npj | systems biology and applications

www.nature.com/npjsba

ARTICLE OPEN

Check for updates

Topological data analysis of spatial patterning in heterogeneous cell populations: clustering and sorting with varying cell-cell adhesion

Dhananjay Bhaskar[1,2,3,7], William Y. Zhang[3,4,8], Alexandria Volkening[5], Björn Sandstede[3,4] and Ian Y. Wong[1,2,3,6]✉

Different cell types aggregate and sort into hierarchical architectures during the formation of animal tissues. The resulting spatial organization depends (in part) on the strength of adhesion of one cell type to itself relative to other cell types. However, automated and unsupervised classification of these multicellular spatial patterns remains challenging, particularly given their structural diversity and biological variability. Recent developments based on topological data analysis are intriguing to reveal similarities in tissue architecture, but these methods remain computationally expensive. In this article, we show that multicellular patterns organized from two interacting cell types can be efficiently represented through persistence images. Our optimized combination of dimensionality reduction via autoencoders, combined with hierarchical clustering, achieved high classification accuracy for simulations with constant cell numbers. We further demonstrate that persistence images can be normalized to improve classification for simulations with varying cell numbers due to proliferation. Finally, we systematically consider the importance of incorporating different topological features as well as information about each cell type to improve classification accuracy. We envision that topological machine learning based on persistence images will enable versatile and robust classification of complex tissue architectures that occur in development and disease.

npj Systems Biology and Applications (2023)9:43 ; https://doi.org/10.1038/s41540-023-00302-8

INTRODUCTION

Animal tissues are spatially organized into complex spatial patterns by varying adhesive interactions between different cell types[1,2]. For instance, mixtures of motile animal cells in a planar geometry can self-sort into their respective types, as well as aggregate into multicellular clusters[3–21]. Historical work by Steinberg attempted to explain these behaviors using a physical analogy with surface tension, where cells exhibit "differential adhesion" with one another[22], which was subsequently updated by Brodland to include contractility[23]. In particular, two different cell types that each exhibit strong homotypic adhesion (to self) but weak heterotypic adhesion (to the other) would eventually segregate into separate clusters, each consisting of a single cell type[24]. On the other hand, two cell types with strong heterotypic adhesion would randomly intermix within a single cluster. Between these limiting cases, intermediate homotypic and heterotypic adhesion would result in a core–shell organization, where a first cell type would aggregate at the interior, and the second cell type would organize as a spread layer around the periphery. More recent work using *Drosophila melanogaster* has addressed the role of cellular contractility[25] and proliferation[26] in tissue patterning, resulting in distinctive topologies including developmental compartment boundaries[27], hierarchical hexagonal patterning in the retina[28,29], and multicellular rosettes during germband extension[30]. Contractility-based sorting has also been characterized in germ-layer organization during zebrafish gastrulation[31]. "Checkerboard" cellular patterns of sensory hair cells and supporting cells have been observed in the auditory epithelium of the mouse cochlea[32]. More complicated finger-like or labyrinth-like tissue patterning have also been attributed to reaction–diffusion (Turing) mechanisms[33]. Further, Lim and coworkers have demonstrated synthetic cell–cell adhesions that are fully modular and tunable, enabling rational design of multicellular architecture[34]. Given this rich diversity of tissue architectures generated experimentally and in silico, an emerging challenge is to achieve an automated and unbiased classification of distinct spatial patterns[35]. An intriguing possibility is to implement an interpretable and computationally efficient machine learning framework that can be generalized to classify spatial patterns comprised of multiple interacting cell types.

Topological data analysis (TDA) is a promising approach for machine learning of high-dimensional architectures that extracts the "shape" of a dataset based on spatial connectivity[36]. TDA considers how discrete data points may be connected pairwise (dimension 0 homology) or linked into closed loops around an empty region (dimension 1 homology). Persistent homology then treats the stability of these connected structures at varying spatial scales[37]. Topological features can then be represented using a characteristic persistence "barcode" or "diagram," and their relative "similarity" can be determined based on the "cost" of rearranging one diagram to resemble the other[38]. Such topological approaches have recently been used to visualize the spatial organization of a single motile species[39–46] and particulate systems[47]. These past investigations mostly utilized connected components (dimension 0 homology) to analyze populations of fixed size that were near confluency. However, a comparison of two populations with differing size will be biased since the number of connected components is not identical. Our recent

[1]School of Engineering, Brown University, Providence, RI, USA. [2]Center for Biomedical Engineering, Brown University, Providence, RI, USA. [3]Data Science Institute, Brown University, Providence, RI, USA. [4]Division of Applied Mathematics, Brown University, Providence, RI, USA. [5]Department of Mathematics, Purdue University, West Lafayette, IN, USA. [6]Legorreta Cancer Center, Brown University, Providence, RI, USA. [7]Present address: Department of Genetics, Yale School of Medicine, New Haven, CT, USA. [8]Present address: Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉email: ian_wong@brown.edu

work has shown that classification based on closed loops (dimension 1 homology) is a robust approach to classify spatial patterns with varying population size, particularly in the presence of empty regions[48].

Alternatively, persistence images represent topological features based on the weighted sum of Gaussian features, which yields a standardized finite vector representation that is amenable to machine learning[49]. These persistence images can be compressed into a fixed-length numerical representation using dimensionality reduction and manifold learning techniques such as Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP)[50], Potential of Heat Diffusion for Affinity-based Transition Embedding (PHATE)[51] and autoencoder (AE)[52,53]. In principle, unsupervised hierarchical clustering can then be applied to determine similar spatial structures and compared to some ground truth. Nevertheless, the effectiveness of persistence images for classifying spatial patterns comprised of multiple species has not been previously addressed.

In this article, we investigate the use of persistence images for the unsupervised classification of multicellular spatial patterns that emerge from two interacting cell types. We systematically tune homotypic and heterotypic interactions between cell types to generate distinct architectures ranging from dispersed individuals to intermixed clusters to partially or wholly sorted clusters, as well as more complex striped phases, hierarchical hexagonal patterns, and rosettes. The spatial organization of these patterns is then represented by persistence images, persistence curves, or classical order parameters. Further, dimensionality reduction and hierarchical clustering was performed based on dimension 0 and/or dimension 1 homology for each cell type, as well as accounting for both cell types. As a case study, we first considered pattern formation in populations of fixed size. We then considered pattern formation in populations where cells can proliferate with contact inhibition. This work establishes the importance of topological features such as connected components (dimension 0 homology) and closed loops (dimension 1 homology) for classifying spatial patterns, as well as information about each cell type. Altogether, we envision this computational approach will enable new quantitative insights into the emergence of complex tissue architectures via spatiotemporal interactions between multiple cell types.

RESULTS

Sorting and clustering of two nonproliferating cell types with varying adhesion

Differential adhesion simulations were performed using a self-propelled particle model with randomly initialized particle positions and periodic boundaries. Two cell types, labeled in blue (τ_B) and orange (τ_O), were simulated at constant population size ($N = 200$ total, 60% blue and 40% orange) with systematically varying parameters governing blue–blue adhesion (J_{BB}), orange–orange adhesion (J_{OO}), and orange–blue adhesion (J_{OB}), respectively. A random self-propulsion force of constant magnitude, $|\mathbf{P}| = 0.005$, was applied to each particle throughout the simulation. The total simulation time was chosen to exceed the time taken to reach stable steady-state configurations for each set of parameter values (Supplementary Figs. 1 and 2).

In the limit of zero blue–orange adhesion ($J_{BO} = 0$) with weak blue–blue adhesion ($J_{BB} < 0.03$) and weak orange–orange adhesion ($J_{OO} < 0.03$), both orange and blue cells remained individually dispersed (Fig. 1ai and Supplementary Figs. 3 and 4). At stronger blue–blue adhesion ($0.03 < J_{BB}$) and weak orange–orange adhesion ($J_{OO} < 0.03$), blue cells aggregated into clusters while orange cells remained individually dispersed (Fig. 1aii). Conversely, at stronger orange–orange adhesion ($0.03 < J_{OO}$) and weak blue–blue adhesion ($J_{BB} < 0.03$), orange cells aggregated into clusters while blue cells remained individually dispersed (Fig. 1aiii). Finally, at stronger blue–blue adhesion ($0.03 < J_{BB}$) and stronger orange–orange adhesion ($0.03 < J_{OO}$), blue cells aggregated into clusters and orange cells aggregated separately into clusters (Fig. 1aiv).

At slightly increased blue–orange adhesion ($J_{BO} = 0.05$), with comparable blue–blue ($J_{BB} = 0.05$) or orange–orange adhesion ($J_{OO} = 0.05$), sorting into separate blue and orange clusters was again observed (Fig. 1biv and Supplementary Figs. 5 and 6). When blue–blue or orange–orange adhesions were stronger than blue–orange ($J_{BO} \le J_{OO}$, $J_{BO} \le J_{BB}$), a new type of cluster was observed with intermixed orange and blue cells (Fig. 1bv). This organization is reminiscent of "checkerboard" cellular patterns of sensory hair cells and supporting cells have been observed in mouse auditory epithelium of the cochlea[32]. Moreover, for weak blue–blue adhesion ($J_{BB} < 0.03$) and weak orange–orange adhesion ($J_{OO} < 0.05$), blue and orange cells organized into clusters with alternating stripes that were roughly 1–2 cells thick (Fig. 1bvi). These striped phases are somewhat reminiscent of those observed during skin patterning in zebrafish[54]. At stronger blue–blue adhesion ($0.03 < J_{BB}$) and weak orange–orange adhesion ($J_{OO} < 0.03$), clusters consisted of orange cells dispersed in a hexagonal configuration, surrounded by blue cells. (Fig. 1bvii). Conversely, at stronger orange–orange adhesion ($0.03 < J_{OO}$) and weak blue–blue adhesion ($J_{BB} \approx 0$), clusters consisted of blue cells dispersed in a hexagonal configuration, surrounded by orange cells. (Fig. 1bviii). Interestingly, this hierarchical patterning has a superficial resemblance to cone cells in the *Drosophila* retina[28]. At stronger orange–orange adhesion ($0.03 < J_{OO}$) and slightly stronger blue–blue adhesion ($J_{BB} \approx 0.01$), clusters consisted of a core of orange cells surrounded by blue cells at the periphery (Fig. 1bix).

