Topological Data Analysis of Collective and Individual Phases in a Minimal Model of Epithelial Cells

Dhananjay Bhaskar[1,2], William Y. Zhang[3] and Ian Y. Wong*[1,2]

[1]School of Engineering, Center for Biomedical Engineering, Brown University, 184 Hope St Box D, Providence, RI 02912
[2]Data Science Initiative, Brown University, 184 Hope St Box D, Providence, RI 02912
[3]Department of Computer Science, Brown University, 184 Hope St Box D, Providence, RI 02912

March 24, 2020

Abstract

Interacting, self-propelled particles such as motile epithelial cells can dynamically self-organize into large-scale patterns. In such living systems, cell number and density can vary dramatically over time due to proliferation, which is not commonly considered in other active matter systems. As a consequence, it remains challenging to determine individual and collective phases over varying populations sizes without *a priori* information. Here, we demonstrate an unbiased machine learning approach to analyze multiparticle clusters based on topological structure, which is robust to changes in population size. For a given particle configuration, topological data analysis (TDA) determines the stability of spatial connectivity at varying length scales (i.e. persistent homology), and can compare different particle configurations based on the "cost" of reorganizing one configuration into another. We show that TDA can accurately map out phase diagrams for interacting particles with varying adhesion and self-propulsion, at constant population size as well as when proliferation is permitted. Next, we use this approach to profile our recent experiments on the clustering of epithelial cells in varying growth factor conditions. Finally, we characterize the statistical robustness of this approach over repeated simulations and with random particle removal. Overall, we envision TDA will be broadly applicable as a model-agnostic approach to analyze active systems with varying population size, from cytoskeletal motors to motile cells to flocking or swarming animals.

1 Introduction

Collective behaviors emerge from multi-particle interactions, resulting in rich self-organizing patterns.[1] For instance, epithelial cells assemble into tightly connected multicellular layers due to strong cell-cell and cell-matrix adhesions, representing a fascinating system of nonequilibrium dynamics.[2] Moreover, multicellular clusters can "scatter" as migratory individuals in response to biochemical stimuli,[3–5] analogous to an epithelial-mesenchymal transition.[6] Instead, dispersed and motile individuals can transition towards collective migration and ultimately arrested states, analogous to a "jamming" transition.[7–19]

Epithelial cells can be computationally modeled as self-propelled particles in two dimensional space.[20] Such models treat cells as disks with some isotropic repulsive potential, which move persistently at constant speed in the absence of additional interactions.[7,13,21–34] These self-propelled particles can further interact via attractive potentials or local alignment (e.g. Vicsek model),[7] resulting in spatiotemporal correlations in position and velocity. However, a potentially confounding behavior of living systems is that the size of the population changes over time due to proliferation or death,[31] which is not commonly considered in active matter systems.

Topological data analysis (TDA) is an emerging mathematical framework for visualizing the underlying "structure" of high-dimensional datasets based on the spatial connectivity between discrete points.[35] In particular, TDA will determine the robustness of connectivity between features over a range of spatial scales (i.e. persistent homology), which can be then represented as a characteristic "topological barcode."[36] Topological barcodes have been used to visualize swarming or patterning behaviors in living entities,[34,37–39] as well as percolation thresholds in 2D disk packing,[40] but have been typically implemented at constant population size. TDA represents an exciting approach for unbiased and unsupervised analysis for (dis)ordered and collective phases in active matter systems, but its validity for time-varying population sizes has not been established.

In this article, we use TDA to elucidate collective and individual phases in self-propelled particles, as a minimal model of epithelial cell migration. We show that TDA enables unbiased and unsupervised classification of distinct phases and transitions. We first apply TDA on a training set of interacting self-propelled particles with varying adhesion but no proliferation. We subsequently generalize TDA for interacting self-propelled particles that exhibit significant proliferation over the course of the simulation. We show that TDA can be utilized for experimental data based on tracking epithelial cell nuclei, and accurately classifies experimental results from different biochemical treatments. Finally, we examine how TDA compares different particle configurations by randomly removing particles one by one from an initial configuration. We envision that TDA will be broadly applicable for visualizing how living units migrate, proliferate, and interact across length scales from molecular motors to mammalian cells to animals.

