amount and duration of the Dox treatment. We tested and imaged different Dox concentrations of 0, 5, 15, and 25 ng/ml for three days (Fig. 2b and Supplementary Fig. S2). In this work, pan-GATA6 quantification represents the summative GATA6 expression levels in each cell of both induced GATA6-HA and endogenous GATA6, whereas the HA levels only represent the subset corresponding to GATA6-HA expression induced by the gene circuit. We observed that higher Dox concentration resulted in increased pan-GATA6 and HA expression levels, and that pan-GATA6 expression was likely to be higher than that of HA (Fig. 2b–c).

Using our pipeline, we segmented the images based on nuclear morphologies and acquired cellular positional information and intensity profiles of the immunofluorescent markers. The green fluorescent signal serves as a marker of NANOG expression, and the red fluorescent signal corresponds to either immunolabeling of pan-GATA6 or HA in the corresponding cell cultures. Every population of cells was discretized into binary groups using threshold values for each signal intensity. Due to computational limitations, we split each discretized image into 16 even, non-overlapping patches. As a result, in each Dox concentration group we have 240 patches (from 15 images) separately in pan-GATA6 and HA populations. We provide full details on our thresholding in the Methods and give an idea of our approach below for a given population of cells. The value corresponding to the 75^{th} percentile of the green intensity distribution for the *highest* Dox treatment group (25 ng/ml) was assigned as the green threshold for all Dox treatment groups. All cells with green signals above this threshold are categorized as green positive, G^+ (Fig. 2c). On the other hand, the values corresponding to the 75th percentile for the *lowest* Dox treatment groups (0 ng/ml) were assigned as the red (pan-GATA6 or HA) thresholds for all Dox treatment groups. All cells in the pan-GATA6 or HA populations with red signals above their corresponding red threshold are categorized as red positive, R^+ (Fig. 2c). We selected the 75th percentile after assessing how the populations were partitioned into each cell type for different percentile choices (Supplementary Table S1).

To verify that the pattern formation observed in the Dox treatment groups (Fig. 2 and Supplementary Fig. S2) does not arise prior to the treatment protocol, we also obtained microscopy images at the start of the protocol (Supplementary Fig. S3) and processed them with our pipeline as well. We make the simplifying assumption that pluripotent cells are $\mathrm{R}^-\mathrm{G}^+$ cells in the treatment groups because GATA6 is a marker for differentiation potential[61] and because NANOG is a marker for pluripotency[62–65]. It follows that $\mathrm{R}^-\mathrm{G}^+$ cells are likely the most biologically similar to the (pluripotent) cells in the day 0, pre-Dox treatment group (i.e. those present at the start of the protocol).

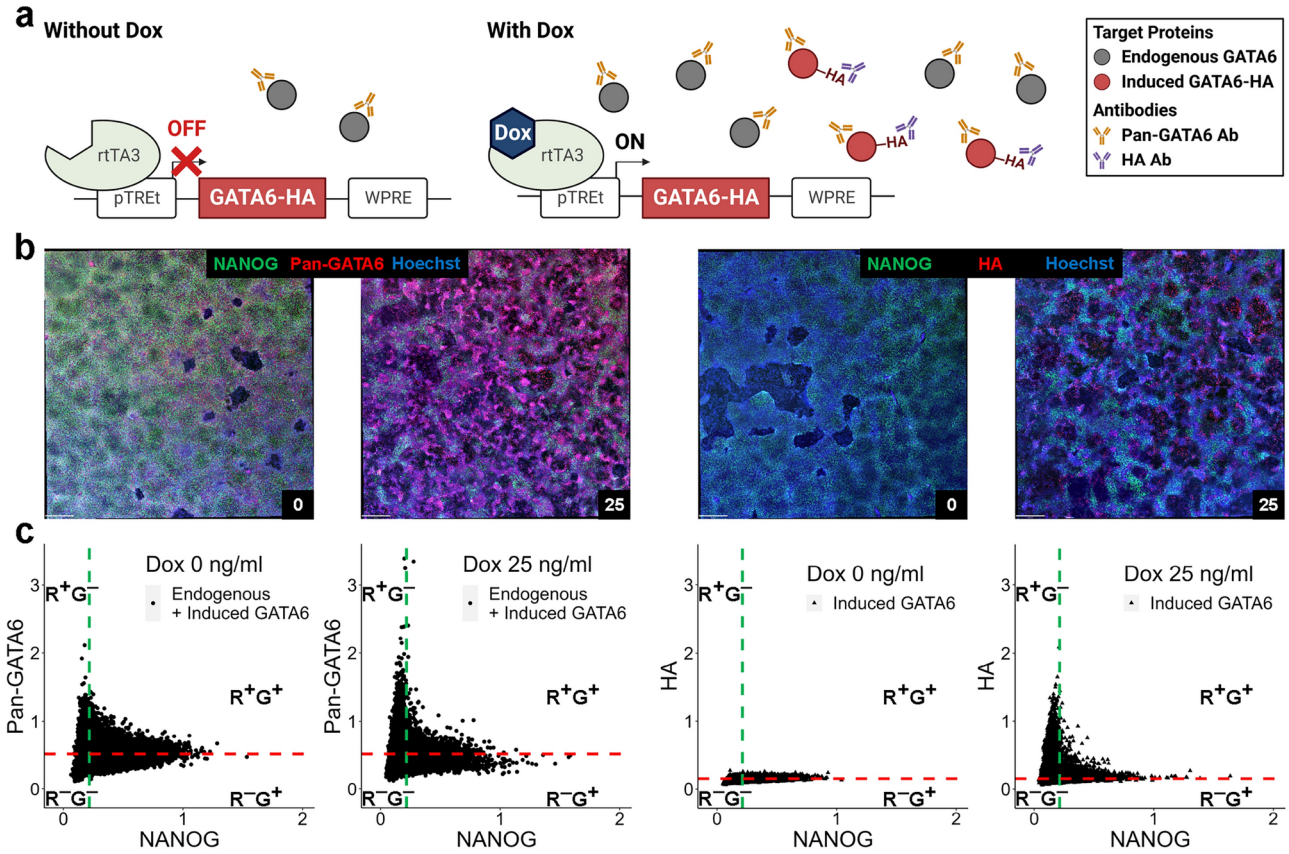

Fig. 2. Artificial induction of exogenous GATA6-HA occurs within the context of endogenous GATA6 expression. (**a**) Gene circuit for chemical induction of GATA6-HA expression. The pan-GATA6 antibody can detect both induced GATA6-HA and endogenous GATA6 (i.e. the total amount of GATA6), whereas the HA antibody can only detect induced GATA6-HA (i.e. a subset produced via induction of the total amount of GATA6). (**b**) Representative immunofluorescence images of NANOG and pan-GATA6 or HA at 0 and 25 ng/ml Dox concentrations (scale bar, $440\mu m$). Using Volocity, we applied the gamma changes (gamma of 1.5) after brightness enhancement on all stitched large images used in the figure for better contrast for representation (see Methods). (**c**) Quantification of segmented images by each channel from (**b**). Green (NANOG) and red (pan-GATA6 or HA) fluorescent intensities are normalized to the corresponding nuclear Hoechst value in blue. The threshold for the signal in the green channel is given by the green dotted line, and the red dotted line indicates the threshold for the signal in the red channel. The four cell types based on these thresholds are given by the labels $\mathrm{R^+G^-}$, $\mathrm{R^+G^+}$, $\mathrm{R^-G^+}$, and $\mathrm{R^-G^-}$ of the corresponding quadrant.

Quantifying differences in pattern formation induced by varying doxycycline concentration

In Fig. 2 and Supplementary Fig. S2, we notice visible differences in the spatial organization of cells across the Dox treatment groups for both the pan-GATA6 and HA populations. As Dox concentration increases, cells exhibiting high pan-GATA6 start forming dense regions separated by an increasing number of empty regions (Supplementary Fig. S2a top); we can observe these dense and empty regions when staining with HA (Supplementary Fig. S2a bottom) even though there are fewer cells with high HA intensity (Fig. 2c). We also notice differences between these treated groups imaged at day 3 and the day 0, pre-treatment group (Supplementary Fig. S3). Aside from the general absence of cells with high pan-GATA6 and HA intensity, we observe a lack of dense regions of cells and a seemingly stochastic assortment of empty regions. We explore differences between the Dox treatment groups and the pre-treatment group in the Supplementary Information and find that the cell patterning structure in all Dox treatment groups differs from the one in the pre-treatment group. This finding agrees with our expectation that the synthetically induced differentiation would trigger various local mechanisms which would affect multicellular organization in turn.

To study the effect of Dox dosing on pattern formation further, we generated the average PL of each cell type in the pan-GATA6 (Supplementary Fig. S4) and HA (Fig. 3 and Supplementary Fig. S5) populations per treatment group; see Methods for details. We also ran two-sample permutation tests on PL vectors of each pair of Dox concentrations (per cell type); see Methods for details on how these permutation tests were performed. The p-values for almost all pairwise permutation tests are less than $5e{-}2$ (Supplementary Table S2), the statistical significance threshold used in this work, substantiating the idea that topological features differ from one Dox concentration to another for both pan-GATA6 and HA groups. The only p-values greater than $5e{-}2$ are associated to the $\mathrm{R^-G^-}$ cell type in the pan-GATA6 group, which implies that topological structure within

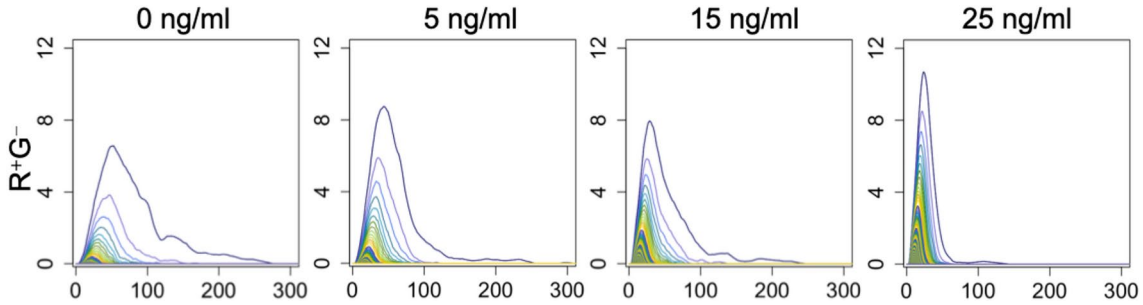

Fig. 3. Comparison of average persistence landscapes across various Dox concentrations for R^+G^- cell type in the HA group. As Dox concentration increases, the average persistence landscape gets taller and narrower.

this cell type does not change much across Dox concentrations. This observation is also supported by the visual similarity of their average PLs (Supplementary Fig. S4). Using our 75th signal percentile thresholds (Fig. 2c), we organize our analysis of the average PLs in Fig. 3, Supplementary Fig. S4, and Supplementary Fig. S5 based on R^+ and R^- cell populations.