At intermediate blue–orange adhesion ($J_{BO} = 0.13$), with stronger blue–blue or orange–orange adhesions ($J_{BO} \le J_{OO}$, $J_{BO} \le J_{BB}$), clusters again were comprised of intermixed orange and blue cells (Fig. 1cv and Supplementary Figs. 7 and 8). For weaker blue–blue or orange–orange adhesions, ($J_{OO} \approx J_{BB} \le J_{BO}$), clusters with alternating orange and blue stripes were observed (Fig. 1cvi). At stronger blue–blue adhesion ($0.03 < J_{BB}$) and weak orange–orange adhesion ($J_{OO} < 0.03$), clusters consisted of orange cells dispersed in a hexagonal configuration, surrounded by tightly packed blue cells (Fig. 1cxi). When blue–blue adhesion and orange–orange adhesion were roughly comparable ($J_{BB} \approx J_{OO} \approx 0.09$), clusters consisted of many tightly packed orange cells dispersed in a hexagonal configuration, surrounded by tightly packed blue cells (Fig. 1cx). At the strongest blue–blue adhesion ($J_{BB} \approx 0.20$) and weak orange–orange adhesion ($J_{OO} < 0.03$), spots of orange cells were again dispersed in a hexagonal configuration, surrounded by slightly sparser blue cells (Fig. 1cvii). At stronger orange–orange adhesion ($0.03 < J_{OO}$) and weak blue–blue adhesion ($J_{BB} \approx 0$), clusters again consisted of blue cells dispersed in a hexagonal configuration, surrounded by orange cells. (Fig. 1cvii). When orange–orange adhesion was comparable to blue–orange adhesion ($J_{OO} \approx J_{B0} = 0.13$) with weaker blue–blue adhesion ($J_{BB} = 0.07$), clusters were observed with finger-like or labyrinth-like patterns of orange and blue cells (Fig. 1cxii), analogous to those generated by reaction–diffusion (Turing) mechanisms[33]. At strong orange–orange adhesion ($J_{OO} \approx 0.2$) with weak blue–blue adhesion ($J_{BB} \in [0.03, 0.11]$), clusters consisted of a core of orange cells surrounded by blue cells at the periphery (Fig. 1cix). Notably, a number of cluster configurations observed at $J_{BO} = 0.13$ (Fig. 1cv–ix) were qualitatively similar to those previously observed at $J_{BO} = 0.05$ (Fig. 1bv–ix), but offset to higher values of J_{BB} or J_{OO}.

At strong blue–orange adhesion ($J_{BO} = 0.25$) with comparable blue–blue and orange–orange adhesion ($J_{BB} \approx J_{OO}$), blue and orange cells organized into clusters with alternating stripes that were roughly 1–2 cells thick (Fig. 1dvi and Supplementary Figs. 9 and 10). At stronger blue–blue adhesion ($0.03 < J_{BB}$) and weak orange–orange adhesion ($J_{OO} < 0.03$), clusters again consisted of spots of orange cells dispersed in a hexagonal configuration,

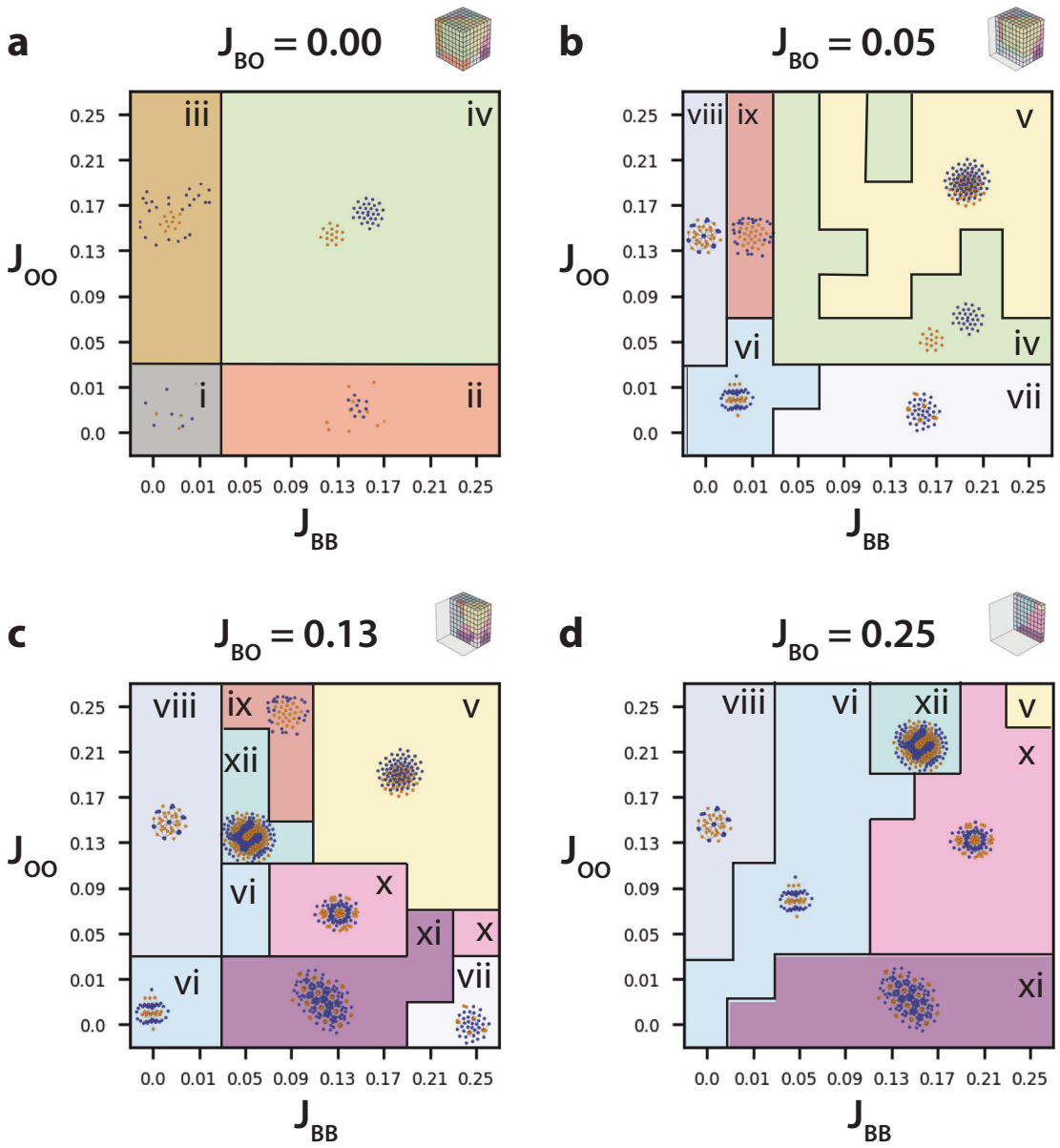

Fig. 1 Comparison of cluster and stripe patterning for two cell types (orange and blue) at constant population size with systematically varying blue–blue adhesion (J_{BB}), orange–orange adhesion (J_{OO}), and blue–orange adhesion (J_{BO}), respectively. Representative slices with J_{BB} vs J_{OO} for **a** $J_{BO} = 0.0$, **b** $J_{BO} = 0.05$, **c** $J_{BO} = 0.13$, and **d** $J_{BO} = 0.25$.

surrounded by tightly packed blue cells (Fig. 1cxi). Conversely, at stronger orange–orange adhesion ($0.03 < J_{OO}$) and weak blue–blue adhesion ($J_{BB} \approx 0$), clusters again consisted of blue cells dispersed in a hexagonal configuration, surrounded by orange cells. (Fig. 1dviii). When orange–orange adhesion was comparable to blue–orange adhesion ($J_{OO} \approx J_{BO} = 0.25$) with weaker blue–blue adhesion ($J_{BB} = 0.15$), clusters were observed with finger-like patterns of orange and blue cells (Fig. 1dxii).

When blue–blue adhesion and orange–orange adhesion were roughly comparable ($J_{BB} \approx J_{OO} \approx 0.13$), clusters consisted of tightly packed orange cells dispersed in a hexagonal configuration, surrounded by tightly packed blue cells (Fig. 1dx). When all three adhesions were comparable ($J_{BO} \approx J_{BB} \approx J_{OO} = 0.25$), clusters were observed with intermixed orange and blue cells (Fig. 1dv). Next, we performed unsupervised classification on the multicellular patterns for varying adhesion parameters (Fig. 1) using persistence images (Supplementary Fig. 11), normalized persistence curves (Supplementary Fig. 12), and order parameters (Supplementary Fig. 13).

Unsupervised classification of two nonproliferating cell types with varying adhesion

Unsupervised classification of multicellular patterns using PCA, PHATE, AE, and UMAP was compared based on our ground truth labeling, and also colored by increasing values of J_{BO}, J_{OO}, and J_{BB} (Fig. 2). Ground truth labels (12 in total, denoted i-xii) were assigned based on manual inspection of particle configurations at the end of the simulation. For ease of visualization, axes were plotted so that the spatial configurations are more qualitatively consistent, as noted in the figure. In general, patterns with individually dispersed cells (e.g., i, ii, iii) were classified farther away from other patterns where both cell types were clustered (Fig. 2). Further, for the crescent-like grouping of the remaining clustered cell patterns, J_{BO} increased from bottom to top, J_{OO}

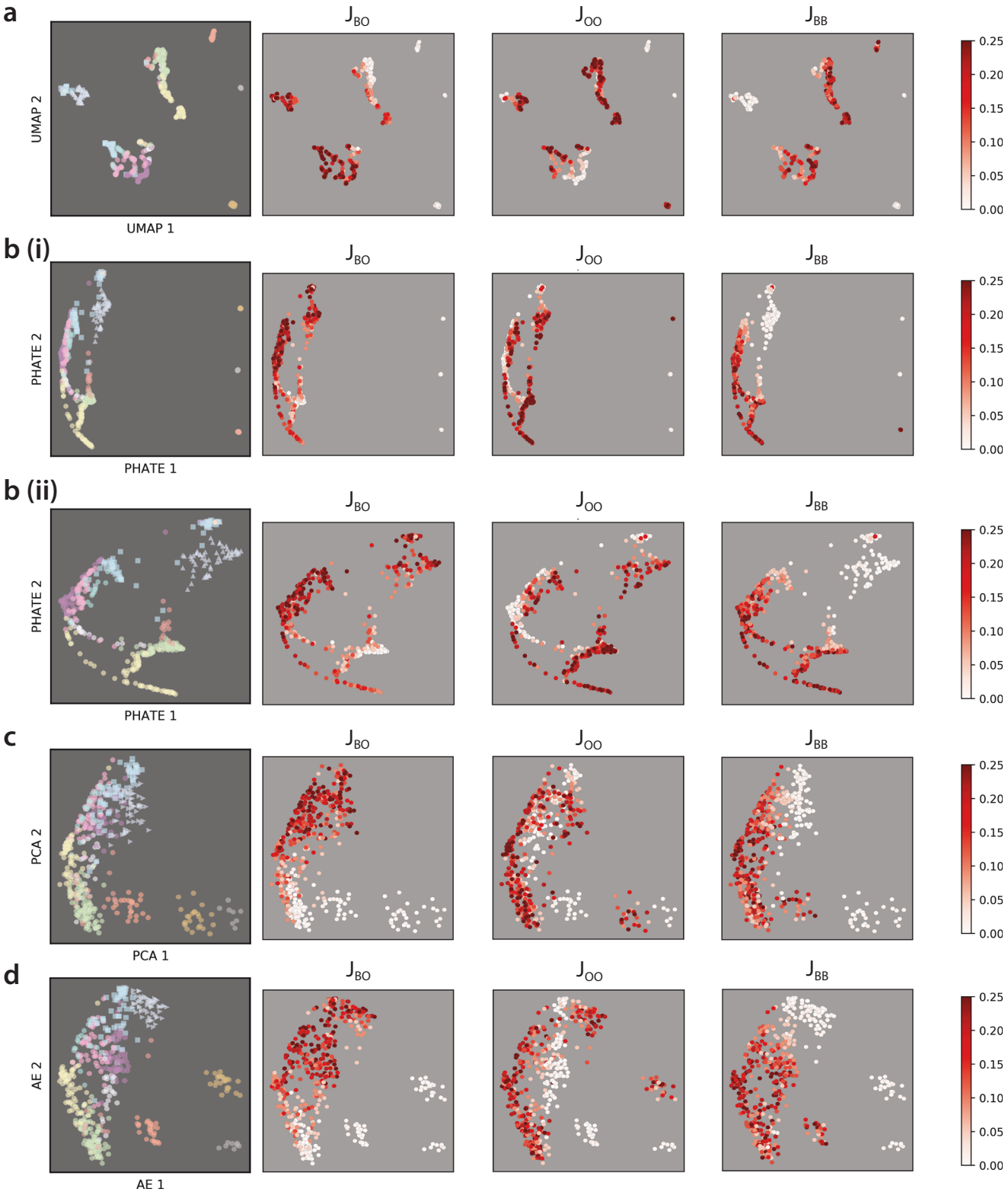

Fig. 2 2D embeddings of persistence images colored by ground truth and adhesion values for simulations with constant population size. 2D embeddings obtained using **a** UMAP, **b** PHATE with all simulations (i) and zoomed in to only simulations with clusters (ii), **c** PCA, and **d** autoencoder.