2 Computational Model

Our model represents epithelial cells as self-propelled particles with three features. First, particles travel at constant velocity but randomly polarize in new directions at constant intervals (offset to different times). Second, particles interact with nearby neighbors through a short-range repulsion corresponding to the particle radius, as well as a tunable attractive interaction. Third, particles can proliferate at regular intervals (offset to different starting times), unless surrounded by four or more neighbors (i.e. contact inhibition of proliferation) (**Fig. 1a**).

Simulations were initialized with 200 particles randomly placed on a square domain ($[-10,10]\times[-10,10]$) with periodic boundary conditions. To ensure that particles were not too close together at

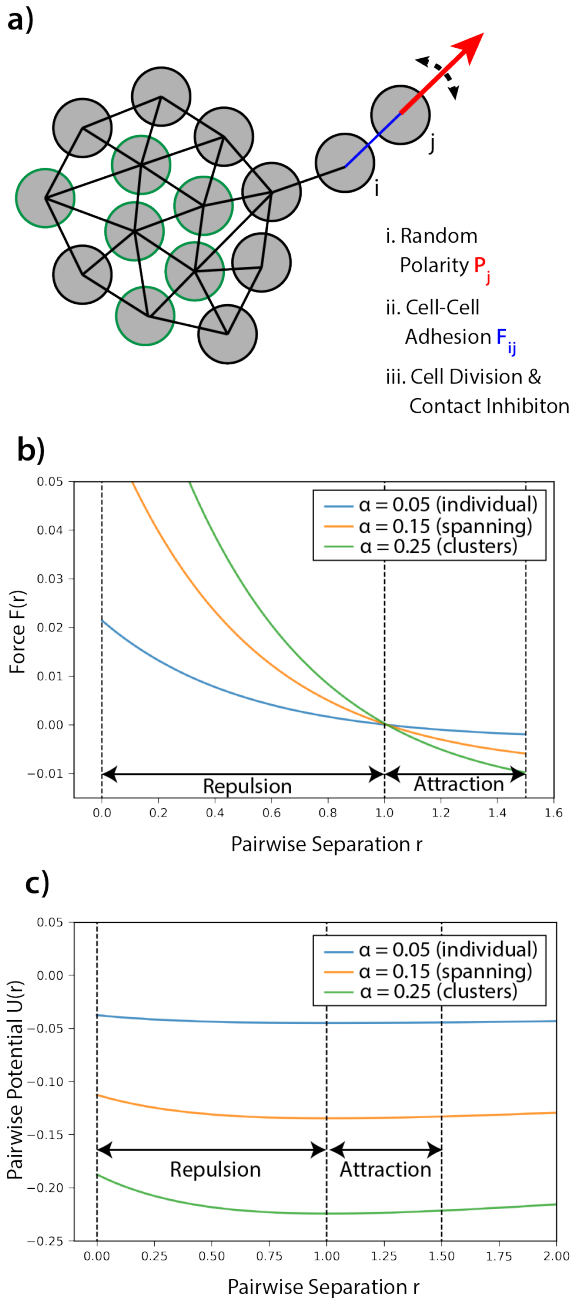

Figure 1: **Self-propelled particle model.** (a) Cells were represented as self-propelled disks subject to time-varying random polarization force $\mathbf{P}$ and cell-cell adhesion force $\mathbf{F}$. A "bond" was drawn between two cells if they were within radius $r = 1$ of one another. Cells with 4 or more neighbors (outlined in green) were not permitted to proliferate. (b) The adhesion force exerted on cell i (located at $r = 0$) due to neighboring cell j, $\mathbf{F}_{ij}$, as a function of radial distance r was plotted for various values of the adhesion parameter α. Long-range attractive force (pointing inwards towards cell i) was negative and short-range repulsion (pointing outwards, away from cell i) was positive. Note that the attraction force was cut-off at $r = \epsilon = 1.5$. (c) The attraction-repulsion kernel U, was plotted for various adhesion parameter values, α. Between $0 \leq r \leq \epsilon$, the kernel was minimized at $r = 1$, the equilibrium distance we use to define neighboring cells indicated by a "bond" drawn between them.

initialization, a rejection sampling algorithm was used. At least three simulations with identical parameter values but distinct initial conditions were run over 150,000 timesteps using the following over-damped equation of motion:

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

where $\mathbf{r}_i^t$ was the position vector of particle i at time t, Δt was the time step (default value $\Delta t = 0.02$), γ represented a drag coefficient (with default value $\gamma = 1$) and $N(t)$ was the number of particles at time t.