We observe trends that correspond to an increase in differentiation from the population distributions (Supplementary Table S1) and the average PLs of R^+ cells in the pan-GATA6 (Supplementary Fig. S4) and HA (Fig. 3 and Supplementary Fig. S5) groups. As expected, the number of R^+ cells increases and the number of R^- decreases as the Dox concentration increases. The average PLs of R^+G^- cells are compressed left, indicating that most cycles in higher Dox treatment groups are born earlier (Supplementary Fig. S4 and Fig. 3). The increasing height of average PLs (across all depths) also demonstrates that the cycles in higher Dox treatment groups are also more persistent, with labyrinthine patterns emerging in the 25 ng/ml group. These trends suggest that differentiated cells stay in close proximity to one another even as their number increases and as they spread into and around empty regions. At the first depth (the outermost function), the height of the average PLs of R^+G^+ cells also increases with Dox concentration in the pan-GATA6 and HA groups; however, after the first depth, the nested heights of the average PLs do not always follow this trend.

In general, the trends of R^-G^+ cells in the pan-GATA6 (Supplementary Fig. S4) and HA (Fig. 3) groups are reversed from those of R^+G^- cells. Average PL heights of R^-G^+ cells decrease for all depths in the HA population but only after the first few depths in the pan-GATA6 population. This behavior corresponds to the fact that most of the cycles in R^-G^+ populations are less persistent in higher Dox treatment groups. The average PLs also widen as the Dox concentration increases, indicating that cycles in R^-G^+ populations have a wide range of births in higher Dox treatment groups. Together, these observations suggest that most pluripotent cells spread out from each other as their number decreases due to differentiation. The average PLs of R^-G^- cells have the same behavior as the average PLs of R^-G^+ cells in the HA group at most depths (Supplementary Fig. S5); however, the average PLs in the pan-GATA6 population look similar across most Dox treatment groups (Supplementary Fig. S4).

Topological data analysis extends observations from fixed-scale approaches to spatial statistics

We compare our results from the previous section with two more traditional approaches for spatial statistics, Voronoi neighborhood size (VNS) and the F-function, to examine the value of using TDA to study how Dox dosing affects pattern formation. Given that the sequence of structures (i.e. simplicial complexes) which we employ for persistence homology is based on Voronoi diagrams, measuring VNS distributions is an appropriate "fixed-scale" alternative to PLs. We indicate that this measurement is at a "fixed" scale because the neighborhoods of each Voronoi cell remain fixed after the diagram is computed. We generated VNS distributions for each patch, and then aggregated these distributions for each cell type in the pan-GATA6 and HA populations per Dox treatment group using our 75th signal percentile thresholds (Fig. 2c); see Supplementary Fig. S6 for examples. We also generated F-function representations per patch which we summarized via a statistic labeled $F_{\max}$; see Methods for full details. Conceptually, this statistic approximates the radius of the largest empty region in the associated patch[8]. Since the F-function (i.e. the "empty-space function") tracks empty circular regions across multiple radii[9], we determined that $F_{\max}$ would be another appropriate fixed-scale alternative to PLs. As with the VNS distributions, we summarized $F_{\max}$ values for each cell type in the pan-GATA6 and HA populations per Dox treatment group; see Supplementary Fig. S7 for the resulting distributions.

When comparing Dox concentrations for each cell type, we found evidence that using PLs allows us to pick up subtle differences missed by using the VNS and $F_{\max}$ distributions. We used pairwise t-tests on the VNS distributions (Supplementary Table S3) and $F_{\max}$ distributions (Supplementary Fig. S7) per cell type for comparison against the corresponding pairwise permutation tests (Supplementary Table S2) on PLs from the previous section. In general, the pairwise t-test results from both approaches agree with, and thus validate, the pairwise permutation tests on PL vectors. However, there are many instances where the difference in the spatial patterning of R^+G^+ or R^-G^- cell populations is not considered statistically significant when using one (or both) of these fixed-scale approaches while the corresponding difference using PL vectors is statistically significant. Because the trends of the average PLs for R^+G^+ and R^-G^- cells are not always consistent

(Supplementary Fig. S4 and Supplementary Fig. S5), it may be the case that any trends which do exist are too subtle or otherwise difficult to detect by studying VNS or $F_{\max}$ distributions alone.

We also found that PLs allow us to make multiscale observations about the R^+G^- and R^-G^+ cell populations which we cannot establish with VNS or $F_{\max}$ distributions individually. For R^+G^- cells, the VNS histograms (Supplementary Fig. S6) only indicate that the number of R^+G^- cells increases with Dox concentration. Even if we consider the density of R^+G^- cells more directly (Supplementary Table S4), we cannot gain any further insight. This trend is not only expected but can be observed by just looking at cell population changes (Supplementary Table S1). When we consider changes in the $F_{\max}$ distributions (Supplementary Fig. S7) alone, we see that the largest empty regions are decreasing in size without disappearing. It is only by jointly combining the observations from VNS and $F_{\max}$ distributions that we can assert that R^+G^- cells remain close to one another as they spread and grow in number (with increasing Dox concentration). Note that this assertion spans multiple scales because of how VNS and $F_{\max}$ are computed (see Methods). However, we are able to draw the same conclusions and establish the aforementioned, multiscale relationship by just considering changes in the height and width of the average PLs in the previous section. For R^-G^+ cells, we again find that the average PLs are able to establish a multiscale relationship (detailed in the previous section) which is jointly corroborated by the two fixed-scale approaches.

Spatial information improves classification and prediction of doxycycline treatment groups

While some PL trends across Dox treatment groups are apparent by eye, we wanted to determine if the variations of the topological signatures produced by our pipeline are sufficient to classify image patches according to their Dox concentration. We applied multiclass support vector machines (SVM) to concatenations of vectors generated from the PLs of all cell types (per patch). For comparison, we also performed the same analyses on cell count data by constructing a four-dimensional vector for each patch which holds the count for each cell type. Note that we normalize both PL and cell count vectors; see Methods for more details about SVM and the train/test data split. We provide the confusion matrices of one instance of multiclass SVM in Supplementary Table S5 for the PL vectors and cell count vectors. Averaging across 20 instances of multiclass SVM, we accurately classified 71.8% patches in the pan-GATA6 group and 73.5% in the HA group with their corresponding Dox concentrations using PLs. The average accuracy dropped to 67% for the pan-GATA6 group and 68.3% for the HA group when we used the cell count vectors instead, indicating that including spatial information improves our ability to distinguish patches between various Dox concentrations.

Since the Dox concentration is a numerical variable, we also performed support vector regression (SVR) to predict the Dox concentration of each image. We conducted SVR separately on the normalized PL vectors and normalized cell count vectors of each image's constituent patches; see Methods for details. We then averaged the Dox concentration predictions of an image's patches. The resulting Dox concentration predictions are shown in Fig. 4. Supplementary Table S6 has additional information on every box plot in Fig. 4, including the deviation of the median from the actual Dox concentration (called "error"). By comparing the two approaches, we can observe that the errors are higher overall for the cell count vectors than for the PL vectors. The Dox concentration predictions for the 5 ng/ml, 15 ng/ml, and 25 ng/ml images are well separated from one another when using PL vectors; in contrast, when using cell count vectors, the fourth quartile of predictions for the 5 ng/ml images overlaps with the first quartile of predictions for the 15 ng/ml images, and the fourth quartile of predictions for the 15 ng/ml images overlaps with the first quartile of the predictions for the 25 ng/ml images (Fig. 4). In most Dox treatment groups, we also observe that the interquartile ranges of predictions are narrower and the whiskers are shorter using PL vectors instead of cell count vectors.

By considering Supplementary Table S5 and Fig. 4a, we notice that it is difficult to distinguish between 0 ng/ml and 5 ng/ml Dox concentrations with SVM and SVR in both pan-GATA6 and HA populations. This suggests that the 5 ng/ml increase in Dox concentration does not accelerate the differentiation of cells enough to change the pattern formation so that the SVM/SVR models can properly learn the difference between these Dox treatment groups. Although the permutation test on 0 ng/ml and 5 ng/ml Dox treatment groups shows that their topological structures are distinct (Supplementary Table S2), the ranges of their Dox concentration predictions greatly overlap (Fig. 4a). Multiclass SVM and SVR on cell count vectors also do not separate well between 0 ng/ml and 5 ng/ml Dox treatment groups (Supplementary Table S5 and Fig. 4b). Because most incorrect classifications in the multiclass SVM confusion matrices involve 0 ng/ml and 5 ng/ml (Supplementary Table S5), we performed pairwise SVM on the PL and cell count vectors of two distinct Dox treatment groups. We conducted 20 runs of SVM for each pairwise classification, and reported average accuracies over these runs; see Methods for more details. The average accuracy of pairwise SVM is high (i.e. greater than $\sim 90\%$) for almost every Dox treatment group pair using PL vectors (Supplementary Table S7); only the 0 ng/ml vs 5 ng/ml classification had a low average accuracy (62.67% for the pan-GATA6 group and 63.48% for the HA group). The average accuracy of pairwise SVM on the cell count vectors are lower than those on the PL vectors, except for the 0 ng/ml vs 5 ng/ml classification which is nevertheless fairly low. In addition, we performed permutation tests on the average accuracies for the two vector types. Most p-values were $1e{-}4$, indicating statistical significant differences in SVM accuracies favoring PL vectors. The only exception was the 0 ng/ml vs 5 ng/ml classification for pan-GATA6, where the p-value of $9.45e{-}2$ indicates that the difference is not statistically significant; for HA, cell count vectors were favored with a p-value of $1e{-}4$ (Supplementary Table S7).