increased from the inside out, and J_{BB} increased from top to bottom. Proceeding counterclockwise from the top right, the top arm of the crescent represented high J_{BO}, low J_{OO}, and low J_{BB}, which included striped configurations (vi, sky blue; xii, turquoise) and hexagonal configurations of blue cells (viii, gray blue). Further, the center of the crescent represented high J_{BO}, low J_{OO}, and high J_{BB}, which included hexagonal configurations of orange cells (vii, gray; x, fuchsia; xi, purple). Finally, the bottom arm of the crescent

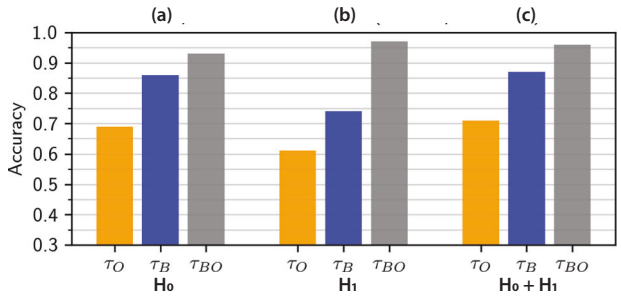

Fig. 3 Unsupervised classification accuracy for persistence images at constant population size. Unsupervised classification of simulations by hierarchical clustering of 20-dimensional autoencoder embeddings of persistence images. Classification accuracy is computed by comparing cluster labels to ground truth for **a** dimension 0 (H_0), **b** dimension 1 (H_1), and **c** dimension 0 and dimension 1 ($H_0 + H_1$) homology.

represented low J_{BO}, high J_{OO} and high J_{BB}, which included sorted clusters (iv, green), partially sorted clusters (v, yellow), and orange clusters surrounded by blue cells (ix, brick red).

The accuracy of unsupervised AE classification (Fig. 2d) relative to our manually annotated ground truth (Fig. 1) was determined based on persistence images that considered only dimension 0 homology, only dimension 1 homology or both dimension 0 and dimension 1 (Fig. 3 and Supplementary Fig. 14). Further, unsupervised classification was performed using persistence images that included information on both blue and orange cells, blue cells only, and orange cells only (Fig. 3).

Classification accuracy based on persistence images using only dimension 0 homology was moderately successful with only one cell type, ranging from 69% for orange cells only, to 86% for blue cells only, but increasing to 93% for blue and orange cells combined (Fig. 3a). In comparison, classification accuracy using only dimension 1 homology was worse when using only orange cells (61%) and blue cells (74%), but considerably improved when considering both orange and blue cells (97%) (Fig. 3b). Finally, classification accuracy using both dimension 0 and dimension 1 homology generally outperformed dimension 0 and dimension 1 only, ranging from 71% for orange cells only, up to 87% for blue cells only, and then 96% for both orange and blue cells (Fig. 3c). Overall, classification accuracy was considerably better when considering blue cells only relative to orange cells only, which could be explained by the 60:40 ratio of blue cells to orange cells in the simulations. Interestingly, combining information from both blue and orange cells incrementally improved classification accuracy for dimension 0 homology only (Fig. 3a), but resulted in much larger improvements for dimension 1 (Fig. 3b) only as well as dimension 0 and dimension 1 (Fig. 3c). For comparison, classification using persistence curves was slightly worse, ranging from 62 to 83% (Supplementary Figs. 15a–c and 16). Classification using radial order parameters was comparable to persistence images, ranging from 68 to 94% (Supplementary Figs. 17a–c and 18), but considerably worse for angular order parameters. Overall, the best classification occurred with persistence images and AE, considering both blue and orange cells for dimension 1 homology only (97%), and comparable performance for dimension 0 and dimension 1 homology (96%) (Supplementary Tables 1–3).

Sorting and clustering of two proliferating cell types with varying adhesion

Next, these differential adhesion simulations were implemented with proliferation, so that a mother cell would divide into two daughter cells after some duration (e.g., 50,000 time steps), randomly offset. Cell division events were implemented so that one daughter cell retained the velocity and direction of the mother cell, while the other daughter cell was placed nearby, but moving at equal velocity in the opposite direction. Moreover, a contact inhibition rule was included so that a cell was not permitted to divide if the local cell density was high (more than four nearest neighbors). Simulations were otherwise performed consistently with the previous scenario, starting with a 60:40 ratio of blue and orange cells while systematically varying blue–blue adhesion (J_{BB}), orange–orange adhesion (J_{OO}), and blue–orange adhesion (J_{BO}). Again, the total simulation time was chosen to exceed the time taken to reach stable steady-state configurations for each set of parameter values (Supplementary Figs. 19 and 20). For most of the representative simulations, the steady-state configuration was reached by 8–10 cell cycles (15% of the overall simulation duration), although certain scenarios involving the merging of clusters required up to 48 cell cycles (76% of the overall simulation duration). Nevertheless, the multicellular stripe or spot organization within these clusters was typically evident by ~10 cell cycles. This scenario with both proliferation and sorting was inspired by tissue repair, such as the regeneration of zebrafish skin patterns after laser ablation[55].

For weak blue–orange adhesion ($J_{BO} = 0.05$) with roughly comparable blue–blue and orange–orange adhesion ($J_{BB} \approx J_{OO}$), clusters appeared intermixed with irregular domain sizes (Fig. 4av and Supplementary Fig. 21). We utilized the same numbering convention as in Fig. 1 for ease of comparison. Nevertheless, for weak blue–blue adhesions ($J_{BB} = 0.05$) and varying orange–orange adhesions ($0.07 \le J_{OO}$), clusters were partially sorted with more blue cells than orange cells (Fig. 4axiv), which was not previously observed. By analogy, for weak orange–orange adhesions ($J_{OO} = 0.05$) and varying blue–blue adhesions ($0.07 \ge J_{BB}$), clusters were partially sorted with more orange cells than blue cells (Fig. 4axv).

At slightly increased blue–orange adhesion ($J_{BO} = 0.09$), with roughly comparable blue–blue and orange–orange adhesion ($0.09 \le J_{BB} \approx J_{OO}$), clusters again appeared intermixed with irregular domain sizes (Fig. 4bv and Supplementary Fig. 22). In comparison, for weak blue–blue and orange–orange adhesion ($J_{BB} \approx J_{OO} \le 0.09$), clusters with alternating orange and blue stripes were observed (Fig. 4bvi). For slightly increased blue–blue or orange–orange adhesions ($J_{OO} = 0.13, J_{BB} = 0.13$), clusters exhibited finger or labyrinth-like morphology (Fig. 4bxii). For strong blue–blue adhesions ($0.17 \le J_{BB}$) and weak orange–orange adhesions ($J_{OO} = 0.05$), clusters were sorted with blue cells in the interior and orange cells at the periphery (Fig. 4bvii). In comparison, for strong orange–orange adhesions ($0.17 \le J_{OO}$) and weak blue–blue adhesions ($J_{BB} = 0.05$), clusters were sorted with orange cells in the interior and blue cells at the periphery (Fig. 4bix).

At intermediate blue–orange adhesion ($J_{BO} = 0.13$), with roughly comparable blue–blue and orange–orange adhesion ($0.13 \le J_{BB} \approx J_{OO}$), clusters continued to be intermixed with irregular domain sizes (Fig. 4cv and Supplementary Figs. 23 and 24). For weak orange–orange adhesions ($J_{OO} < 0.13$), clusters exhibited hexagonally arrayed spots of packed orange cells surrounded by blue cells (Fig. 4cx). Instead, for weak blue–blue adhesions with comparable orange–orange and blue–orange adhesions ($J_{BB} < J_{OO} \approx J_{BO} = 0.13$), clusters exhibited a stripe phase (Fig. 4cvi). For weak blue–blue adhesions ($J_{BB} < 0.13$) with strong orange–orange adhesions ($J_{OO} \approx 0.21$), clusters exhibited a labyrinth or finger-like morphology (Fig. 4cxii). Finally, for slightly increased blue–blue adhesion ($J_{BB} = 0.09, J_{OO} = 0.25$), clusters exhibited a core of orange cells with peripheral blue cells (Fig. 4cix).