The second term $\mathbf{P}_i^t$ represented a random polarity force on particle i at time t, with constant magnitude varying from $0.005-0.025$, and direction chosen uniformly at random once every $2,500$ timesteps. To prevent cells from repolarizing at the same time, an offset (a random value chosen uniformly between 0 and 500 timesteps) was initially subtracted from the total time to repolarization for each cell.

The third term $\mathbf{F}_{ij}$ represented pairwise cell-cell interactions for cell i with other cells j, and is plotted in **Fig. 1b,c** for three representative adhesion values:

$$\mathbf{F}_{ij} = -\nabla U(\|\mathbf{r}_j - \mathbf{r}_i\|)\frac{\mathbf{r}_j - \mathbf{r}_i}{\|\mathbf{r}_j - \mathbf{r}_i\|}\mathbb{1}_{0\leq\|\mathbf{r}_j-\mathbf{r}_i\|\leq\epsilon} \tag{2}$$

where the attraction-repulsion kernel U governed the overall magnitude of adhesion and repulsion between any pair of cells. Note that the cell-cell interaction force is only active at radial distances between 0 and $\epsilon = 1.5$, preventing cells from attracting other cells located far away. The interaction force was obtained by computing the gradient of this potential function, which included 4 parameters:

$$U(r) = -c_A e^{-r/L_A} + c_R e^{-r/L_R} \tag{3}$$

which specified length scales for long range attraction ($L_A = 14.0$) and short range repulsion ($L_R = 0.5$) as well as the relative strength of attraction and repulsion ($c_A = \alpha$ and $c_R = 0.25\alpha$, respectively). The parameter α, varying from $0.07-0.25$, controls the strength of the adhesion and repulsion force.

For some simulations, proliferation was also included by adding a new "daughter" particle placed close to the "parent" with a polarization vector in the opposite direction. For all particles, the total cell cycle time was the same (50,000 timesteps), with an initial randomly chosen offset (between 0 and 10,000 timesteps) to avoid biologically unrealistic synchrony in cell division. Particles with 4 or more nearest neighbors were not permitted to undergo division, representing contact inhibition of proliferation.[41]

Finally, particles were defined as neighbors if they were positioned within a radius of $r = 1$ from each other, which is indicated by plotting a "bond" between these particles. A group of 4 or more neighboring cells, with cell-cell adhesion bonds that persist over many simulation time-steps was considered as a cluster.

All simulations were conducted at the Brown Center for Computation and Visualization. Persistence of topological features was quantified by extracting a barcode (also represented as a persistence diagram) using the Vietoris-Rips complex, implemented in Julia's Eirene package.[42] Both simulation code and TDA code will be made available (upon publication) on Github.

3 Topological Barcodes and Wasserstein Distance

Topological data analysis (TDA) computationally visualizes the "shape" of data from the spatial connectivity between discrete points.[35] We provide a brief primer here, and refer interested readers to more comprehensive texts on this topic.[43] Essentially, some arbitrary set of discrete data points (i.e. point cloud) can be understood as a (noisy) sampling of some underlying, lower-dimensional topological space. In order to extract this information, the point cloud can be represented by connected components at varying length scales as a simplicial complex. For instance, two points within some distance ϵ (filtration) can be linked together by an edge (forming a connected component characterized by Betti number, $\beta_0 = 1$). Next, a circular set of points that are pairwise within a separation distance ϵ can be linked into a closed loop enclosing a one-dimensional hole (characterized by Betti number, $\beta_1 = 1$). For our 2-D point cloud representing cell positions, edges and loops are sufficient to capture topological structure, but analogues in higher dimensions (e.g. 3-D voids, n-dimensional holes for $n > 3$) can be extracted from simplicial complexes to quantify more complex topological spaces.