Distinguishing pattern formation of total GATA6 vs synthetically induced GATA6

We next considered the following question: "Does the choice of a biological marker, pan-GATA6 or HA, affect how we perceive the structure of differentiated cell colonies?" Recall that the HA-specific antibody only targets synthetically induced GATA6 expression whereas the pan-GATA6 antibody stains for total GATA6 expression (detecting both endogenous GATA6 plus synthetically-induced GATA6). Because prior work suggests the local

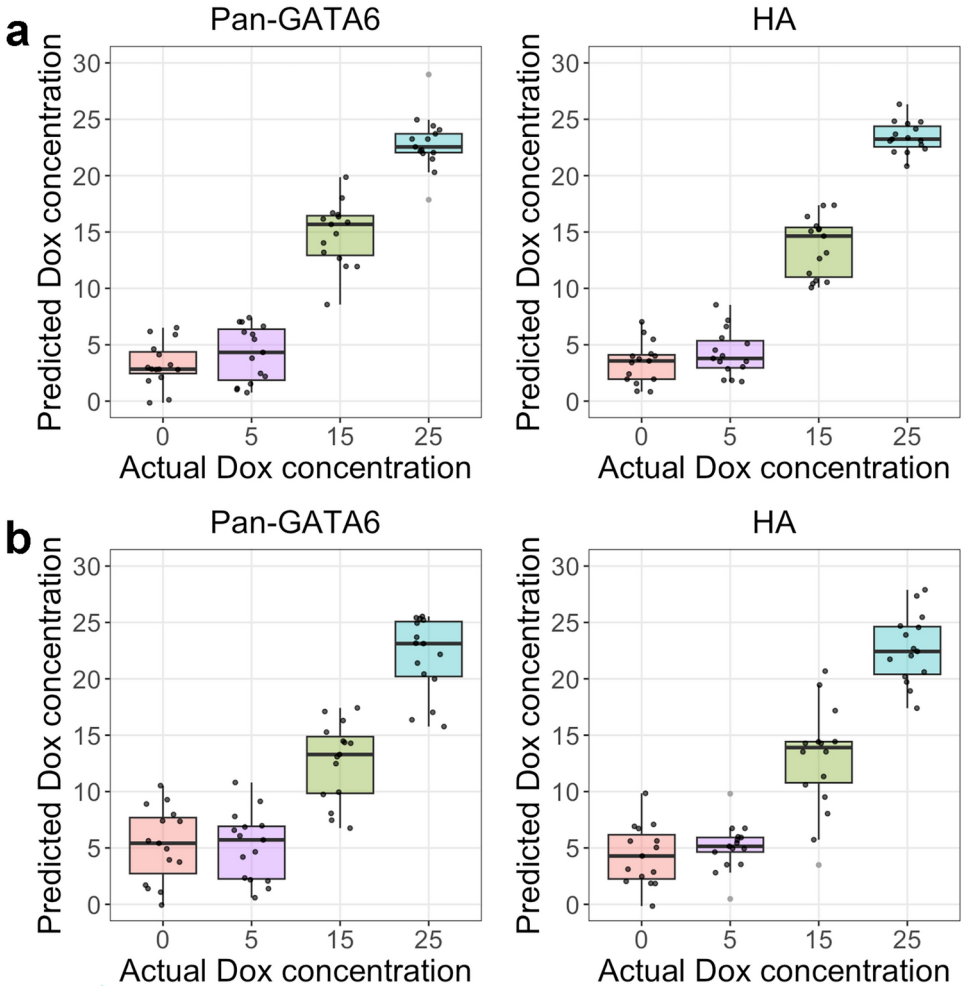

Fig. 4. We performed support vector regression on (**a**) persistence landscape vectors and (**b**) cell count vectors extracted from patches; we then averaged the Dox concentration predictions of each image's patches. The image predictions of each Dox treatment group are spread horizontally for better visualization, and the outliers are marked with gray points. The Dox concentration predictions of 5 ng/ml, 15 ng/ml, and 25 ng/ml images are close to their actual values and well separated from one another when using persistence landscape vectors (**a**, top). When using cell count vectors (**b**, bottom) the results are broadly similar – there is no overlap between the predictions in the second and third quartiles (the boxes in the box and whisker plots) for the 5 ng/ml, 15 ng/ml, and 25 ng/ml images. However, in contrast to the persistence landscape predictions, the cell count predictions in the fourth quartile for the 5 ng/ml images overlap with those in the first quartile for 15 ng/ml images (the whiskers in the box and whisker plots) and the predictions in the fourth quartile for the 15 ng/ml images overlap with those in the first quartile for 25 ng/ml images. The Dox concentration predictions of 0 ng/ml images are not well distinguished from 5 ng/ml images in both cases.

expression of GATA6 in neighboring cells may result in organization and cell fate specification[47,48], answering this question positively could imply the presence of intercellular communication among neighborhoods of cells expressing (endogenous and/or synthetic) GATA6. To address the question above, we examine cells with high red (R^+, pan-GATA6^{high} or HA^{high}) and low green (G^-, NANOG^{low}) intensity in the highest (25 ng/ml) Dox treatment group. We observed subtle, visual differences (e.g. number and size of holes) between microscopy images from both pan-GATA6 and HA populations (Fig. 2b and Supplementary Fig. S8). However, our goal is to quantitatively determine if there is a difference in the pattern formation of these two populations.

For baseline validation, we compared the G^+ PL vectors of the pan-GATA6 and HA populations in the 25 ng/ml Dox treatment group using a permutation test and obtained a non-significant p-value of $2.38e{-}1$. This indicates that the G^+ (NANOG^{high}) cells in both the pan-GATA6 and HA populations exhibit topologically similar patterning, which aligns with the expectation that NANOG expression should remain consistent regardless of pan-GATA6 or HA staining. To answer the question proposed above, we performed a permutation test on R^+G^- PL vectors from the pan-GATA6 and HA populations to determine if the cell differentiation patterns are statistically distinct; see Methods for details. We obtained a p-value of $1e{-}4$ which strongly supports the claim that the topological structures of these patterns are different. Thus, the choice of a biological marker does affect how we perceive colony structure, with GATA6 antibodies showing distinct differences.

We also performed t-tests on the VNS distributions and the $F_{\max}$ distributions of R^+G^- cells in the pan-GATA6 and HA populations from the 25 ng/ml Dox treatment group to compare these fixed-scale approaches with TDA with mixed results. We obtained a p-value of $5.03e{-}1$ for the t-test comparing the VNS distributions, indicating that the difference between the pan-GATA6 and HA populations is not statistically significant when we only focus on local neighborhood sizes. On the other hand, we obtained a p-value of $8.06e{-}8$ for the t-test comparing $F_{\max}$ distributions, validating the permutation test results by PLs. This result builds on our intuition that the difference in patterning of the pan-GATA6 and HA populations comes from the empty regions within these populations. Unlike with PLs, a shortcoming of using $F_{\max}$distributions is that we cannot precisely discern how differently the R^+G^- cells are distributed around empty regions in the HA and pan-GATA6 populations from this method alone (Supplementary Fig. S7). As we discuss in the next section, we gain more information by examining the average PLs of both populations.

Interpreting spatial differences in the patterning of total GATA6 and synthetically induced GATA6

Our next goal is to understand the dissimilarities between the cell differentiation patterns of pan-GATA6 and HA groups; see Supplementary Fig. S8 for example. We can visualize these topological dissimilarities by subtracting the HA average PL (Fig. 5a middle) from the pan-GATA6 average PL (Fig. 5a left) and plotting the difference