At the strongest blue–orange adhesion ($J_{BO} = 0.25$), intermixed clusters were only observed for comparably strong blue–blue and orange–orange adhesions ($J_{BB} \approx J_{OO} = 0.25$) (Fig. 4dv and Supplementary Figs. 25 and 26). When blue–blue and orange–orange adhesions were weaker but comparable ($J_{BB} \approx J_{OO} \le 0.21$), clusters with alternating orange and blue stripes were observed (Fig. 4dvi). At one location ($J_{BB} = 0.09, J_{OO} = 0.05$),

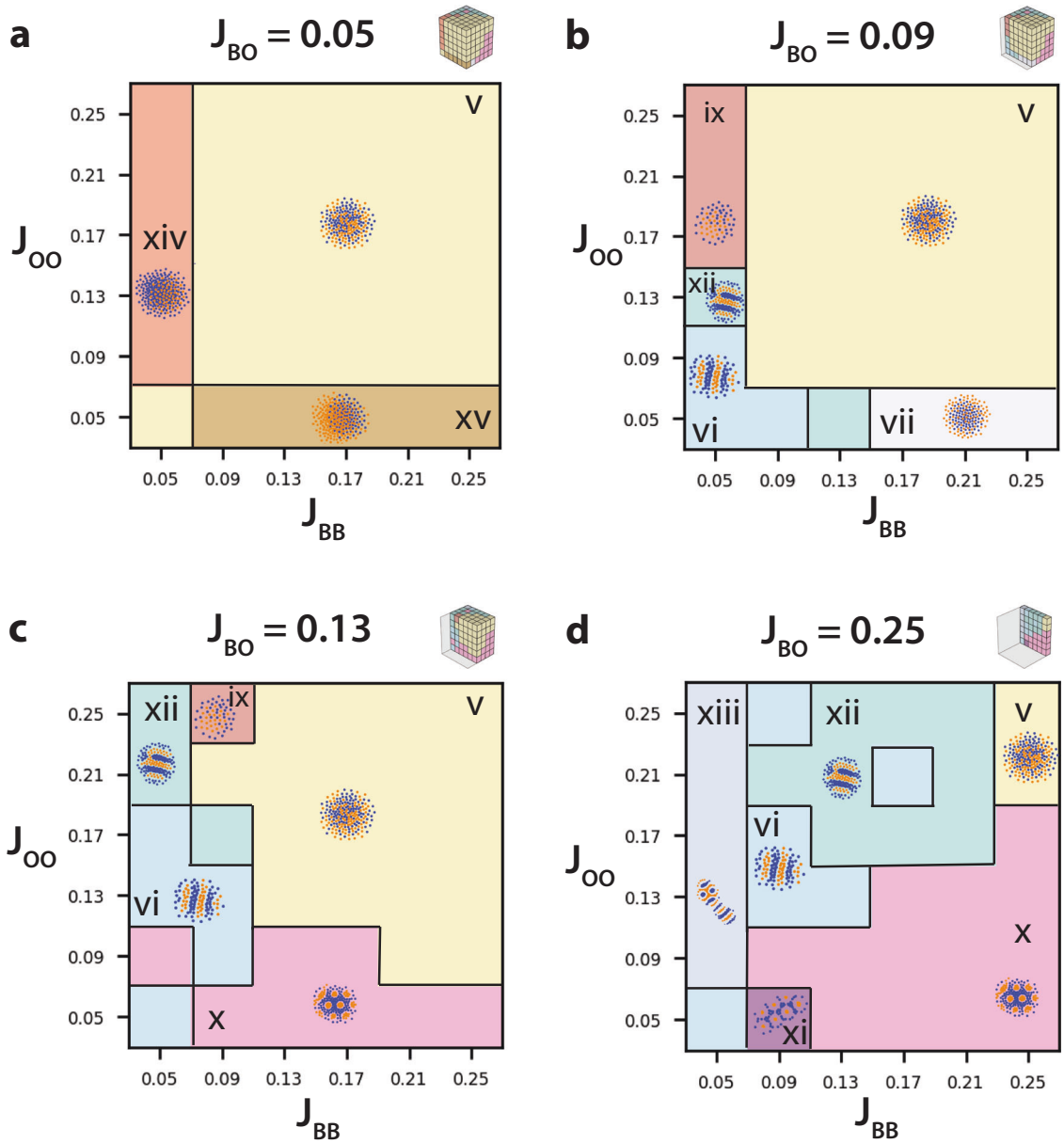

Fig. 4 Comparison of cluster and stripe patterning for two cell types (orange, blue) at varying population sizes with systematically varying blue–blue adhesion (J_{BB}), orange–orange adhesion (J_{OO}), and blue–orange adhesion (J_{BO}), respectively. Representative slices with J_{BB} vs J_{OO} for **a** $J_{BO} = 0.05$, **b** $J_{BO} = 0.09$, **c** $J_{BO} = 0.13$, **d** $J_{BO} = 0.25$.

clusters consisted of spots of orange cells dispersed in a hexagonal configuration, surrounded by tightly packed blue cells (Fig. 4dxi). For stronger blue–blue adhesions relative to orange–orange adhesions ($J_{OO} < J_{BB} \leq 0.17$, clusters consisted of circular spots of orange cells surrounded by blue cells (Fig. 4dx). For weak blue–blue adhesions ($J_{BB} = 0.05$), a new cluster morphology was observed with mixed stripes and spots (Fig. 4dxiii). Finally, for stronger orange–orange adhesions relative to blue–blue adhesions ($J_{BB} < J_{OO} \leq 0.25$), clusters exhibited a labyrinth or finger-like morphology (Fig. 4dxii).

Unsupervised classification of two proliferating cell types with varying adhesion

Unsupervised classification of these multicellular patterns with proliferation was implemented using PCA, PHATE, AE, and UMAP for comparison with ground truth labeling, then plotted by increasing values of J_{BB}, J_{OO}, and J_{BO} (Fig. 5). Ground truth labels (10 in total, denoted v-vii, ix-xv) were assigned based on manual inspection of the final configurations. For ease of visualization, axes were again plotted so that the spatial configurations are more qualitatively consistent. In general, it was apparent that similarly classified conditions were more widely dispersed after dimensionality reduction for these simulations with proliferation relative to no proliferation (Fig. 2), which can be attributed in part to differences in total cell numbers. Since cells continued to proliferate until contact-inhibited "steady state", no simulations were observed with individually dispersed cells.

After 2D embedding, UMAP and PHATE yielded roughly similar distributions (Fig. 5a, b), while PCA and AE also were comparable (Fig. 5c, d). Nevertheless, conditions with different colors were poorly separated by UMAP and PHATE relative to PCA and AE (e.g., v, yellow and xv, tan). Roughly, J_{BO} increased from left to right, J_{OO} increased inward from the periphery, and J_{BB} increased outwards from the interior (Fig. 5). Proceeding from the left, the leftmost

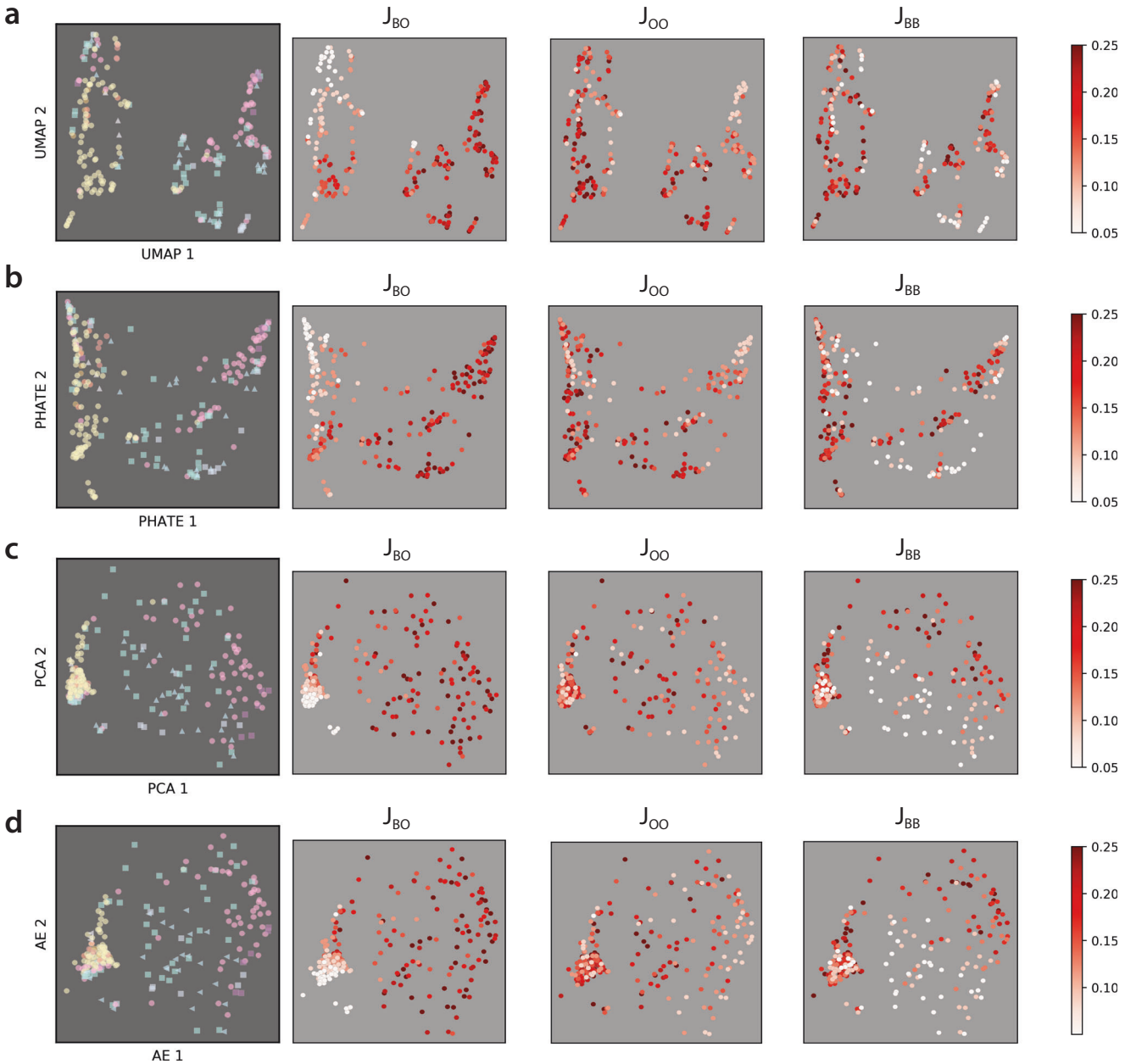

Fig. 5 2D embeddings of persistence images colored by ground truth and adhesion values for simulations with varying population size. 2D embeddings were obtained using **a** UMAP, **b** PHATE, **c** PCA, and **d** autoencoder.

group represented low J_{BO}, high J_{OO} and intermediate J_{BB}, included intermixed clusters (v, yellow), partially sorted clusters (xiv, brick red; xv, tan), and clusters with orange cells at the interior and blue cells at the periphery (ix, coral). Further, the middle grouping of cells represented high J_{BO}, high J_{OO}, and low J_{BB} which included striped patterns (vi, sky blue; xii, turquoise; xiii, gray blue). Finally, the rightmost grouping represented high J_{BO}, low J_{OO}, and high J_{BB}, which included spots of orange cells in a hexagonal configuration (x, fuchsia and xi, purple) (Fig. 2).

For these simulations with contact-inhibited proliferation, the accuracy of unsupervised AE classification (Fig. 5) relative to our manually annotated ground truth (Fig. 4) was again determined based on persistence images that considered only dimension 0 homology, only dimension 1 homology or both dimension 0 and dimension 1 (Supplementary Fig. 27). Further, unsupervised classification was performed using persistence images that included information on both blue and orange cells, blue cells only, and orange cells only. Notably, unsupervised AE classification for simulations with proliferation was considerably worse compared to simulations with constant cell number, particularly for dimension 1 homology or dimension 0 and 1 homology, which ranged from 37 to 78% (Fig. 6). Qualitatively similar trends were also observed where classification accuracy tended to be worse when considering orange cells only, with some improvement for blue cells, as well as orange and blue cells, respectively.