Topological barcodes visualize the robustness (via persistent homology) of topological features such as edges and loops across varying length scales (i.e. filtration values) ϵ.[36] For example, consider a set of 18 points at varying filtration, illustrated by red disks with radius ϵ_1 centered at each point (**Fig. 2a**). As the radius increases to ϵ_2, certain red disks overlap, indicating that the corresponding points should be connected by edges into a simplicial complex (represented by i) at this scale. Moreover, these connected edges form a closed loop around an empty region, denoted I. A further increase in radius to ϵ_3 results in the formation of a second connected component (ii) with a closed loop II, but the first closed loop I collapses. Finally, at the largest radius of ϵ_4, all the points are connected into a single connected component (iii), which persists even as the radius approaches infinity.

The corresponding topological barcode uses horizontal bars to visualize the persistence of features that appear or disappear at varying ϵ intervals (**Figure 2b**). Essentially, the barcode presents Betti intervals whose start corresponds to the ϵ-value for the appearance of a topological feature (i.e. the lowest ϵ-value at which the feature is "born"), and whose end corresponds to the ϵ-value for the disappearance of the topological feature (i.e. the highest ϵ-value at which the feature is present, after which it "dies"). For example, at ϵ_1, there are 18 distinct blue bars corresponding to the number of discrete points, which are not connected at this length scale. At ϵ_2, there are only nine blue bars, since ten points have linked together forming a connected component (i), and there is the appearance of a closed loop, denoted I, indicated by a red bar. At ϵ_3, there are only two blue bars are left, since the remaining 8 points have linked together into a single connected component (ii), and the presence of another closed loop, II, is indicated by another red bar. However, note that loop I has collapsed, and this bar does not exist at ϵ_3. Finally, at ϵ_4, there are only two bars - one blue bar corresponding to the connected component (iii) that consists of all points linked together, and one red bar corresponding to the continuation of loop II. Note that the second red bar persists longer, indicating that loop II is bigger than loop I and therefore more significant. The same information

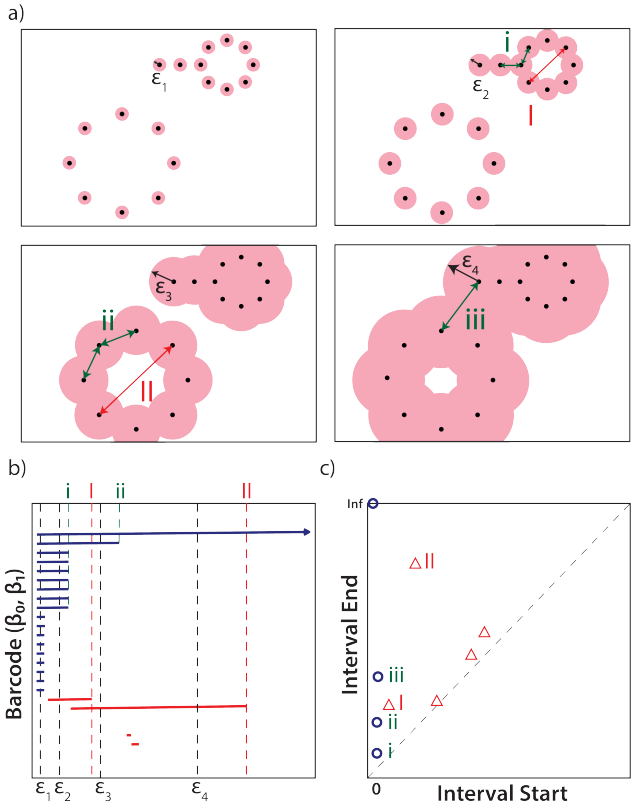

Figure 2: **Computation of persistence homology.** (a) Visualization of connectivity between points at varying values of spatial parameter ϵ. (b) Topological barcode. (c) Persistence diagram.

can also be organized in a persistence diagram (birth ϵ values on x-axis and death ϵ values on y-axis), where the distance from the diagonal is indicative of the significance or importance of a topological feature (**Fig. 2c**). For instance, note that features ii, iii, I, II are appreciably offset from the diagonal, signifying that they are relatively stable topological structures.