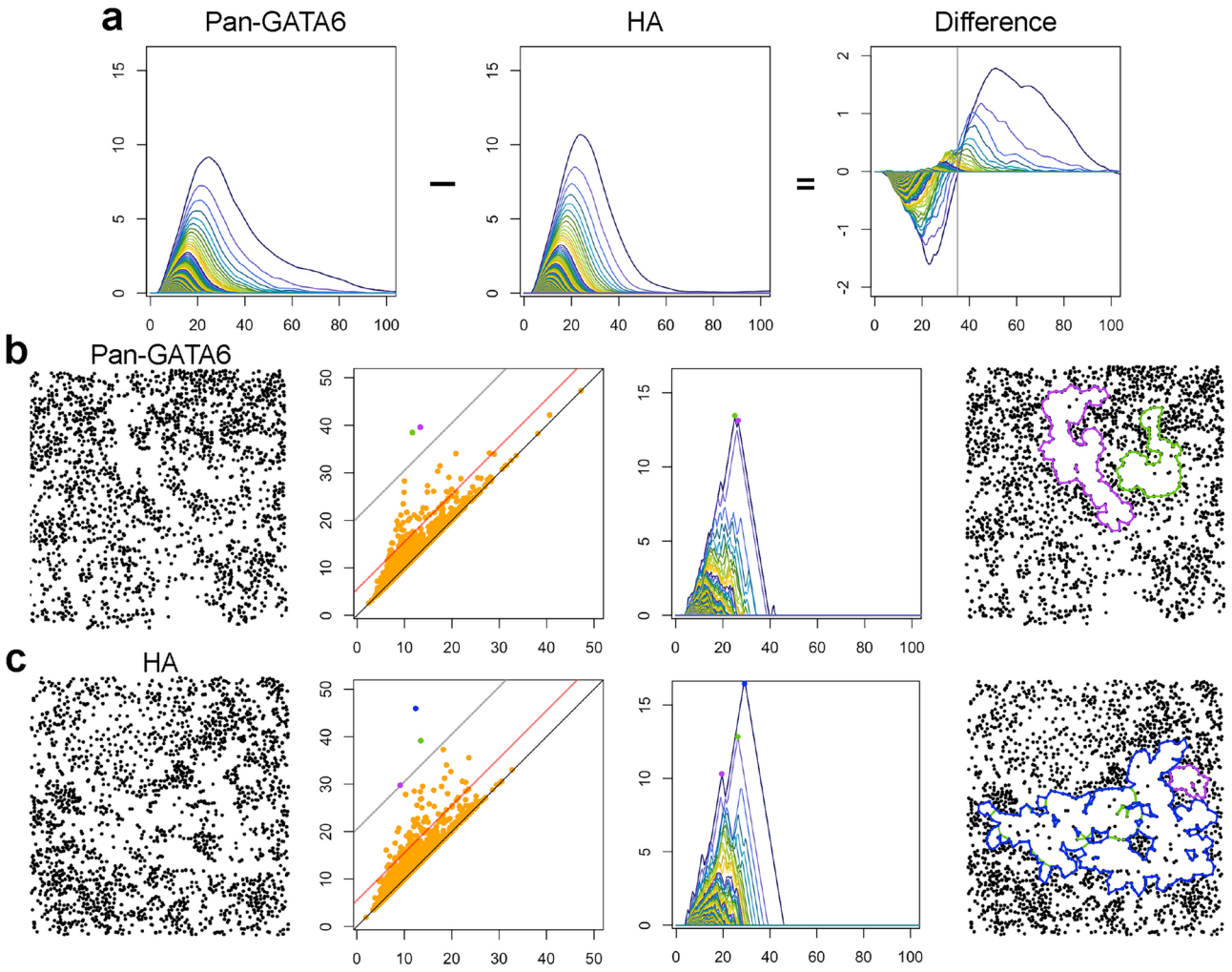

Fig. 5. Comparison between pattern formations in the pan-GATA6 and HA groups for R^+G^- cell type with 25 ng/ml Dox concentration. (**a**) Average persistence landscapes of the pan-GATA6 and HA groups and their difference. The gray vertical line splits the difference plot into two parts showing distinct types of dissimilarities between the cell differentiation patterns. (**b**) The TDA pipeline applied to a representative patch of the pan-GATA6 group, and (**c**) a representative patch of the HA group. For each patch, its persistence diagram, persistence landscape, and most persistent cycles are shown. In the persistence diagrams, the gray line indicates the persistence threshold of 20 and the red line marks the threshold of 5. There are 2 points above the gray line in the pan-GATA6 patch and 3 points in the HA patch; the corresponding representative cycles of these points are shown in the rightmost column. Note that the green cycle lies within the region enclosed by the blue cycle in the HA patch.

(Fig. 5a right). We observe two distinct parts in the difference plot separated by the vertical (gray) line defined by the equation $(\text{birth} + \text{death})/2 = 35$. We begin by observing that the side left of the vertical line is mostly negative (Fig. 5a), indicating that the cycles that appear early in the HA group are more persistent than those in the pan-GATA6 group. Cycles that are more persistent surround larger empty regions, so HA populations have more of these regions than pan-GATA6 populations. This difference can be explained by the fact that HA antibody only detects induced GATA6 expression; thus, neighboring R^+G^- cells in the HA populations may be kept from entering empty regions by other cells with high levels of total GATA6 expression.

To better grasp the difference in number of large empty regions in the pan-GATA6 and HA groups, we analyze the points in their persistence diagrams. In particular, we count the points which are above certain persistence thresholds and satisfy the inequality $(\text{birth} + \text{death})/2 < 35$, which corresponds to the left side of the difference plot (Fig. 5a). We choose low (5) and high (20) thresholds to identify points corresponding to medium and large holes; while these thresholds are not necessarily optimal, they effectively capture significant topological features while minimizing the impact of noise. Fixing the persistence threshold at 20, there are 182 more of these points in the HA group than in the pan-GATA6 group; the average number of such points is 3 in HA and 2 in pan-GATA6. If we lower the persistence threshold to 5, then the difference in number increases to 2,368, and the averages become 84 for HA and 74 for pan-GATA6. We choose two representative patches from each group with about 1,900 cells each to visualize these dissimilarities (Fig. 5b–c).

Now, we focus on the right side of the difference plot (Fig. 5a), which is positive because the average PL of the pan-GATA6 group is wider than the average PL of the HA group. This implies there are more cycles that are born later in the pan-GATA6 populations than in the HA populations; most appear in patches from the pan-GATA6 group where cells enclosing large empty regions are relatively far apart. By analyzing the histograms of cell counts per patch of both groups (Supplementary Fig. S9), we observe that the count distribution of the pan-GATA6 group is also wider than that of the HA group. This observation provides further evidence for our assertion about the heterogeneous spread of R^+G^- cells in the pan-GATA6 group. There are several patches in the pan-GATA6 group whose cell counts are below the count distribution of the HA group; most contain large empty regions which are enclosed by few cells. The existence of these patches results in the elevated region on the right side of the difference plot (Fig. 5a). If we omit patches with less than 1,000 cells and re-compute the two average PLs for the remaining patches, then the elevated region mostly diminishes in the difference plot. Specifically, the total height of the elevated region decreases by approximately four times, and its PL depth reduces by at least the same factor (Supplementary Fig. S10). Thus, the presence of large holes which appear later is not representative of the pan-GATA6 pattern formation. We conducted additional analysis using stitched image files without brightness and shading corrections (see Methods) to verify that our findings were not affected by this correction process; see the Supplementary Information for details.

Discussion

Emergent organization in multicellular systems occurs through mechanisms that operate on multiple scales, ranging from juxtacrine communication and mechanosensing of attached neighboring cells to long-range diffusible morphogens across longer distances. We introduced a general-purpose pipeline to quantify cell pattern formation from microscopy images using topological features (persistence homology) that capture holes in the spatial organization of cells and applied it to images of differentiating human induced pluripotent stem cell (hiPSC) colonies. The functional representations of persistence homology, called persistence landscapes (PLs), allowed us to successfully distinguish features that reflect experimental conditions which varied the loss of pluripotency within an imaged colony as well as the subtle differences between immunofluorescence patterning associated with use of two antibodies that target different populations of GATA6 proteins.

Our topological summaries revealed trends of how the spatial organization of each cell type changes in hiPSC colonies that are capable of synthetically inducing GATA6 as Dox concentration increases. This system is ideal for testing topological data analysis due to prior reports of Turing-like patterning that occurs during early differentiation of GATA6 expression[47] that were attributed to heterogeneity in cellular decisions based upon high GATA6 (reflected in the synthetic HA population) or low GATA6 levels. In particular, our results suggest that differentiated cells remain near each other even as their number grows with increasing synthetic induction. On the other hand, most pluripotent cells seem to spread away from each other as their numbers shrink due to differentiation. By comparing our multiscale TDA approach with two traditional, fixed-scale approaches to spatial statistics, we validated our results and uncovered instances where TDA is able to connect observations from these approaches and/or provide additional insight. SVM classification and SVR prediction on PL vectors of different Dox treatment groups yielded better accuracy than on simpler cell count vectors, indicating that multicellular spatial information is richer and more valuable for distinguishing between various Dox concentrations.

Our work also demonstrated that dissimilarities in the spatial patterning of the pan-GATA6 and HA-tagged populations are statistically significant. We were able to detect that the HA group, which reflects synthetic induction of GATA6, has more persistent cycles than the pan-GATA6 group using PLs; these cycles correspond to ample (relative to the distance between neighboring cells) empty regions within the microscopy images. Further investigation showed that the HA populations had more large empty regions than the pan-GATA6 populations. Because pan-GATA6 detects induced and endogenous GATA6 levels, the difference in patterning between these groups may be caused by heterogeneous expression of endogenous GATA6 in the pan-GATA6^{high}NANOG^{low} populations and the organization of HA^{high}NANOG^{low} cells towards each other because of neighbor-to-neighbor signaling in the context of morphogen diffusion. While it is possible that the synthetically induced GATA6 may activate the endogenous GATA6 transcription, this activation would increase the aforementioned heterogeneity and would not explain the topological differences of the HA group. To investigate this matter further, we imaged cell populations co-stained with both pan-GATA6 and HA antibodies; see Fig. 6a. Combined with our

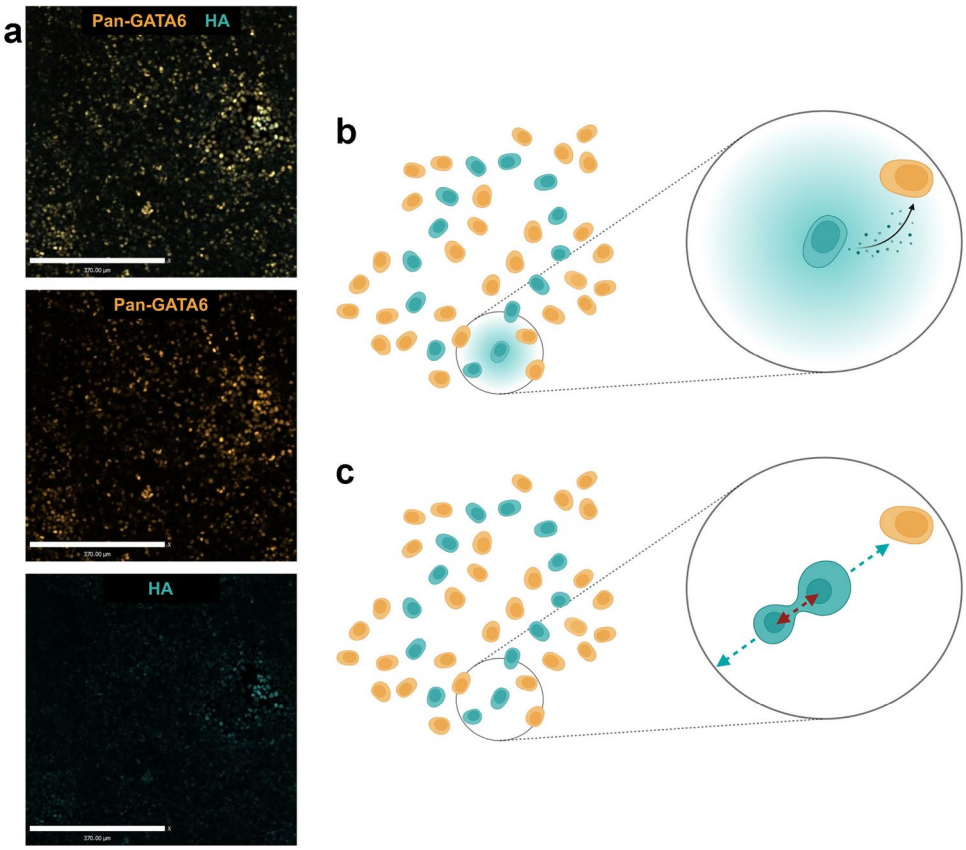

Fig. 6. The topological structures of the pan-GATA6 and HA populations are statistically different according to our results. This difference may be due to cellular organization and intracellular and/or intercellular communication in response to morphogen gradients. (**a**) Immunofluorescence images of co-stained pan-GATA6 and HA antibodies at 25 ng/ml Dox concentration (scale bar, $370\mu m$). The greater presence of cycles in the HA group which are more persistent suggests that cells whose differentiation is synthetically induced may be in close proximity due to chemotaxis or mitosis; we show illustrations of these mechanisms on the right (**b** and **c**). (**b**) We propose that morphogen-mediated signaling driven by heterogeneity in endogenous GATA6 levels could give rise to less persistent holes in the pan-GATA6 group. (**c**) An alternate hypothesis is that cell division alone allows $HA^{high}NANOG^{low}$ cells to stay close to one another.

quantitative analysis, these images suggest that holes in $HA^{high}NANOG^{low}$ cell populations may be partially occupied by other cells expressing (endogenous) GATA6, which would be detected with the pan-GATA6 marker.