Persistence diagrams and images may be biased by different point cloud sizes, corresponding here to different numbers of cells, which likely affected the unsupervised classification. Thus, we normalized the persistence images across all simulations by dividing each persistence image by its maximum intensity. As a consequence, classification accuracy for normalized images improved considerably by 5–10%. For example, classification accuracy based on normalized persistence images improved to 61–84% for dimension 0, to 56–81% for dimension 1, and to

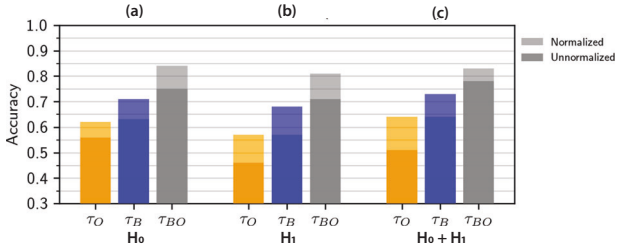

Fig. 6 Unsupervised classification accuracy for persistence images at varying population size. Unsupervised classification by hierarchical clustering of 20-dimensional autoencoder embedding of persistence images for **a** dimension 0 (H_0), **b** dimension 1 (H_1), and **c** dimension 0 and dimension 1 ($H_0 + H_1$) homology. Darker bars represent persistence images without normalization, whereas lighter bars represent the additional improvement after normalization by peak intensity per image. Classification accuracy of images with and without normalization is computed by comparing cluster labels to ground truth.

63–83% for dimensions 0 and 1, respectively (Fig. 6). The general trend of improving classification from orange only to blue only to orange and blue remained consistent with normalization. For comparison, classification using persistence curves was comparable, ranging from 49% to 84% (Supplementary Figs. 15d–f and 28). Classification using radial order parameters was also comparable to persistence images, ranging from 66% to 88% (Supplementary Figs. 17d–f and 29), but considerably worse for angular order parameters. Further, the classification of persistence curves and order parameters using PHATE was also associated with poorly separated groupings (Supplementary Fig. 30). Overall, dimension 0 only classification tended to outperform dimension 1 only classification, and dimension 0 and dimension 1 together tended to give the best classification. Moreover, accounting for both orange and blue cells tended to give comparable classification for persistence images, persistence curves, and order parameters (Supplementary Tables 4–6).

DISCUSSION

In this work, we investigated how two interacting and motile cell types ("blue" and "orange") self-organize into multicellular patterns when the strength of blue–blue, blue–orange, or orange–orange adhesions are varied systematically. Based on the relative positions of blue cells with respect to other blue cells and orange cells, as well as orange cells to other orange cells, these patterns could be represented using persistence images, persistence curves, and classical order parameters, then analyzed using dimensionality reduction and hierarchical clustering. After optimization, unsupervised classification showed excellent agreement with our manually annotated ground truth (85–95%) (Figs. 3 and 6). Moreover, this classification is in good agreement with Steinberg's scaling arguments for differential adhesion[24]. For instance, the classifier reveals distinct regimes where both cell types are intermixed, which occurs when homotypic adhesions are weaker than heterotypic adhesion ($J_{BB}, J_{OO} < J_{OB}$, Fig. 1v). Conversely, cells are sorted apart when heterotypic adhesions are much weaker than homotypic adhesion, ($J_{OB} << J_{BB}, J_{OO}$, Fig. 1iv). We also classify a pattern where a core of orange cells surrounded by a shell of blue cells for mismatched homotypic adhesion with intermediate heterotypic adhesion ($J_{BB} < J_{OB} < J_{BB}$, Fig. 1ix). Our computational analysis is more granular, identifying distinct patterns driven by more subtle differences in homotypic and heterotypic adhesion. For example, we observe discrete clusters of blue or orange cells of varying size arranged in checkerboard, stripe, or labyrinth patterns (Fig. 1vi–xii). These patterns are very robust for identical adhesion parameters with different initial conditions (Supplementary Fig. S31). More recent conceptual models are based on interfacial tension, which address both cell–cell adhesion as well as cortical tension[1]. Our classified patterns are also in qualitative agreement with scaling arguments for interfacial tension (see refs. [2,23]), noting that interfacial tension decreases with increasing adhesion. However, we note that Brodland's work uses a vertex-based model of cell shape, which yields certain numerical prefactors for the scaling argument. Thus, there are some quantitative differences with our agent-based model which does not consider interfacial tension and polygonal cell shape.

Based on manual inspection of the remaining discrepancies (5–15%), we recognized that some conditions were located at the "phase boundary" between different regions, and could exhibit a mixture of different spatial patterns. Indeed, misclassified conditions were often far from the centroid of each grouping, which was also apparent from the 2D AE embedding (Supplementary Figs. 32–37). Although some caution is warranted when visually interpreting dimensionality-reduced embeddings, we note that different colored groupings were relatively well separated by AE, particularly for the simulations with proliferation (Fig. 5d). In comparison, UMAP biased towards discrete clusters with some loss of global structure, while PHATE "squeezed" data points together into continuous branch-like structures[51]. Thus, for simulations with proliferation, different colors are widely dispersed without clean separation in UMAP and PHATE embedding, including intermixed (v, yellow) and partially mixed particles (xv, brown), as well as stripes (vi, light blue) and spots (x, pink) (Fig. 5a, b), which likely contributed to their markedly worse performance for classifying simulations with proliferation. For comparison, we have included supervised classification results obtained using a "softer" probabilistic algorithm where conditions can be classified by multiple adjacent regions. Briefly, we used a soft margin support vector machine (SVM), using the radial basis function for nonlinear transformation of the input, at various values of C and computed the accuracy using fivefold cross-validation, which also exhibits excellent performance (Supplementary Tables 7 and 8).

Moreover, the classification was occasionally confounded by multicellular patterns with similar spatial organization but where the positions of the "blue" and "orange" cells were switched. For instance, consider a scenario where hexagonal arrays of orange clusters are surrounded by blue cells occur at high J_{BB} and low J_{OO}, (Fig. 1bvii). Switching the relative positions of blue and orange cells in this scenario results in hexagonal arrays of blue clusters surrounded by orange cells at low J_{BB} and high J_{OO} (Fig. 1bviii). The input to the classifier consists of feature vectors representing the persistence images of blue, orange, or blue and orange cells, which are weighted equally without explicitly specifying a "cell type". Thus, from the perspective of the unsupervised classifier, these two configurations appeared indistinguishable, especially given that the effective blue and orange cell sizes were identical based on the interaction potential (Eq. (3)). Analogous issues occurred for unsupervised classification using only one cell type (e.g., orange) without consideration of the other cell type (e.g., blue). For example, if only orange cells were considered, hexagonal arrays of blue clusters surrounded by orange cells appeared relatively similar to clusters with blue cells at the interior and concentric rings of orange and blue, particularly in dimension 1 homology that only considers topological loops (Supplementary Fig. 38a). Instead, if only blue cells were considered (and dimension 1 homology), hexagonal arrays of orange clusters would appear very similar to hexagonal arrays of orange individuals (Supplementary Fig. 38b). Similarly if the spatial configuration of orange cells was comparable in scenarios with very different blue cell configurations, consideration of only orange cells in dimension 0 homology would also result in misclassification (Supplementary Fig. 39b). The use of feature vectors representing spatial information about both blue and

orange cells typically yielded improved classification. Nevertheless, classification using dimension 0 or dimension 1 homology usually resulted in comparable accuracy (using both colors), without significant improvement when both dimension 0 and dimension 1 were considered.

These simulations were initialized with a 60:40 ratio of blue to orange cells, which indirectly encodes a difference between cell types into the classifier. To check the robustness of these results, we repeated these simulations with varying domain sizes and verified that multicellular patterns occurred consistently, particularly the organization of cells within clusters (Supplementary Figs. 40 and 41). We note that classification accuracy remains consistently high across different domain sizes, both for constant population size as well as with proliferation (Supplementary Fig. 42). We also varied the cell type ratio from 90:10 to 10:90 over a comparable range of homotypic and heterotypic adhesion strengths (Supplementary Fig. 43). Notably, increasing the fraction of one cell type (e.g., blue) relative to the other (e.g., orange) was equivalent to increasing the relative homotypic adhesion (e.g., strengthening blue–blue adhesion relative to orange–orange). For example, a 90:10 ratio of blue:orange cells typically resulted in sparse orange cells surrounded by blue cells (Fig. 1d–h). In comparison, a 10:90 ratio of blue:orange cells exhibited sparse blue cells surrounded by orange cells, also equivalent to switching the positions of blue and orange cells. Moreover, cells that could undergo contact-inhibited proliferation exhibited additional patterns with both stripes and spots (Fig. 4cxiii), as well as partially sorted patterns (Fig. 4cxiv, xv). We further verified that varying the proliferation rate with contact inhibition resulted in qualitatively similar spatial patterns when homotypic and heterotypic adhesions were held constant (Supplementary Fig. 44). Overall, these artifacts of interchanging cells are unlikely to arise under more realistic conditions, since different cell types within a tissue exhibit appreciable differences in size or biomarker expression that would further inform unsupervised classification. We further considered simulations with three interacting cell types (50% blue, 30% orange, 20% green), which resulted in analogous pattern formation as with two interacting cell types (Supplementary Figs. 45–47). This algorithm also demonstrated excellent classification of these patterns based on the positions of two out of the three cell types, especially with blue and orange cells which comprise 80% of the population (Supplementary Fig. 48). We also classified blue–green and orange–green pairings, using the blue–orange ground truth phase classification as a reference. This resulted in a few misclassifications, primarily resulting from overlapping cell positions or blue and green cells (Supplementary Fig. 46d), orange and green cells (Supplementary Fig. 46e), as well as previously unseen spatial arrangements that emerged due to competition between two cell types for maximizing interaction with the third cell type (Supplementary Fig. 46f).

Although self-organization into multicellular patterns occurs relatively robustly in development, the motility and interactions of individual cells are stochastic. Thus, an unsupervised classifier must be stable against variability that arises from biological "noise". The transformation of coordinates to persistence diagrams in topological data analysis is provably stable with respect to bottleneck distance and Wasserstein distance[37]. However, direct comparison of cell positions via bottleneck distance is undesirable, since the L_∞ norm is entirely determined by a single topological feature. Comparisons based on the Wasserstein metric are computationally expensive, requiring the calculation of an optimal transport plan between pairs of persistence diagrams. Here, we have focused on persistence images as an underexplored approach for topological data analysis of multicellular pattern formation, relative to classical order parameters and persistence curves. Persistence images enable excellent unsupervised classification at intermediate computational cost with theoretical guarantees on stability. The Euclidean distance between persistence images based on the 2D Gaussian kernel is bounded by the 1-Wasserstein distance between the corresponding persistence diagrams[49]. The kernel bandwidth and image resolution can be adjusted to achieve the desired trade-off between computational efficiency and stability guarantees. In comparison, persistence curves can be computed more efficiently but lead to worse classification. The loss of information by summing over the Gaussian kernel along the diagonal is offset by the greater computational efficiency of using a lookup table of cumulative distribution values in the calculation. In comparison, classical order parameters perform very well for classification, but are quite expensive computationally. One explanation for the relative success of classical order parameters, especially for simulations with proliferation, is that they have been normalized to account for different numbers of (identical) particles. In comparison, the optimum normalization for persistence images remains unresolved, and will be considered more thoroughly in a follow-up manuscript.