To compare two persistence diagrams X and Y, the notion of distance between these diagrams is defined using the Wasserstein metric as follows. First, a bijection, $s : X \to Y$, is defined by matching all off-diagonal points in X with an off-diagonal point in Y. Points on the diagonal (corresponding to very short-lived and insignificant topological features) do not contribute to the distance between persistence diagrams. In case the two diagrams contain an unequal number of points, we also permit points to be matched to their projection on the diagonal, effectively ignoring them. Matching points across diagrams requires solution of an assignment problem, which is easier if the number of points in both diagrams are identical.[44] Therefore, in practice, projections of off-diagonal points to the diagonal are exchanged between persistence diagrams before matches are obtained (**Fig. 3, a-c**). The Wasserstein distance, $W(X, Y)$ is then defined as the infimum over all possible bijections, s:

$$W_{q,p}(X, Y) = \inf_{s:X\to Y} \left(\sum_{x\in X} \|x - s(x)\|_p^q \right)^{\frac{1}{q}} \quad (4)$$

where for $p, q = 2$ we minimize the sum of squared Euclidean distances.

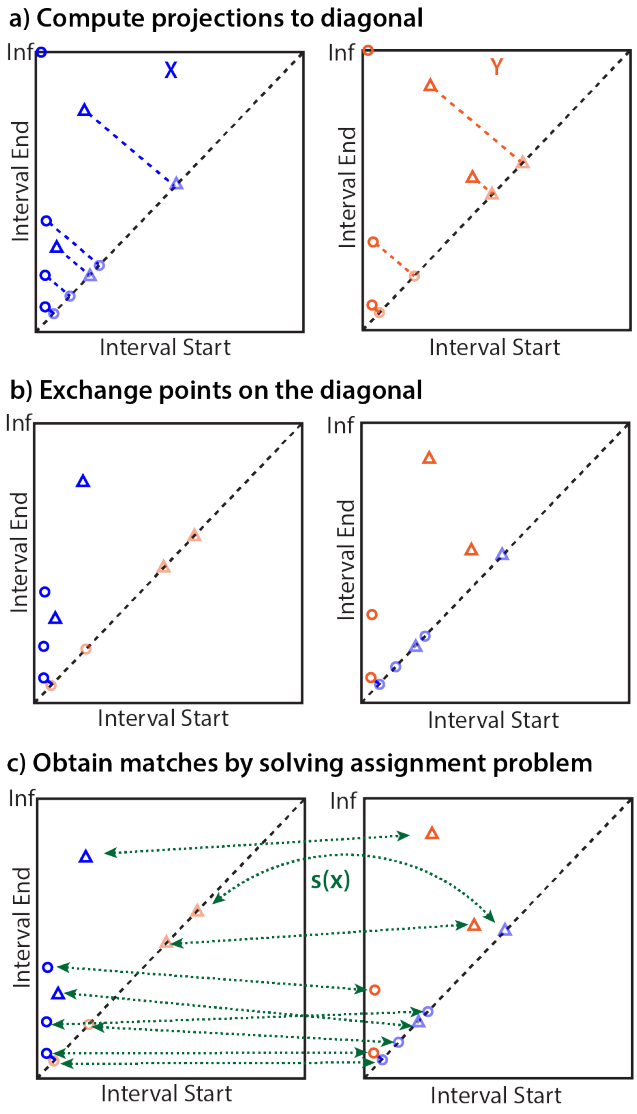

Figure 3: **Computation of Wasserstein distance.** (a) Projections of off-diagonal points to the diagonal are computed. Circles represent edges or connected components and triangles represent loops. Point at $[0, \infty)$ representing 1 connected component for high values of spatial parameter ϵ is not considered. (b) Projections on the diagonal are exchanged between persistence diagrams. (c) Points are matched to their closest neighbor in the other diagram. Note that points can also be matched to their diagonal projection. Circles can only be matched to other circles and triangles can only be matched to other triangles.

To compare multiple simulations and experimental data, pairwise Wasserstein distances between persistence diagrams derived from particle positions (at the end of the simulation or experiment) were computed.