Observations from Guye *et al* 2016 and Carter *et al* 2020[47,48] about the hiPSC line used in this work indicate that the local expression of GATA6 in neighboring cells and/or the presence of neighboring differentiated cells influence cell fate acquisition, localization, and emergent differentiation patterning. Accordingly, we propose that the difference in the number and persistence of cycles in the pan-GATA6 and HA populations could imply the existence of cellular organization and intracellular and/or intercellular communication in response to morphogen gradients (Fig. 6b). It is instead possible that $HA^{high}NANOG^{low}$ cells remain relatively close to one another over the course of pluripotency loss because of mitosis (Fig. 6c); however, it is not known whether this cell line undergoes symmetric or asymmetric cell division[66–69]. Further experimentation using time-course data would be required to determine if such preferential cellular organization occurs and/or if morphogen gradients drive pattern formation. Alternatively, simulations from a computational model could be used to support one or both of these hypotheses, with our pipeline serving as a means to quantitatively compare model outputs with experimental data. As a whole, our multi-scale analysis serves as an example of how measuring and interpreting topological features in multicellular pattern formation can inform the discovery and study of the molecular mechanisms driving cell development and behavior.

Our pipeline has several advantages, including its ability to capture the structure of cell patterns across all scales, but it has some limitations as well. The discretization process in the cell type identification module relies on the selection of a threshold; minor alterations of this parameter produce small changes in the populations of each cell type and thus may affect downstream analysis. We chose this threshold carefully by examining the population distributions of each cell type; see Methods for details. More broadly, the abstraction of cell outlines from the segmentation module into points in the discretization module may not be appropriate in instances where the shapes and sizes of cells are as important as their spatial distribution. These circumstances can occur when studying populations of heterogeneous cell morphology, especially elongated cells like neurons where

nuclear coordinates do not properly reflect neighboring cell distances. In these instances, an altered discretization approach could be applied wherein each cell is identified as a collection of points based on its shape and size at a fixed resolution. Given the relative uniformity of cell size and shape in our data, we considered point abstraction to be appropriate for the pluripotent populations analyzed here.

A benefit of TDA is that the persistence diagram and PL are stable for perturbations of the cells in an image with respect to the Hausdorff distance[29]. It follows that they are insensitive to certain imperfections in the microscopy images and the segmentation of these images, such as small changes in the positions of the cells and the misidentification of a pair of neighboring cells as a single cell, or vice versa. However, the persistence diagram and PL are sensitive to other changes, such as a cell with no nearby cells being omitted, being erroneously added, or being mislabeled due to a small change in its color intensity. Careful attention to imaging protocols can mitigate the issue of cells being omitted or added, especially given the strength of nuclear stain DAPI[70]. Color intensity changes may arise during image correction and/or be caused by stitching microscopy image tiles to produce large images (as done in our data set). There are emerging technologies which allow for microscopy-based imaging of a large field of view without the use of stitching[71,72]; our pipeline could be readily applied to such image outputs because of its general-purpose design. The sensitivity to changes in signal intensity near the channel threshold may be overcome with more computationally intensive TDA methods such as perturbation and averaging[73] or multiparameter persistent homology[34]. The sensitivity to outliers may be overcome by subsampling[74] or denoising[75]. In our approach, these shortcomings are alleviated by averaging over a large number of patches for each image.

The differentiation of stem cells is one of several contexts in which patterns arise based on interactions within a microenvironment. Other contexts include tumor architecture, bacterial biofilm formation, and particle organization on surface materials[76–79]. Research on spatial organization has been drastically growing within the past 5 years, primarily led by spatial transcriptomics and spatial proteomics[80]. Therefore, there is a need for quantitative techniques which assist with interpreting spatial relations between sub/cellular elements and deciphering the complexity of cell-cell interactions and signaling pathways. Multiscale descriptors extracted from structural features in these settings could help elucidate a myriad of underlying mechanisms associated with pattern formation. Our pipeline facilitates the use of TDA to extract these descriptors from fluorescence microscopy images because of its general-purpose, modular, and accessible design. Moreover, the pipeline possesses general applicability toward the patterns acquired in both *in vitro* and *in silico* images, RNA profiling, and medical images.

Methods

Cell culture and differentiation

The PGP-GATA6 hiPSC line was a gift from Ron Weiss (MIT)[47]; please refer to this source for further biological material availability. Cells were maintained in mTeSR Plus (STEMCELL Technologies) at $37^{\circ}\mathrm{C}$ and 5% CO_2 with media being changed every other day and cultures being passaged at 80% confluency. Cell culture plates were coated with Matrigel (Growth Factor Reduced, Corning) diluted in KnockOut DMEM (Gibco) for $1hr$ at $37^{\circ}\mathrm{C}$. During the passage, cells were treated with Accutase (STEMCELL Technologies) for 3 min at $37^{\circ}\mathrm{C}$. The dissociated cells were put in PBS and harvested with centrifuging at 200 g for 5 min. The pellet was resuspended in mTeSR Plus with $10mM$ Y-27632 Dihydrochloride (STEMCELL Technologies) at a final concentration of $10\mu\mathrm{M}$, and the fraction of resuspension was seeded.

Differentiation experiments were performed at an initial cell seeding density of 25,000 cells per cm^2 in mTeSR Plus with $10\mu\mathrm{M}$ Y-27632. The next day, the media was changed to mTeSR Plus with Doxycycline (Sigma-Aldrich) diluted at different concentrations from 0, 5, 15, and 25 ng/ml respectively, and replaced daily for 3 days. The experiments had three replicates per Dox concentration (0, 5, 15, or 25 ng/ml) and antibody choice that detects GATA6 expression (pan-GATA6 or HA).

Immunofluorescence and imaging

Cells were grown and treated on Matrigel-coated μ-Slide 8 Well chamber slide (ibidi) and fixed for 10 min in 4% formaldehyde diluted in PBS at room temperature. Each chamber was washed three times in PBS for 5 min. Cells were then permeabilized for 10 min in 0.1% Triton X-100 dissolved in PBS with a subsequent washing step. Cells were blocked for $1hr$ in Odyssey buffer (LI-COR Biosciences) and incubated with the primary antibodies overnight at $4^{\circ}\mathrm{C}$, where each dilution factor varied over antibodies. The next day, each chamber was washed three times in PBS for 5 min, and the cells were incubated with the secondary antibodies for 30 min at room temperature. Each chamber was washed three times in PBS for 5 min. As a final step, nuclei in cells were stained with Hoechst 33342 (Invitrogen) for 5 min followed by one wash and the slide was ready for imaging with PBS.

Primary antibodies are Nanog (Cell Signaling Technology 4893S, 1:2000), Gata6 (Abcam ab22600, 1:200, R&D Systems AF1700, which was only used in co-staining, 1:200), and HA (Novus Biologicals NB600-363R, 1:200). Secondary antibodies are donkey anti-mouse Alexa Fluor™Plus 488, donkey anti-rabbit Alexa Fluor™546 (Invitrogen, 1:1000), and donkey anti-goat Alexa Fluor™647 (Invitrogen, 1:1000). All microscopy images were generated by 6×6 multiple fields and acquired by PerkinElmer UltraVIEW VoX Spinning Disk Confocal Microscope and Volocity software (Quorum Technologies) at 10X objective and stitched together with 10% overlap between each tile. The resulting image size is approximately $4,132.5 \times 4,132.5\mu m$ on average. To avoid stitching artifacts, we applied brightness and shading corrections when stitching. Images were obtained from five different locations in every replicate. When acquiring the images, the Volocity software converts real-size measurements to pixel values, $1.45\mu m$ to 1 pixel at 10X objective. Additionally, before starting the protocol to treat the cells with or without Dox, we obtained the images of the day 0, pre-Dox treatment using Nikon CSU-W1 Spinning Disk Confocal Microscope and the image acquisition software called NIS-Elements. The images were obtained at 10X objective and stitched with 10% overlap between each tile, converting $1.3\mu m$ to 1

pixel as real-size measurements to pixel values at the objective. Using Volocity, the stitched large images in the figures were processed with the gamma settings (gamma of 1.5) after brightness enhancement. However, all the quantitative analyses were performed with raw image files.

Labeling cells and signal intensities

We adapted an existing, histogram-thresholding based cell segmentation pipeline[49] to quantify cell-specific signal intensities from immunofluorescence microscopy images. A nuclear-localized stain (e.g. Hoechst or DAPI) must be used with this pipeline to identify relative cell locations and to measure other signal intensities. We normalize non-nuclear signals (e.g. Nanog, Gata6, and HA) based on this nuclear-localized stain to account for physical factors (e.g. z-positioning) that affect how signal intensities are detected on a microscope. Our main modifications to the existing segmentation pipeline allow for the consecutive processing of multiple microscopy images across different treatment conditions. This module in our pipeline produces a CSV file for each image which contains cell locations, signal intensities, and other information.