In conclusion, we show that combining persistence images with autoencoders enables the unsupervised, computationally efficient classification of spatial patterns associated with two interacting cell types. This approach represents topological features of the multicellular architecture as the weighted sum of Gaussian features, yielding a standardized finite vector representation. We show that optimized dimensionality reduction using AE and hierarchical clustering can reveal topologically similar simulation conditions, in excellent agreement with our manually annotated ground truth, particularly for populations of fixed size. However, persistence images of simulations with varying population size required normalization for unbiased comparison, and performed slightly worse. In a follow-up manuscript, we will apply this approach to analyze stripe and spot patterning in zebrafish development, as well as to gain deeper mechanistic insight into persistence images. Overall, we envision that topology-based machine learning represents a powerful and human-interpretable framework to explore the diversity of complex tissue shapes and patterns that emerge from self-sorting and collective cell migration.

METHODS

We investigated how two different types of interacting cells (discrete agents) self-organize into multicellular patterns as cell–cell adhesion was varied (Fig. 7a). For simplicity, cells were defined as either "blue" or "orange," and a total of 512 different combinations with varying adhesion between blue–blue, orange–orange, or blue–orange cell types were simulated at constant population size (i.e., no proliferation) (Fig. 7ai). These simulations were then implemented for another 343 different conditions (3 replicates each) where cells were permitted to proliferate when there was unoccupied space nearby (i.e., contact-inhibited proliferation). The equations of motion for each cell was solved for self-propulsion with random reorientation, until cells reached some "steady-state" configuration after 5,000,000 time steps (Fig. 7aii). These varying multicellular patterns were subsequently considered for classification.

Steady-state multicellular patterns were then manually classified to define a "ground truth" (Fig. 7bi). These patterns were further converted to persistence images, which can also be represented as feature descriptor vectors (Fig. 7bii). For comparison, these patterns were also converted to feature vectors based on conventional order parameters that represent radial or angular symmetries of blue, orange, or both sets of cells (Fig. 7biii). Finally, these multicellular patterns were converted to normalized persistence curves. Unsupervised classification of these feature

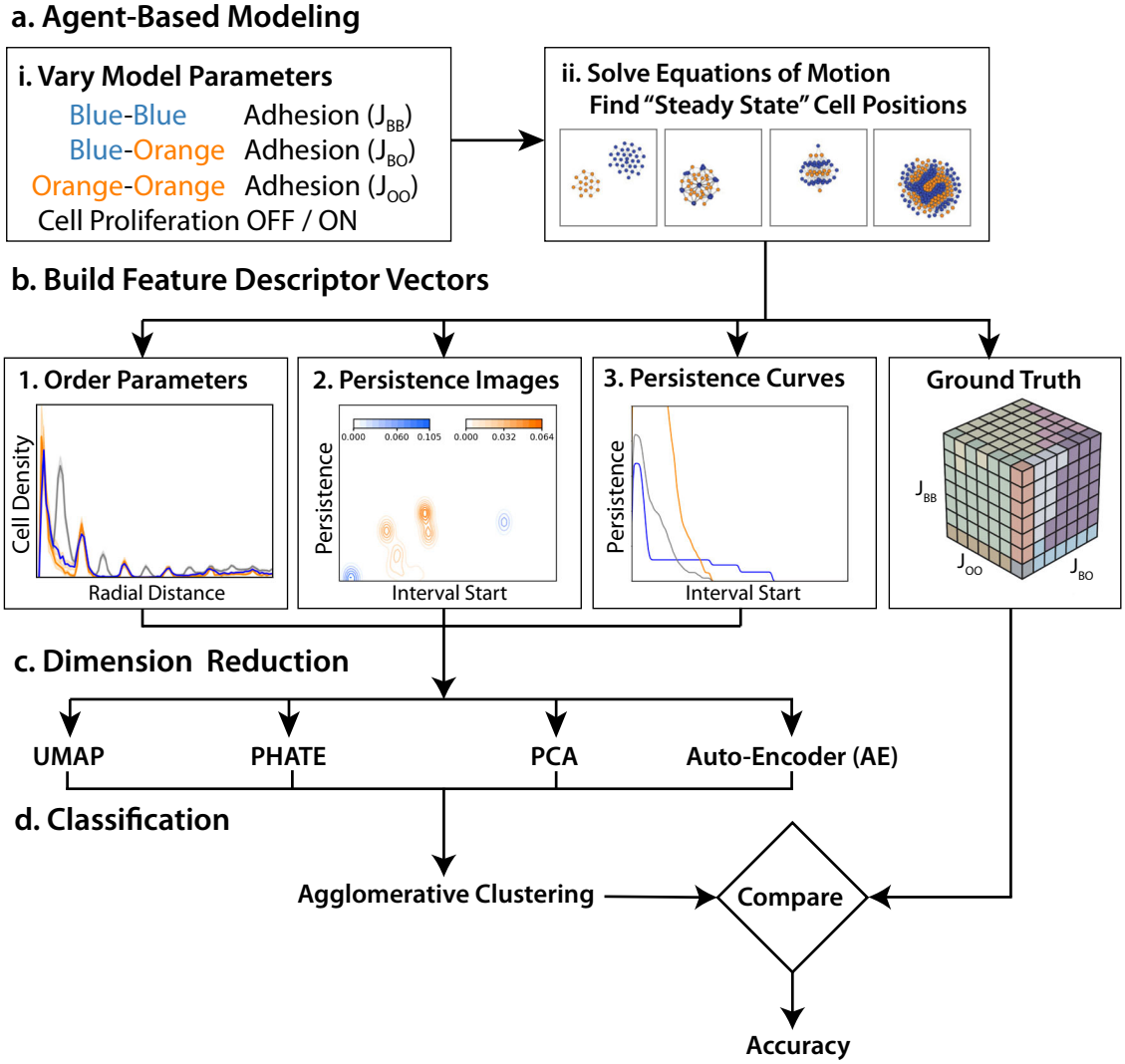

Fig. 7 Flow diagram of the proposed methodology for unsupervised classification of cellular patterns in the coculture model. a Adhesion parameter sweep simulations of an agent-based model at varying and constant population sizes. **b** Cell positions at steady state are featurized using persistent homology and order parameters. **c** Fixed-length feature vectors are dimension-reduced and classified using hierarchical clustering. **d** Accuracy is computed by agglomerative clustering after dimensionality reduction and comparing to ground truth.

descriptor vectors was then implemented using UMAP[50], PHATE[51], PCA and autoencoder (AE)[52,53] (Fig. 7c) for comparison with the ground truth classification (Fig. 7bi).

Agent-based modeling of two interacting cell types

Cells were represented as rigid-body agents that undergo non-inertial motion (Eq. (1)) due to a random self-propulsion force, $\mathbf{P}_i^t$, and cell–cell interactions governed by the attraction–repulsion force, $\mathbf{F}_{ij}^t$ (Fig. 8a). Cell positions were initialized in a $[-20, 20] \times [-20, 20]$ simulation box with periodic boundaries on all sides. The equation of motion was given by:

$$\mathbf{x}_i^{t+\Delta t} = \mathbf{x}_i^t + \frac{\Delta t}{\eta}\left(\mathbf{P}_i^t + \sum_{j \neq i}^{N(t)} \mathbf{F}_{ij}^t\right) \tag{1}$$

where $\mathbf{x}_i^t$ denotes the position of the ith cell at time t, Δt denotes the time step, and η denotes a friction coefficient.

Two cells i and j in close proximity (within neighborhood distance $r_{\max} = 1.5$) would attract or repel each other in accordance with an the attraction–repulsion force, $\mathbf{F}_{ij}$, which is the gradient of some potential $U(||\mathbf{x}_j - \mathbf{x}_i||)$ (Eq. (2)) (Fig. 8b):

$$\mathbf{F}_{ij} = -\nabla U(||\mathbf{x}_j - \mathbf{x}_i||)\frac{\mathbf{x}_j - \mathbf{x}_i}{||\mathbf{x}_j - \mathbf{x}_i||}\mathbf{1}_{||\mathbf{x}_j - \mathbf{x}_i|| \leq r_{\max}} \tag{2}$$

In particular, we define U in terms of a Morse potential with a first term that represents long-range attraction and a second term that represents short-range repulsion (Eq. (3)):

$$U(r_{ij}) = -J_{ij}\left[\exp\left(\frac{-r_{ij}}{l_A}\right) - \frac{1}{4}\exp\left(\frac{-r_{ij}}{l_R}\right)\right] \tag{3}$$

where the adhesion parameter $J_{ij} = J(T(i), T(j))$ determines the magnitude of the attraction–repulsion potential, and depends on cell types $T(i)$ and $T(j)$ for cells i and j respectively, which we denote as blue or orange types (e.g., τ_B and τ_O, respectively). We further define the characteristic length scale of the first long-range attraction term as $l_A = 14.0$, and the second short-range repulsion term as $l_R = 0.5$, based on our previous work with epithelial cells[15,48]. Finally, the second short-range repulsion term was further scaled to be 1/4 of the first long-range attraction term.

Cell positions were initialized at $t = 0$ from a uniform distribution with rejection sampling to ensure a minimum separation of

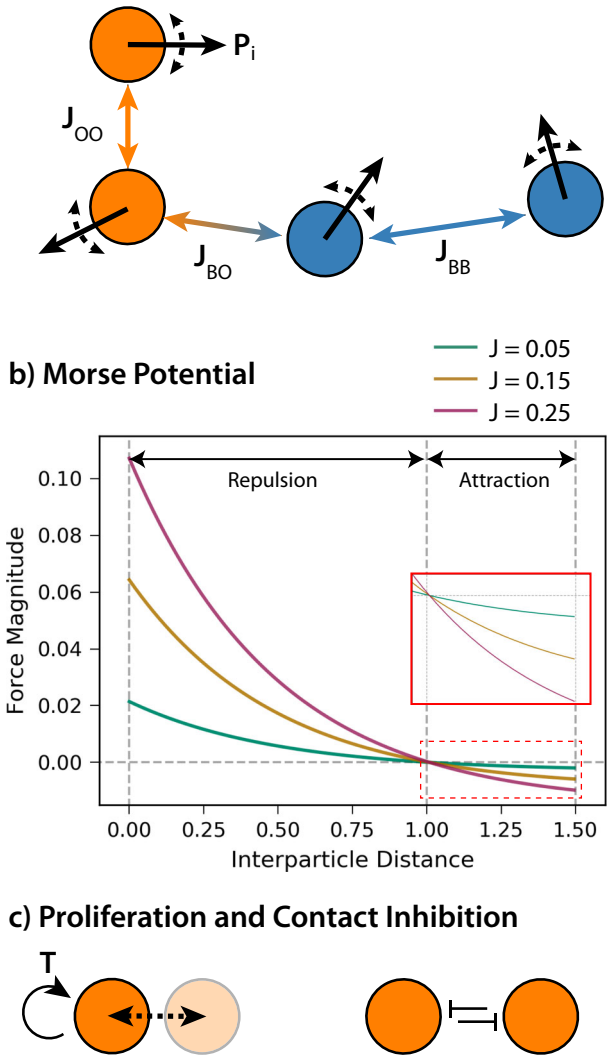

Fig. 8 Schematic of the agent-based model. Cells are modeled as rigid bodies undergoing non-inertial motion. **a** Homotypic and heterotypic cell–cell adhesion parameters. **b** Cell–cell adhesion varies with inter-particle distance based on the Morse potential. **c** Cell proliferation is implemented using an internal cell cycle timer. In addition, cells with greater than four neighbors are unable to proliferate.