4 Results

4.1 Comparing Topological Structures using Wasserstein Distance

We validated TDA for classifying collective and individual phases in systems of interacting, self-propelled particles through the following approach. First, we established a computational model of self-propelled particles that traveled with constant velocity but randomly changed direction, and interacted with a tunable short-range repulsion and longer-ranged attraction (**Fig. 4a**). We implemented a second variant of this model where particles proliferated at regular time intervals (offset to different times to prevent synchronous division), unless surrounded by several neighbors (which mimics contact inhibition of proliferation). Next, we analyzed the phase behavior of self-propelled particles using known order parameters (i.e. local particle density) (**Fig. 4b**). For comparison, we also analyzed particle configuration using a topological barcode, and determined the "similarity" between persistence diagrams using the Wasserstein distance (**Fig. 4c**). Finally, we used these two complementary approaches to analyze experimental data on epithelial cell migration from our previous publication,[18] showing distinct individual and collective migratory behaviors as well as cluster morphologies (**Fig. 4d**).

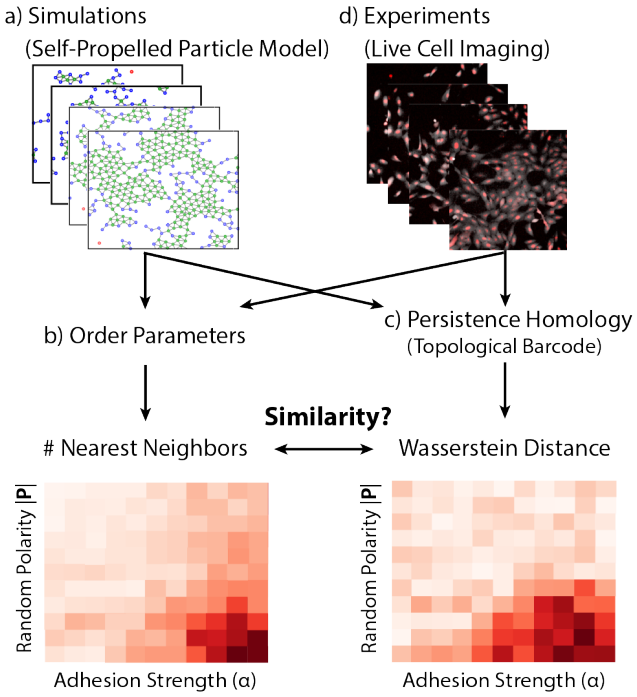

Figure 4: **Methodology.** Self-propelled particle simulations and experimental cell nuclei positions were analyzed using an order parameter (mean local density) and persistent homology (topological barcode). Distinct individual or collective phases were identified by comparing across simulations or experimental conditions.

4.2 Individual and Clustered Phases Exhibit Distinct Topological Structure at Constant Population Size

First, we considered a system consisting of self-propelled particles, where the speed of the ith particle at time t is specified by a polarity force $\mathbf{P}_i^t$ with random orientation, which repolarizes in a different direction after some duration. Moreover, we varied the relative adhesive interactions through the parameter α, which sets the magnitude of the pairwise potential. As the polarity force $\mathbf{P}$ and adhesion strength α were varied, three representative phase behaviors were qualitatively observed at the completion of the simulation ($t = 150,000\Delta t$). First, for

strong polarity force $\mathbf{P}$ and weak adhesion strength α, particles remained individual or interacted transiently as unstable clusters (**Fig. 5a,i**). Next, when polarity force $\mathbf{P}$ and adhesion strength α were comparably strong or weak, a spanning phase was observed where clusters exhibited extended branched morphology (**Fig. 5a,ii,iii**). Finally, for weak polarity force $\mathbf{P}$ and strong adhesion strength α, all particles were incorporated within larger rounded clusters(**Fig. 5a,iv**). It should be noted that the particle dynamics at the completion of the simulations had reached some steady state, where particles either remained as individuals throughout the simulation (**Fig. S1a**), were associated with a spanning network in dynamic equilibrium (**Fig. S1b,c**) or isolated clusters (**Fig. S1d**), before the completion of the simulation ($t = 150,000\Delta t$).