Signal thresholding and cell type identification

The cell type identification module of our pipeline generates a signal intensity threshold for each input marker (e.g. HA) based on a user-defined percentile value *n* and a baseline microscopy image set. Each threshold is equal to the *n*th percentile of the corresponding signal intensity distribution from the baseline images. The module uses these thresholds on *all* input images to determine whether each of a cell's signal intensities should be considered "high/positive" or "low/negative". Then, this module assigns these discretized values to the cell by updating the CSVs produced in the segmentation module, thereby classifying each cell into a specific cell type. The set of baseline images may be the entire set of input microscopy images or a proper subset. For *m* signals (one signal per channel), the module will generate 2^m cell types for each possible combination. For example, there will be four cell types ($\mathrm{R}^+\mathrm{G}^-$, $\mathrm{R}^+\mathrm{G}^+$, $\mathrm{R}^-\mathrm{G}^+$, and $\mathrm{R}^-\mathrm{G}^-$) for two channels: red (R) and green (G).

In our work, we generated thresholds for NANOG, pan-GATA6, and HA expression while considering that signal intensities may be affected by the choice of imaging location in each well replicate. We partitioned our microscopy images into subsets according to the red marker (pan-GATA6 or HA) and imaging location and then generated green (NANOG) and red thresholds for each subset using the same percentile value *n*. For each subset, we chose the images from the highest Dox treatment group (25 ng/ml) as our baseline image set for the green threshold given that NANOG expression is expected to be lowest when GATA6 expression is highest. We chose the images from the lowest Dox treatment group (0 ng/ml) as our baseline image set for the red thresholds for the opposite reason.

Percentile threshold and patch size selection

We discuss the following parameters chosen for our microscopy images: percentile threshold and patch size. Signal intensity thresholds rely on a user-defined percentile *n*, so the choice of *n* affects the population distributions of each cell type. To determine which *n* to use, we considered how the population distributions of the four cell types change when $n = 65, 70, 75, 80, 85$ using the cell type identification module of our pipeline. We report this information in Supplementary Table S1. Our goal in choosing *n* was to maximize the number of cells classified as $\mathrm{R}^-\mathrm{G}^+$ or $\mathrm{R}^+\mathrm{G}^-$ across both red markers (pan-GATA6 and HA) because NANOG and GATA6 are indicators for pluripotency and differentiation, respectively. The *n* that maximized this number was directly proportional to the Dox concentration in general, so we chose the middle value $n = 75$ for all Dox concentrations.

Our choice of patch size balances (i) the presence of big, rare regions devoid of cells and (ii) the ratio of empty patches for each cell type. Large patches may anomalous, vacant regions which might not be representative of the multicellular patterning associated with differentiation; see Supplementary Fig. S11 for example. On the other hand, splitting images into smaller patches increases the number of empty patches in every cell type (Supplementary Table S8). For instance, the number of empty patches of $\mathrm{R}^-\mathrm{G}^+$ cell type in the pan-GATA6 group is 43.3% when every image is partitioned into 25 patches.

Persistence homology and persistence landscapes

In this section, we describe basic details of persistence homology[28] which is employed in the last module of our pipeline. Let *P* be a collection of points on the plane. A *Voronoi cell* of a point $p \in P$ is the region of the plane such that every point in that region is at least as close to *p* as to any other point in *P*. The union of Voronoi cells covers the whole plane and is called a *Voronoi diagram* of *P*. We gradually connect the points in *P* in the following way: we grow balls of radius *r* around every point in *P*. For every $p \in P$, we denote the intersection of its Voronoi cell and its *r*-ball by $V_r(p)$. For distinct $x, y \in P$, if $V_r(x)$ intersects $V_r(y)$, we connect *x* and *y* with a line segment. If for some $z \in P$, $V_r(x)$, $V_r(y)$, and $V_r(z)$ have a nonempty intersection, then we connect them in a triangle. If no four points in *P* lie on the same circle, then there will be no quadruple intersections. At a fixed *r*, we obtain an object consisting of vertices, edges, and triangles called a *simplicial complex*.

By allowing *r* to grow, we get an increasing sequence of simplicial complexes called the *Delaunay filtration*[59]. To every element of the filtration, we apply homology in degree 1 with binary coefficients, a method from algebraic topology that detects (1-dimensional) holes, namely, sequences of edges arranged in a cycle. Persistence homology records the "births" and "deaths" of empty regions surrounded by cycles as the scale parameter *r* increases. This information is encoded as a collection of $(\,\mathrm{birth}\,,\,\mathrm{death}\,)$-pairs on the upper-left half-plane of a plot called a *persistence diagram*. The distance from the points in the persistence diagrams to the diagonal is proportional to their *persistence*, which is the difference between their birth and death values. In order to apply statistics and machine learning to persistence diagrams, we convert them to *persistence landscapes* (PLs); PLs have unique averages unlike persistence diagrams[30,60].

A PL is a decreasing sequence of piece-wise linear functions (λ_k) with slopes 1, -1, or 0 plotted on the $((\text{birth} + \text{death})/2, (\text{death} - \text{birth})/2)$-plane. More precisely, consider the function $f_{b,d}$: $\mathbb{R} \to \mathbb{R}$ for $b < d$ defined by $f_{b,d}(t) = t - b$ if $b \leq t \leq (b + d)/2$; $f_{b,d}(t) = d - t$, if $(b + d)/2 \leq t \leq d$; and $f_{b,d}(t) = 0$ otherwise. This function captures the persistence of each topological feature recorded in the persistence diagram. Let $\mathcal{D}$ be a persistence diagram. Its corresponding PL is given by the sequence of functions (λ_k), $k = 1, 2, 3, \ldots$, where $\lambda_k : \mathbb{R} \to \mathbb{R}$ is given by defining $\lambda_k(t)$ to be the kth-largest value of $f_{b,d}(t)$ over the points (b, d) in $\mathcal{D}$. The function λ_k is called the kth PL function of $\mathcal{D}$ and the parameter k is referred to as its *depth*. We visualize the PL by graphing its functions on the same plot and depicting them with a set of colors that repeat every 15 depths.

Persistence landscapes for patches of discretized microscopy images

Each discretized, $2,850 \times 2,850$ pixel ($4,132.5 \times 4,132.5$ μm) microscopy image was split into 16 even non-overlapping square patches. For each patch, we considered each cell type individually and applied the following TDA pipeline. We computed the Delaunay filtration, persistence diagrams, and representative cycles using R package 'TDA'[81], which wraps C++ libraries 'GUDHI'[82] and 'Dionysus'[83]. PL computations used the R package 'tda-tools'[84], which wraps the persistence landscape toolbox[85].

To convert PLs into numerical arrays (PL vectors), we discretize each PL function at regular intervals, with a step size equal to 0.3, which is fine enough to capture fine spatial features and provide nice visualizations. For instance, a PL $\lambda_k(t)$ at depth k, defined on the interval [0, 10] with a step size of 0.3, yields a sequence of 34 values, representing $\lambda_k(t_i)$ at $t_i = 0, 0.3, 0.6, \ldots, 9.9$. These sampled values form a numerical array for each depth of the PL. We used all depths of the PLs that are nonzero for plots. However, we limited the analysis to PLs up to depth 30 for computational purposes, concatenating them into a single vector. For example, if each depth yields 34 sampled values, concatenating 30 depths results in a vector of length $(30 \cdot 34) = 1,020$. This approach strikes a balance between capturing significant topological features and ensuring computational efficiency, while minimizing the impact of noise. As a result, we obtained four vectors for every patch, one for each cell type. When analyzing multiple data samples, the average PL is computed by averaging individual PLs pointwise. The average PL enables comparison between datasets and helps identify significant patterns or differences in topological structures, and is visualized and interpreted in the same manner as individual PLs.

Statistics

To check for a statistically significant difference between two groups, we used the permutation test (one-tailed)[86] on PL vectors. Each of the Dox treatment groups consisted of 240 PL vectors (16 patches per image, 15 images), while the pre-treatment group had only 48 PL vectors (16 patches per image, 3 images). A permutation test is a non-parametric test on two or more samples which is often appropriate when the underlying distributions of the samples are unknown[86]. After choosing a test statistic for the permutation test, the *observed* test statistic is computed using the samples as given. Then, the members from both samples are combined and assigned to the two groups randomly, called a *shuffle*. This is repeated n times and the test statistic is recomputed after each shuffle. Our test statistic was the Euclidean distance between the means of the PL vectors of the two groups, and we shuffled the PL vectors 10,000 times. Under the assumption that both groups cannot be distinguished by the test statistic, the p-value was estimated as the proportion of permutations for which the distance was at least as large as the observed distance. We used the threshold of $5\text{e}{-2}$ to determine the significance of the p-value.

For fixed-scale approaches for spatial statistics, we computed the Voronoi neighborhood size (VNS) distributions, the mean number of neighbors in a given radius (i.e. "mean local density"), and F-functions for each patch. We used the cell positional information (i.e., X and Y coordinates) from each patch for our computations and aggregated the output of these computations for each cell type and each Dox treatment group in the pan-GATA6 and HA populations. After we obtained the Voronoi diagram of each patch, the number of adjacent cells to each cell was calculated to obtain the VNS distributions. These VNS distributions were compared between Dox treatment groups in pan-GATA6 and HA populations per cell type using Welch's two-sample t-test with the level of significance as $5\text{e}{-2}$. To acquire the mean local density, we selected the search radius of 25 pixels (i.e. 36.25 μm) and computed the density of each cell.

The F-function is the cumulative distribution function (CDF) of distances from random positions to their nearest neighbors on a given point cloud with a certain pattern. It is also known as the empty space function because it measures the "empty space" between each random position and its nearest neighbor in the point cloud. While the empirical CDF is computed for the given point cloud, the theoretical CDF for a point cloud with a fully random pattern has a predetermined formula. By comparing the two CDFs and whether the empirical F-function falls below, around, or above the theoretical F-function, one can conclude whether the given point cloud has a pattern which is clustered, random, or evenly dispersed, respectively[6]. As described in Bull *et al* 2020[8], we obtained the radius size of the largest empty region (i.e. $F_{\max}$). By only reporting $F_{\max}$, we are not able to fully capture whether the empirical function falls above or below the theoretical function. However, $F_{\max}$ allows us to assign a value per point cloud, enabling us to automatically compute and then perform statistics on distributions of this value across point clouds with various patterns. Using the R package 'spatstat'[6], each patch is converted to a point pattern called a "ppp" object. We then computed the F-function of the object with simulation envelopes that generated 1000 simulated random patterns. Note that the number of points in the simulated patterns was the same as those in the input pattern. We reported $F_{\max}$; if $F_{\max} = 1$, the minimum radius that corresponds to $F_{\max} = 1$ was reported. For comparisons of VNS and $F_{\max}$ distributions between pan-GATA6 and HA groups with R^+G^- cells at 25 ng/ml, we also utilized Welch's two-sample t-test. Statistical differences were determined using a significance level of $5\text{e}{-2}$.