1.0 between all cells. The magnitude of the polarization force, $|\mathbf{P}|$, was fixed at 0.005 and the adhesion parameter, J, was varied between 0.01 and 0.25. As a consequence, cells moved directionally at a constant speed for some duration, followed by random reorientation. This behavior has been experimentally validated for epithelial cells[56], and is analogous to classical "run and tumble" models of bacteria.

The friction coefficient $\eta = 1.0$, and time step $\Delta t = 0.02$, were held constant throughout the simulation. Each simulation ran for 5,000,000 time steps. Repolarization takes place every 2500 time steps, with an initial offset to prevent synchronization. In scenarios with proliferation, cell division was modeled with a cell cycle duration of 80,000 time steps, and permitted to occur as long as cells have less than four neighbors (i.e., contact inhibition of proliferation) (Fig. 8c). This scenario with both proliferation and sorting occur is reminiscent of tissue repair after damage, such as the regeneration of zebrafish skin patterns after laser ablation[55].

Altogether, for scenarios at constant cell number (without proliferation), a total of 512 combinations of adhesion parameters were simulated (8 different values of $J_{BB} \times 8$ different values of $J_{OO} \times 8$ different values of J_{BO}), with 3 independently initialized replicates. However, for scenarios with proliferation, simulations with low adhesion were not included since they were too computationally expensive and did not reach steady state over the timescales considered for the other simulations ($J \approx 0.00$, 0.001). Thus, only 216 combinations of adhesion parameters were simulated with proliferation (6 different values of $J_{BB} \times 6$ different values of $J_{OO} \times 6$ different values of J_{BO}), with three independently initialized replicates.

Computation of persistence diagrams, persistence images, and order parameters

Given a point cloud representing cell positions, we extracted topological features by constructing a chain of simplicial complexes based on a "proximity parameter" ϵ (separation distance between cell centroids) (Fig. 9a). Specifically, we used Euclidean distance between the cell positions to compute persistent homology via the Vietoris-Rips filtration. To avoid confusion, we denote these distances as "interval start" and "interval end" rather than "birth" and "death," which have different meanings in cell biology and topology literature. We used TDA to obtain a list of (*start, end*) pairs for topological features. These *start* and *end* pairs can be represented using a barcode diagram, or a persistence diagram for visualization (Fig. 9b). In the worst-case scenario, this computation is performed in $O(n^3)$ time, where n is the number of generators of the filtered complex[57], although this has been sped up using the standard reduction algorithm. In order to perform machine learning on simulation data, it was helpful to encode the topological information into vectors, so persistence images were introduced as a vectorized representation of topological features that are encoded in persistence diagrams.

Persistence images summarize the topological features using an intensity function over a measurement of length connected to the ϵ radius[49]. Generally, in order to compute a persistence image, we take the points from the persistence diagram and place Gaussians centered around each point, additionally weighting the Gaussians so that more significant points have larger weight. Then we take the sum over these Gaussians to produce an intensity. However, depending on the dimensionality of the feature we are interested in, this process, as well as the shape of the output, varies. For each simulation at "steady state", persistence images were computed, using only orange cell positions, only blue cell positions, and both cell type positions together. In each case, we calculated the persistence images in dimensions 0 and 1. As an example, we now discuss further details for the computation of persistence images for dimension 0 and dimension 1.

In dimension 0 homology, we consider only the connected component features and place Gaussians at each point weighted by (*end–start*) resulting in a one-dimensional intensity function (Fig. 9c). Mathematically,

$$\mathrm{PI}(x) = \sum_{(start,end)\in\mathscr{D}} (end - start) g_{(start),\sigma}(x)$$

where $\mathscr{D}$ is the persistence diagram and $g_{(start),\sigma}$ is a Gaussian with mean *start* and variance σ^2.

In dimension 1 homology, the presence/absence of topological loops is indicated by (*start, end*) coordinates. We transform the coordinate system to measure persistence along the *y*-axis, (*start, end–start*), and place Gaussians around these new points, weighted by the distance to the diagonal (Fig. 9d). In other words, weight each Gaussian by $\sqrt{\left(\frac{start-end}{2}\right)^2 + \left(\frac{end-start}{2}\right)^2}$. The result is a two-dimensional intensity function, with higher intensity

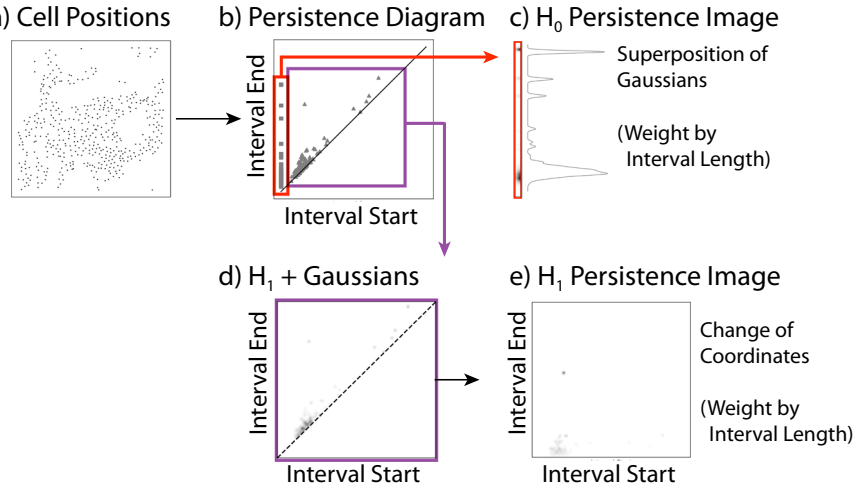

Fig. 9 Flow diagram to generate persistence images. a, **b** A Rips filtration is used on point cloud representing cell positions to compute the persistence diagram. **c**–**e** Persistence images are generated by replacing interval start/end coordinates in the persistence diagram with a Gaussian weighted by distance from the diagonal. In dimension 0 homology (notated H_0, the interval start is always 0 (each cell position is a connected component), resulting in a 1D image. In dimension 1 homology (notated H_1), the intervals represent topological loops arising at nonzero values of the filtration radius, resulting in a 2D image.

for more persistent features (Fig. 9e). Mathematically,

$$\mathrm{PI}\,(x,y) = \sum_{(start,end)\in\mathscr{D}} (end-start)g_{(start,end-start),\Sigma}(x,y)$$

where $g_{(start,\,end-start),\Sigma}$ is a two-dimensional Gaussian centered at (*start, end–start*) with covariance matrix Σ.

Persistence curves are an alternative method of summarizing the persistence diagram by defining a function along its diagonal. Consider a point (t, t) on the diagonal of the persistence diagram $\mathscr{D}$, which defines a rectangular region where (*start, end*) $\in [0,t]\times[t,\infty)$. We apply a function ψ to all points (interval coordinates) within this region and compute a summary statistic T, resulting in a scalar value PC(t). More formally, we define the persistence curve, PC as a real-valued function over the diagonal,[58]

$$\mathrm{PC}\,(\mathscr{D},\psi,T)(t) = T(\psi(\mathscr{D};b,d,t)|(b,d)\in\mathscr{D}_t), t\in\mathbb{R}$$

There are various options for choosing the functions ψ and T. One example is the *Betti curve*, where ψ is the indicator function and T is the summation operator. The Betti curve counts the number of points in the $[0,t]\times[t,\infty)$ rectangular region for all points t along the diagonal. Another example is the *life curve*, where ψ is *(end–start)* and T is the summation operator. Here, we implement the Gaussian persistence curve[59], where ψ places a Gaussian of fixed bandwidth at each point in the $[0,t]\times[t,\infty)$ region, weighted by distance from the diagonal, and T is the integral sum. One benefit of using Gaussian persistence curves is efficiency: they are easy to implement and fast to compute, with theoretical guarantees on stability and injectivity[59]. They have previously been applied to problems in image classification and neuroscience[58,60–62]. Although we focus on persistence images in the main text, persistence curves are benchmarked in the supporting information and represent a promising avenue for future work.

Dimensionality reduction and classification

Persistence images computed from cell positions were concatenated and compressed to a low-dimensional feature vector. Dimensionality reduction was then performed on the feature vectors using principal component analysis (PCA), Uniform Manifold Approximation and Projection (UMAP)[50], Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE)[51], and an autoencoder (AE)[52], which facilitated visualization (in 2D) and unsupervised classification using hierarchical clustering. Briefly, PCA projects data in low dimensions using a linear combination of features to maximize variance along each principal component. UMAP is a learned embedding based on ideas from manifold learning and TDA that constructs a high-dimensional graph representation of the data and optimizes a low-dimensional graph embedding to preserve structural similarities. PHATE computes an information distance based on a diffusion operator constructed from affinity between data points. It leverages multi-dimensional scaling (MDS) to perform dimensionality reduction using the information distance metric. An autoencoder learns the identity map by passing data, x, through an encoder network, E, to obtain a low-dimensional latent representation, $z = E(x)$, which can be decoded to reconstruct the original data via the decoder network, $\hat{x} = D(z) = D(E(x))$. The autoencoder is trained without supervision using the reconstruction loss penalty, $\mathscr{L} = \|x-\hat{x}\|$. The classification results were determined by comparing cluster labels obtained using hierarchical clustering to ground truth phase classification. Agglomerative hierarchical clustering (ward linkage)[63] was performed on 20-dimensional representations obtained by dimensionality reduction, with a stopping criterion (i.e., dendrogram cutoff) to prevent merging when the number of clusters reached the total number of distinct phases in the ground truth.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

All simulation data generated for this study are available via an Open Science Foundation repository at https://osf.io/md86n/?view_only=66ac8655cf1842e4bfe52ca0b8b59a04[64] under an MIT License.

CODE AVAILABILITY

All code used to perform the simulations, compute order parameters and persistent homology, and generate classification results is available on GitHub at https://github.com/dbhaskar92/Coculture-ABM-Model under an MIT license.