These distinct phases were classified using an order parameter that counted the number of nearest neighbors within a distance of 1.0, representing the local particle density (**Fig. 5b**). This was a more useful readout of clustering, since total particle number remained fixed. In the limit of strong polarity force $0.01 < \|\mathbf{P}\|$ and weak adhesion strength $\alpha < 0.2$, particles were observed to be mostly migratory individuals (**Fig. 5b,i**), with an ensemble averaged number of nearest neighbors $\langle n \rangle \approx 1$. Instead, in the limit of weak polarity force $\|\mathbf{P}\| < 0.015$ and strong adhesion strength $0.14 < \alpha$, particles were typically organized into large clusters (**Fig. 5b,iv**), with $\langle n \rangle \approx 5$ nearest neighbors. Finally, when polarity force $\mathbf{P}$ and adhesion strength α were comparable between these two regimes, a spanning phase was observed with $\langle n \rangle \approx 3$ (**Fig. 5b,ii,iii**). In order to determine the statistical distribution of $\langle n \rangle$, these values were calculated for 10 independent simulations with different initial particle configurations, but identical polarity force $\mathbf{P}$ and adhesion strength α. For instance, individual phases typically showed a mean $\pm$ standard deviation of $\langle n \rangle \pm \sigma_n = 1.2 \pm 0.29$. Moreover, spanning phases showed $\langle n \rangle \pm \sigma_n = 3.1 \pm 0.38$, and clustered phases showed $\langle n \rangle \pm \sigma_n = 5.3 \pm 0.39$ (**Fig. S2a**). One drawback of this approach is that $\langle n \rangle$ was defined based on *a priori* information, since the expected interparticle spacing required knowledge of the pairwise interaction potential.

For comparison, we also computed the pairwise Wasserstein distances between the persistence diagrams for all 121 simulations over varying polarity force $\mathbf{P}$ and adhesion strength α. Hierarchical clustering of Wasserstein distance grouped simulations by clustered, individual, spanning, and a mixed spanning + clusters phase along the diagonal (**Fig. S3**). This analysis also revealed several noteworthy off-diagonal entries, indicating some similarity between clustered and "spanning with clusters" phases, as well as individual and spanning phases (**Fig. S3**). Based on this classification, distinct parameter regimes were mapped out corresponding to individual, spanning, spanning with clusters, as well as clustered phases. Indeed, these phases calculated using TDA show good agreement with the phases defined based on nearest neighbors $\langle n \rangle$ (**Fig. 5c**). Nevertheless, several conditions were misclassified, including a clustered simulation that would have been expected to be individual at $\alpha = 0.05$ and $\|\mathbf{P}\| = 0.021$, or expected to be spanning at $\alpha = 0.23$ and $\|\mathbf{P}\| = 0.021$. Moreover, another simulation was labelled as an outlier (indicated in yellow) at $\alpha = 0.11$ and $\|\mathbf{P}\| = 0.017$ because it did not show similarity to any other parameter value. Overall, TDA can be used for unsupervised classification of individual, spanning, and clustered phases in snapshots of self-propelled particles, in excellent agreement with the phases and phase boundaries defined by a predefined order parameter.