Machine learning

For our machine learning computations, we concatenated the PL vectors of the four cell types to obtain a single vector of length 127,770 for each patch. For every PL vector, we removed the coordinates that were zero. Next, we normalized these vectors coordinate-wise; for each coordinate, we subtracted the mean and divided it by the standard deviation. Afterward, we used normalized PLs as inputs for machine learning.

For binary classification, we used support vector machines (SVM). When there were more than two classes, we used multiclass SVM using the "one-against-one" approach[87]. In this method, binary classifiers are trained for every pair of classes. The predicted class is the one that was chosen the most often by these classifiers. We performed our computations using the R package 'kernlab'[88], using default settings, linear kernel, and cost equal to 10. To estimate SVM model accuracy, we split the data into training and testing sets using 5-fold cross-validation in the binary case and 10-fold cross-validation in the multiclass case. Since our focus was on local structure, we partitioned the patches independently of their source image. (For future use, we recommend partitioning by images to avoid the possibility of data leakage.) In each case, we repeated the classification 20 times and reported the average accuracy.

For regression, we used a variant of SVM called support vector regression (SVR)[87] using the ε-insensitive loss function. For this loss function, the value of ε defines a margin of tolerance where no penalty is given to errors. We chose $\varepsilon = 0.01$ and otherwise used the same hyper-parameters as with SVM. We also used 10-fold cross-validation with 20 repetitions and plotted the average predicted values as box plots (Fig. 4). A box plot provides information about the distribution of data based on five quartiles (excluding outliers): 100th (tip of top line), 75th (top of box), 50th (horizontal line), 25th (bottom of box), and 0th (tip of bottom line). The interquartile range is the difference between the 75th and 25th percentiles. Gray points in a box plot are outliers.

Data availibility

The microscopy images associated with our work, including an example image set, are publicly available on Figshare: https://figshare.com/projects/TDA_Microscopy_Data/148855.

Code availability

Our computational pipeline is freely available via the Apache 2.0 license on GitHub along with example scripts: https://github.com/kemplab/TDA-Microscopy-Pipeline.

Received: 7 May 2024; Accepted: 13 February 2025
Published online: 04 April 2025

References

1. Sanderson, M. J., Smith, I., Parker, I. & Bootman, M. D. Fluorescence microscopy. *Cold Spring Harb. Protoc.* **10**, 1042–1065. https://doi.org/10.1101/pdb.top071795 (2014).
2. Ettinger, A. & Wittmann, T. Quantitative imaging in cell biology. In *Fluorescence live cell imaging Vol. 123 of Methods in Cell Biology* (eds Waters, J. C. & Wittman, T.) 77–94 (Academic Press, 2014). https://doi.org/10.1016/B978-0-12-420138-5.00005-7.
3. Bannon, D. et al. DeepCell kiosk: scaling deep learning-enabled cellular image analysis with Kubernetes. *Nat. Methods* **18**, 43–45 (2021).
4. Amitay, Y. et al. Cell Sighter: A neural network to classify cells in highly multiplexed images. *Nat. Commun.* https://doi.org/10.1038/s41467-023-40066-7 (2023).
5. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
6. Baddeley, A., Rubak, E. & Turner, R. *Spatial Point Patterns: Methodology and Applications with R* (Chapman and Hall, 2015).
7. Ghosal, A., Nandy, A., Das, A. K., Goswami, S. & Panday, M. Emerging technology in modelling and graphics. In *A short review on different clustering techniques and their applications* (eds Mandal, J. K. & Bhattacharya, D.) 69–83 (Springer Singapore, 2020).
8. Bull, J. A. et al. Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Sci. Rep.* **10**, 18624. https://doi.org/10.1038/s41598-020-75180-9 (2020).
9. Parra, E. R. Methods to determine and analyze the cellular spatial distribution extracted from multiplex immunofluorescence data to understand the tumor microenvironment. *Front. Molecular Biosci.* https://doi.org/10.3389/fmolb.2021.668340 (2021).
10. Huang, H., Wang, Y., Rudin, C. & Browne, E. P. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun. Biol.* **5**, 1–11. https://doi.org/10.1038/s42003-022-03628-x (2022).
11. Behanova, A., Klemm, A. & Wählby, C. Spatial statistics for understanding tissue organization. *Front. Physiol.* https://doi.org/10.3389/fphys.2022.832417 (2022).
12. Summers, H. D., Wills, J. W. & Rees, P. Spatial statistics is a comprehensive tool for quantifying cell neighbor relationships and biological processes via tissue image analysis. *Cell Rep. Methods* **2**, 100348. https://doi.org/10.1016/j.crmeth.2022.100348 (2022).
13. Blevins, A. S., Kim, J. Z. & Bassett, D. S. Variability in higher order structure of noise added to weighted networks. *Commun. Phys.* **4**, 1–12 (2021).
14. Libby, A. R. et al. Automated design of pluripotent stem cell self-organization. *Cell Syst.* **9**, 483–495 (2019).
15. White, D. E., Kinney, M. A., McDevitt, T. C. & Kemp, M. L. Spatial pattern dynamics of 3D stem cell loss of pluripotency via rules-based computational modeling. *PLoS Comput. Biol.* **9**, e1002952 (2013).
16. White, D. E. et al. Quantitative multivariate analysis of dynamic multicellular morphogenic trajectories. *Integr. Biol.* **7**, 825–833 (2015).
17. Glen, C. M., McDevitt, T. C. & Kemp, M. L. Dynamic intercellular transport modulates the spatial patterning of differentiation during early neural commitment. *Nat. Commun.* **9**, 1–13 (2018).
18. Mahadevan, A. S. et al. cytoNet: Spatiotemporal network analysis of cell communities. *PLoS Comput. Biol.* **18**, e1009846 (2022).
19. Helmuth, J. A., Paul, G. & Sbalzarini, I. F. Beyond co-localization: inferring spatial interactions between sub-cellular structures from microscopy images. *BMC Bioinformatics* **11**, 1–12 (2010).
20. Basu, S., Kolouri, S. & Rohde, G. K. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proc. Natl. Acad. Sci.* **111**, 3448–3453 (2014).
21. Schnitzbauer, J. et al. Correlation analysis framework for localization-based superresolution microscopy. *Proc. Natl. Acad. Sci.* **115**, 3219–3224 (2018).
22. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **46**, 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X (2009).