Received: 17 January 2023; Accepted: 14 August 2023;
Published online: 14 September 2023

REFERENCES

1. Lecuit, T. & Lenne, P.-F. Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis. *Nat. Rev. Mol. Cell. Biol.* **8**, 633–644 (2007).
2. Tsai, T. Y.-C., Garner, R. M. & Megason, S. G. Adhesion-based self-organization in tissue patterning. *Annu. Rev. Cell Dev. Biol.* **38**, 349–374 (2022).
3. Graner, F. & Glazier, J. Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys. Rev. Lett.* **69**, 2013–2016 (1992).
4. Rieu, J.-P., Kataoka, N. & Sawada, Y. Quantitative analysis of cell motion during sorting in two-dimensional aggregates of dissociated hydra cells. *Phys. Rev. E* **57**, 924–931 (1998).
5. Belmonte, J. M., Thomas, G. L., Brunnet, L. G., de Almeida, R. M. C. & Chaté, H. Self-propelled particle model for cell-sorting phenomena. *Phys. Rev. Lett.* **100**, 220–4 (2008).
6. Hogan, C. et al. Characterization of the interface between normal and transformed epithelial cells. *Nat. Cell. Biol.* **11**, 460–467 (2009).
7. Beatrici, C. P. & Brunnet, L. G. Cell sorting based on motility differences. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **84**, 031927 (2011).
8. Méhes, E., Mones, E., Németh, V. & Vicsek, T. Collective motion of cells mediates segregation and pattern formation in co-cultures. *PLoS ONE* **7**, e31711 (2012).
9. Kabla, A. J. Collective cell migration: leadership, invasion and segregation. *J. R. Soc. Interface* **9**, 3268–3278 (2012).
10. Strandkvist, C., Juul, J., Baum, B., Kabla, A. J. & Duke, T. A kinetic mechanism for cell sorting based on local variations in cell motility. *Interface Focus* **4**, 20140013 (2014).
11. Nielsen, A. V., Gade, A. L., Juul, J. & Strandkvist, C. Schelling model of cell segregation based only on local information. *Phys. Rev. E* **92**, 488–4 (2015).
12. Gamboa Castro, M., Leggett, S. E. & Wong, I. Y. Clustering and jamming in epithelial-mesenchymal co-cultures. *Soft Matter* **12**, 8327–8337 (2016).
13. Carrillo, J. A., Colombi, A. & Scianna, M. Adhesion and volume constraints via nonlocal interactions determine cell organisation and migration profiles. *J. Theor. Biol.* **445**, 75–91 (2018).
14. Carrillo, J. A., Murakawa, H., Sato, M., Togashi, H. & Trush, O. A population dynamics model of cell-cell adhesion incorporating population pressure and density saturation. *J. Theor. Biol.* **474**, 14–24 (2019).
15. Leggett, S. E. et al. Motility-limited aggregation of mammary epithelial cells into fractal-like clusters. *Proc. Natl. Acad. Sci. USA* **116**, 17298–17306 (2019).
16. Li, X., Das, A. & Bi, D. Mechanical heterogeneity in tissues promotes rigidity and controls cellular invasion. *Phys. Rev. Lett.* **123**, 058101 (2019).
17. Krajnc, M. Solid–fluid transition and cell sorting in epithelia with junctional tension fluctuations. *Soft Matter* **16**, 3209–3215 (2020).
18. Sahu, P. et al. Small-scale demixing in confluent biological tissues. *Soft Matter* **16**, 3325–3337 (2020).
19. Dey, S. & Das, M. Differences in mechanical properties lead to anomalous phase separation in a model cell co-culture. *Soft Matter* **17**, 1842–1849 (2021).
20. Lucia, S. E., Jeong, H. & Shin, J. H. Cell segregation via differential collision modes between heterotypic cell populations. *Mol. Biol. Cell* **33**, ar129 (2022).
21. Skamrahl, M. et al. Cellular segregation in cocultures is driven by differential adhesion and contractility on distinct timescales. *Proc. Natl. Acad. Sci. USA* **120**, e2213186120 (2023).
22. Steinberg, M. S. Differential adhesion in morphogenesis: a modern view. *Curr. Opin. Genet. Dev.* **17**, 281–286 (2007).
23. Brodland, G. W. The differential interfacial tension hypothesis (DITH): a comprehensive theory for the self-rearrangement of embryonic cells and tissues. *J. Biomech. Eng.* **124**, 188–197 (2002).
24. Steinberg, M. S. Reconstruction of tissues by dissociated cells. Some morphogenetic tissue movements and the sorting out of embryonic cells may have a common explanation. *Science* **141**, 401–408 (1963).
25. Kasza, K. E. & Zallen, J. A. Dynamics and regulation of contractile actin-myosin networks in morphogenesis. *Curr. Opin. Cell Biol.* **23**, 30–38 (2011).
26. Gibson, M. C., Patel, A. B., Nagpal, R. & Perrimon, N. The emergence of geometric order in proliferating metazoan epithelia. *Nature* **442**, 1038–1041 (2006).
27. Major, R. J. & Irvine, K. D. Localization and requirement for Myosin II at the dorsal-ventral compartment boundary of the Drosophila wing. *Dev. Dyn.* **235**, 3051–3058 (2006).
28. Hayashi, T. & Carthew, R. W. Surface mechanics mediate pattern formation in the developing retina. *Nature* **431**, 647–652 (2004).
29. Hilgenfeldt, S., Erisken, S. & Carthew, R. W. Physical modeling of cell geometric order in an epithelial tissue. *Proc. Natl. Acad. Sci. USA* **105**, 907–911 (2008).
30. Blankenship, J. T., Backovic, S. T., Sanny, J. S. P., Weitz, O. & Zallen, J. A. Multicellular rosette formation links planar cell polarity to tissue morphogenesis. *Dev. Cell* **11**, 459–470 (2006).
31. Krieg, M. et al. Tensile forces govern germ-layer organization in zebrafish. *Nat. Cell Biol.* **10**, 429–436 (2008).
32. Togashi, H. et al. Nectins establish a checkerboard-like cellular pattern in the auditory epithelium. *Science* **333**, 1144–1147 (2011).
33. Kondo, S. & Miura, T. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science* **329**, 1616–1620 (2010).
34. Stevens, A. J. et al. Programming multicellular assembly with synthetic cell adhesion molecules. *Nature* **614**, 144–152 (2023).
35. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2016).
36. Carlsson, G. Topology and data. *Bull. Amer. Math. Soc.* **46**, 255–308 (2009).
37. Edelsbrunner, H. *Computational Topology: An Introduction* (American Mathematical Society, 2009).
38. Amézquita, E. J., Quigley, M. Y., Ophelders, T., Munch, E. & Chitwood, D. H. The shape of things to come: topological data analysis and biology, from molecules to organisms. *Dev. Dyn.* **249**, 816–833 (2020).
39. Topaz, C. M., Ziegelmeier, L. & Halverson, T. Topological data analysis of biological aggregation models. *PLoS ONE* **10**, e0126383 (2015).
40. Ulmer, M., Ziegelmeier, L. & Topaz, C. M. A topological approach to selecting models of biological experiments. *PLoS ONE* **14**, e0213679 (2019).
41. Atienza, N., Escudero, L. M., Jimenez, M. J. & Soriano-Trigueros, M. Persistent entropy: a scale-invariant topological statistic for analyzing cell arrangements. Preprint at https://arxiv.org/pdf/1902.06467v4.pdf (2019).
42. Bhaskar, D. et al. Analyzing collective motion with machine learning and topology. *Chaos* **29**, 123125 (2019).
43. McGuirl, M. R., Volkening, A. & Sandstede, B. Topological data analysis of zebrafish patterns. *Proc. Natl. Acad. Sci. USA* **117**, 5113–5124 (2020).
44. Skinner, D. J. et al. Topological metric detects hidden order in disordered media. *Phys. Rev. Lett.* **126**, 048101 (2021).
45. Nardini, J. T., Stolz, B. J., Flores, K. B., Harrington, H. A. & Byrne, H. M. Topological data analysis distinguishes parameter regimes in the anderson-chaplain model of angiogenesis. *PLoS Comput. Biol.* **17**, e1009094 (2021).
46. Stolz, B. J. et al. Multiscale topology characterizes dynamic tumor vascular networks. *Sci. Adv.* **8**, eabm2456 (2022).
47. Kramár, M., Goullet, A., Kondic, L. & Mischaikow, K. Quantifying force networks in particulate systems. *Phys. D Nonlin. Phenomena* **283**, 37–55 (2014).
48. 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 (2021).
49. Adams, H. et al. Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**, 1–35 (2017).
50. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/pdf/1802.03426.pdf (2018).
51. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
52. Bank, D., Koenigstein, N. & Giryes, R. Autoencoders. In *Machine Learning for Data Science Handbook* (eds Rokach, L., Maimon, O. & Shmueli, E.). (Springer, Cham, 2023). https://doi.org/10.1007/978-3-031-24628-9_16.
53. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 37–49 (JMLR Workshop and Conference Proceedings, 2012).
54. Volkening, A. Linking genotype, cell behavior, and phenotype: multidisciplinary perspectives with a basis in zebrafish patterns. *Curr. Opin. Genet. Dev.* **63**, 78–85 (2020).
55. Yamaguchi, M., Yoshimoto, E. & Kondo, S. Pattern regulation in the stripe of zebrafish suggests an underlying dynamic and autonomous mechanism. *Proc. Natl. Acad. Sci. USA* **104**, 4790–4793 (2007).
56. Potdar, A. A., Jeon, J., Weaver, A. M., Quaranta, V. & Cummings, P. T. Human mammary epithelial cells exhibit a bimodal correlated random walk pattern. *PLoS ONE* **5**, e9636 (2010).
57. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
58. Chung, Y.-M., Hull, M. & Lawson, A. Smooth summaries of persistence diagrams and texture classification. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3667–3675 (IEEE, 2020).
59. Chung, Y.-M., Hull, M., Lawson, A. & Pritchard, N. Gaussian persistence curves. Preprint at http://arxiv.org/abs/2205.11353 (2022).
60. Gönen, M. & Alpaydin, E. Multiple kernel learning algorithms. *J. Machine Learn. Res.* **12**, 2211–2268 (2011).
61. Chung, Y.-M., Hu, C.-S., Lawson, A. & Smyth, C. Topological approaches to skin disease image analysis. in *2018 IEEE International Conference on Big Data (Big Data)*, 100–105 (IEEE, 2018).

62. Barnes, D., Polanco, L. & Perea, J. A. A comparative study of machine learning methods for persistence diagrams. *Front. Artif. Intell.* **4**, 681174 (2021).
63. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at http://arxiv.org/abs/1109.2378 (2011).
64. Bhaskar, D. Code and data for "TDA of Spatial Patterning in Heterogeneous Cell Populations". https://osf.io/md86n/ (2023).

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences (R01GM140108), a Brown University Data Science Initiative Seed Grant (D.B., W.Y.Z., and I.Y.W.), and the National Science Foundation (CCF-1740741, DMS-2038039, DMS-2106566) (B.S.). This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

AUTHOR CONTRIBUTIONS

D.B.: conceptualization, methodology, software, formal analysis, investigation, visualization, and writing—original draft preparation. W.Y.Z.: methodology, software, formal analysis, investigation, visualization, and writing—review and editing. A.V.: methodology, formal analysis, and writing—review and editing. B.S.: methodology, formal analysis, and writing—review and editing. I.Y.W.: conceptualization, methodology, formal analysis, visualization, writing—original draft preparation, and supervision.

COMPETING INTERESTS

D.B. is currently supported by a Boehringer Ingelheim Fellowship at Yale University. The remaining authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-023-00302-8.

Correspondence and requests for materials should be addressed to Ian Y. Wong.

Reprints and permission information is available at http://www.nature.com/reprints

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