23. Adler, R. J., Agami, S. & Pranav, P. Modeling and replicating statistical topology and evidence for CMB nonhomogeneity. *Proc. Natl. Acad. Sci.* **114**, 11878–11883 (2017).
24. Hiraoka, Y. et al. Hierarchical structures of amorphous solids characterized by persistent homology. *Proc. Natl. Acad. Sci.* **113**, 7035–7040 (2016).
25. Giusti, C., Pastalkova, E., Curto, C. & Itskov, V. Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Natl. Acad. Sci.* **112**, 13455–13460 (2015).
26. Stolz, B. J., Tanner, J., Harrington, H. A. & Nanda, V. Geometric anomaly detection in data. *Proc. Natl. Acad. Sci.* **117**, 19664–19669 (2020).
27. Thomas, A. et al. Topological data analysis of C. elegans locomotion and behavior. *Front. Artif. Intell.* **4**, 668395. https://doi.org/10.3389/frai.2021.668395 (2021).
28. Edelsbrunner, H. & Harer, J. Surveys on discrete and computational geometry: Twenty years later. In Goodman, J. E., Pach, J. & Pollack, R. (eds.) *Persistent homology - a survey*, vol. 453, 257–282, https://doi.org/10.1090/conm/453/08802 (American Mathematical Society, Providence, 2008).
29. Chazal, F., de Silva, V. & Oudot, S. Persistence stability for geometric complexes. *Geom. Dedicata.* **173**, 193–214. https://doi.org/10.1007/s10711-013-9937-z (2014).
30. Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**, 77–102 (2015).
31. McGuirl, M. R., Volkening, A. & Sandstede, B. Topological data analysis of zebrafish patterns. *Proc. Natl. Acad. Sci.* **117**, 5113–5124. https://doi.org/10.1073/pnas.1917763117 (2020).
32. Bhaskar, D., Zhang, W. Y. & Wong, I. Y. Topological data analysis of collective and individual epithelial cells using persistent homology of loops. *Soft Matter* **17**, 4653–4664. https://doi.org/10.1039/D1SM00072A (2021).
33. Pritchard, Y. et al. Persistent homology analysis distinguishes pathological bone microstructure in non-linear microscopy images. *Sci. Rep.* https://doi.org/10.1038/s41598-023-28985-3 (2023).
34. Vipond, O. et al. Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors. *Proc. Natl. Acad. Sci.* **118**, e2102166118. https://doi.org/10.1073/pnas.2102166118 (2021).
35. Bonilla, L. L., Carpio, A. & Trenado, C. Tracking collective cell motion by topological data analysis. *PLoS Comput. Biol.* **16**, 1–43. https://doi.org/10.1371/journal.pcbi.1008407 (2020).
36. Nguyen, K. C. et al. Quantifying collective motion patterns in mesenchymal cell populations using topological data analysis and agent-based modeling. *Math. Biosci.* **370**, 109158. https://doi.org/10.1016/j.mbs.2024.109158 (2024).
37. 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**, 1–14. https://doi.org/10.1038/s41540-023-00302-8 (2023).
38. Edwards, P. et al. TDAExplore: Quantitative analysis of fluorescence microscopy images through topology-based machine learning. *Patterns* https://doi.org/10.1016/j.patter.2021.100367 (2021).
39. Manrique-Castano, D., Bhaskar, D. & ElAli, A. Dissecting glial scar formation by spatial point pattern and topological data analysis. *Sci. Rep.* **14**, 19035. https://doi.org/10.1038/s41598-024-69426-z (2024).
40. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872. https://doi.org/10.1016/j.cell.2007.11.019 (2007).
41. Choi, J. et al. A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. *Nat. Biotechnol.* **33**, 1173–1181 (2015).
42. Liu, X. et al. Modelling human blastocysts by reprogramming fibroblasts into iBlastoids. *Nature* **591**, 627–632 (2021).
43. Lander, A. D. How cells know where they are. *Science* **339**, 923–927. https://doi.org/10.1126/science.1224186 (2013).
44. Green, J. B. & Sharpe, J. Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development* **142**, 1203–1211 (2015).
45. Fu, J., Warmflash, A. & Lutolf, M. P. Stem-cell-based embryo models for fundamental research and translation. *Nat. Mater.* **20**, 132–144 (2021).
46. Morales, J. S., Raspopovic, J. & Marcon, L. From embryos to embryoids: How external signals and self-organization drive embryonic development. *Stem Cell Rep.* **16**, 1039–1050 (2021).
47. Guye, P. et al. Genetically engineering self-organization of human pluripotent stem cells into a liver bud-like tissue using Gata6. *Nat. Commun.* **7**, 1–12. https://doi.org/10.1038/ncomms10243 (2016).
48. Carter, S. R., Hislop, J., Hsu, J., Velazquez, J. J. & Ebrahimkhani, M. R. Programmed morphogenesis: Methods and protocols. In *Neighborhood impact factor to study cell-fate decision-making in cellular communities* (eds Ebrahimkhani, M. R. & Hislop, J.) 17–28 (Springer, 2021). https://doi.org/10.1007/978-1-0716-1174-6_2.
49. Nikitina, A. et al. A co-registration pipeline for multimodal MALDI and confocal imaging analysis of stem cell colonies. *J. Am. Soc. Mass Spectrom.* **31**, 986–989. https://doi.org/10.1021/jasms.9b00094 (2020).
50. Garcia-Lamont, F., Cervantes, J., López, A. & Rodriguez, L. Segmentation of images by color features: A survey. *Neurocomputing* **292**, 1–27. https://doi.org/10.1016/j.neucom.2018.01.091 (2018).
51. Thomas, R. M. & John, J. A review on cell detection and segmentation in microscopic images. *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)***Kollam, India**, 1–5, https://doi.org/10.1109/ICCPCT.2017.8074189 (2014).
52. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70. https://doi.org/10.1038/s41592-018-0261-2 (2019).
53. Vicar, T. et al. Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinformatics* **20**, 1–25. https://doi.org/10.1186/s12859-019-2880-8 (2019).
54. Haase, R. et al. CLIJ: GPU-accelerated image processing for everyone. *Nat. Methods* **17**, 5–6. https://doi.org/10.1038/s41592-019-0650-1 (2020).
55. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: A generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106. https://doi.org/10.1038/s41592-020-01018-x (2021).
56. Weigert, M., & Schmidt, U. Nuclei instance segmentation and classification in histopathology images with stardist. In: IEEE International Symposium on Biomedical Imaging Challenges Kolkata. *India***1–4**, 2022. https://doi.org/10.1109/ISBIC56247.2022.9854534 (2022).
57. Park, J. et al. Cell segmentation-free inference of cell types from in situ transcriptomics data. *Nat. Commun.* **12**, 1–13. https://doi.org/10.1038/s41467-021-23807-4 (2021).
58. Turing patterns, 70 years later. *Nature Computational Science***2**, 463–464, https://doi.org/10.1038/s43588-022-00306-0 (2022).
59. Carlsson, G. & Vejdemo-Johansson, M. *Topological data analysis with applications* (Cambridge University Press, 2022).
60. Bubenik, P. Topological data analysis. In *The persistence landscape and some of its properties* (eds Baas, N. A. et al.) 97–117 (Springer International Publishing, 2020). https://doi.org/10.1007/978-3-030-43408-3_4.
61. Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375**, 54–64. https://doi.org/10.1016/j.ydbio.2012.12.008 (2013).
62. Kalmar, T. et al. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7**, 1–16. https://doi.org/10.1371/journal.pbio.1000149 (2009).
63. Wu, J. & Tzanakakis, E. S. Contribution of stochastic partitioning at human embryonic stem cell division to NANOG heterogeneity. *PLoS ONE* **7**, e50715. https://doi.org/10.1371/journal.pone.0050715 (2012).

64. Ochiai, H., Sugawara, T., Sakuma, T. & Yamamoto, T. Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells. *Sci. Rep.* **4**, 7125. https://doi.org/10.1038/srep07125 (2014).
65. Abranches, E. et al. Stochastic NANOG fluctuations allow mouse embryonic stem cells to explore pluripotency. *Development* **141**, 2770–2779. https://doi.org/10.1242/dev.108910 (2014).
66. Knoblich, J. A. Mechanisms of asymmetric stem cell division. *Cell* **132**, 583–597. https://doi.org/10.1016/j.cell.2008.02.007 (2008).
67. Brown, K., Loh, K. M. & Nusse, R. Live imaging reveals that the first division of differentiating human embryonic stem cells often yields asymmetric fates. *Cell Rep.* **21**, 301–307. https://doi.org/10.1016/j.celrep.2017.09.044 (2017).
68. Nakamura, S. et al. Asymmetricity between sister cells of pluripotent stem cells at the onset of differentiation. *Stem Cells Develop.* **27**, 347–354. https://doi.org/10.1089/scd.2017.0113 (2018).
69. Kapinas, K. et al. The abbreviated pluripotent cell cycle. *J. Cell. Physiol.* **228**, 9–20. https://doi.org/10.1002/jcp.24104 (2013).
70. Jonkman, J., Brown, C. M., Wright, G. D., Anderson, K. I. & North, A. J. Tutorial: guidance for quantitative confocal microscopy. *Nat. Protoc.* **15**, 1585–1611. https://doi.org/10.1038/s41596-020-0313-9 (2020).
71. Jhaveri, N. et al. Mapping the spatial proteome of head and neck tumors: Key immune mediators and metabolic determinants in the tumor microenvironment. *GEN Biotechnol.* **2**, 418–434. https://doi.org/10.1089/genbio.2023.0029 (2023).
72. Kumar, T. et al. A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181–191. https://doi.org/10.1038/s41586-023-06252-9 (2023).
73. Bendich, P., Bubenik, P. & Wagner, A. Stabilizing the unstable output of persistent homology computations. *J. Appl. Comput. Topol.* **4**, 309–338. https://doi.org/10.1007/s41468-019-00044-9 (2020).
74. Chazal, F. *et al.* Proceedings of the 32nd international conference on machine learning. In Bach, F. & Blei, D. (eds.) *Subsampling Methods for Persistent Homology*, vol. 37, 2143–2151 (JMLR: W &CP, Lille, France, 2015).
75. Silverman, B. W. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability (Chapman & Hall, London, 1986).
76. Smillie, C. S. et al. Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730. https://doi.org/10.1016/j.cell.2019.06.029 (2019).
77. Hachey, S. J. et al. An in vitro vascularized micro-tumor model of human colorectal cancer recapitulates in vivo responses to standard-of-care therapy. *Lab Chip* **21**, 1333–1351. https://doi.org/10.1039/D0LC01216E (2021).
78. Melia, C. E. et al. Architecture of cell-cell junctions in situ reveals a mechanism for bacterial biofilm inhibition. *Proc. Natl. Acad. Sci.* **118**, e2109940118. https://doi.org/10.1073/pnas.2109940118 (2021).
79. Mahanta, S., Vallejo-Ramirez, P., Karedla, N., Puczkarski, P. & Krishnan, M. Wide-field optical imaging of electrical charge and chemical reactions at the solid-liquid interface. *Proc. Natl. Acad. Sci.* **119**, e2209955119. https://doi.org/10.1073/pnas.2209955119 (2022).
80. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
81. Fasy, B. T., Kim, J., Lecci, F. & Maria, C. Introduction to the R package TDA (2014). ArXiv preprint arXiv:1411.1830.
82. Maria, C., Boissonnat, J.-D., Glisse, M. & Yvinec, M. International congress on mathematical software. In *The gudhi library: Simplicial complexes and persistent homology* (eds Hong, H. & Yap, C.) 167–174 (Springer, 2014).
83. Morozov, D. Dionysus (2007). (https://www.mrzv.org/software/dionysus/; accessed 2025/02/14 09:10:17).
84. Bouza, J. tda-tools (2018). (https://github.com/jjbouza/tda-tools; accessed 2025/02/14 09:10:17).
85. Bubenik, P. & Dłotko, P. A persistence landscapes toolbox for topological statistics. *J. Symb. Comput.* **78**, 91–114 (2017).
86. Good, P. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics (Springer-Verlag, New York, 2004), third edn.
87. Awad, M. & Khanna, R. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (Apress, 2015).
88. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab - an S4 package for kernel methods in R. *J. Statist. Software* **11**, 1–20. https://doi.org/10.18637/jss.v011.i09 (2004).

Acknowledgements

The authors would like to thank Hector Baños, Alex Elchesen, Afaf Saidi, and Ashleigh Thomas for helpful discussions at the initial stage of this project. The authors also thank the members of the Laboratory for Systems Medicine at the University of Florida for their feedback during manuscript revisions. This research was supported by the NSF-Simons Southeast Center for Mathematics and Biology (SCMB) through the grants National Science Foundation DMS1764406 and Simons Foundation/SFARI 594594.

Author contributions

All authors collaborated in designing research. E.P. conducted the experiments and imaging. E.P. and J.T. preprocessed the experimental data. I.H. and E.P. performed quantitative data analysis. I.H., E.P., and D.A.C. discussed and interpreted the results. I.H. and J.T. implemented the computational pipeline; P.B. and D.A.C. contributed code and commentary. All authors discussed the results and wrote the manuscript.

Declarations

Competing interests

The authors declare no competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-90592-1.

Correspondence and requests for materials should be addressed to D.A.C.