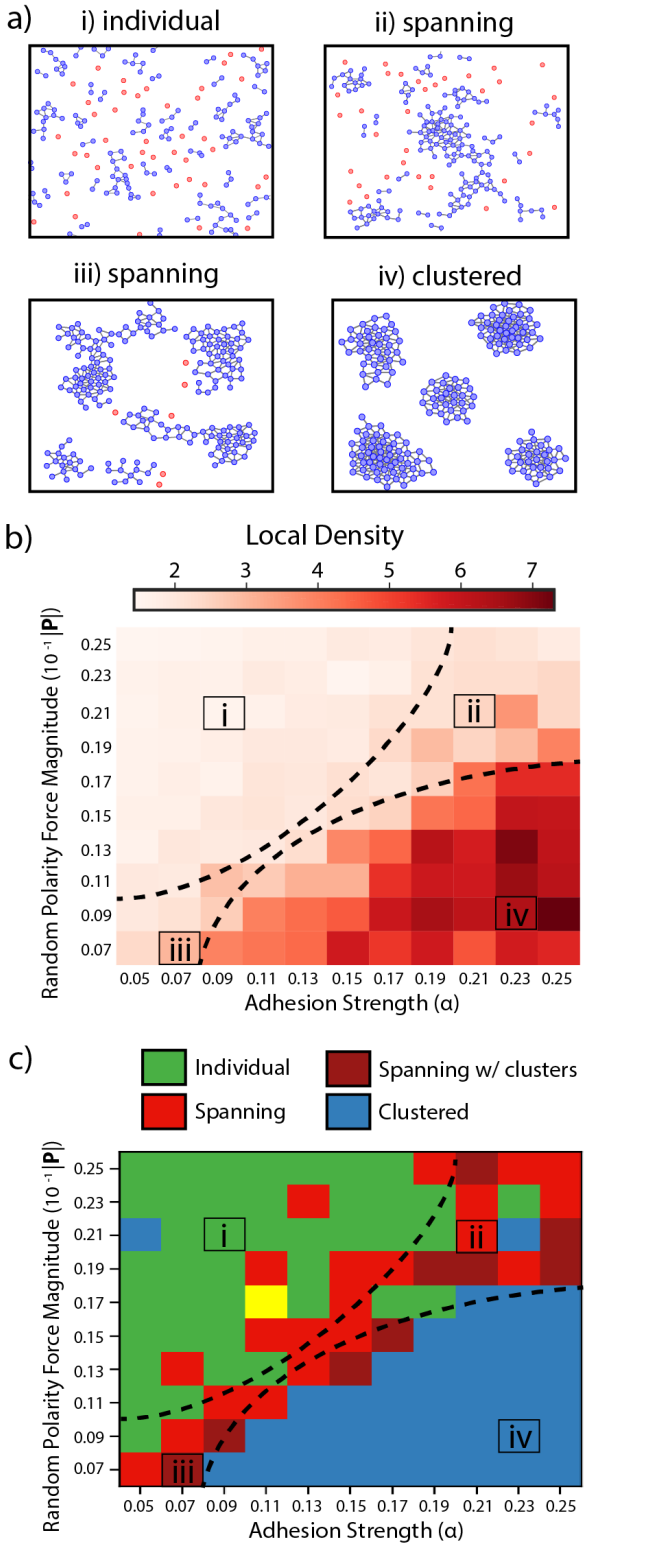

Figure 5: **Individual and clustered phases exhibit distinct topological structure at constant population size.** (a) Snapshots of final configurations observed in simulations of the self-propelled particles for various adhesion and polarization values. (b) Comparison of individual, spanning, and clustered phases based on counting the ensemble averaged number of nearest neighbors within $r = 1$. (c) Comparison of individual, spanning, and clustered phases classified by topological data analysis.

4.3 Spanning and Clustered Phases Exhibit Distinct Topological Structures in Proliferating Populations

Next, we considered a system consisting of proliferating self-propelled particles, where a parent particle divided after a fixed duration (50,000 timesteps), randomly offset. This proliferation was implemented by maintaining the parent particle with the same velocity and direction, but adding a second daughter particle (close to the parent) moving with equal velocity but opposite in direction to the parent. Moreover, the parent particle could not divide if it had more than four neighbors, which mimics the contact inhibition of proliferation of epithelial cells at high density.[41] The polarity force $\mathbf{P}$ and adhesive interaction α were again systematically varied over the same range as in the previous simulations without proliferation. Simulations were analyzed at the final timestep after $150,000\Delta t$.

In the limit of weak adhesion strength α, proliferating particles were observed as a spanning phase with small branched clusters and $\langle n \rangle \approx 3$ (**Fig. 6a,i,iii; b**), which differs from the migratory individuals observed previously without proliferation (**Fig. 5a,i; b**). This difference occurred since individual cells (with few neighbors) were permitted to proliferate, whereas cells within a large cluster (with many neighbors) were not allowed to proliferate based on contact inhibition of proliferation. Next, when polarity force $\mathbf{P}$ and adhesion strength α were comparably strong, a spanning phase was observed with larger clusters that exhibited extended, branching conformations and $\langle n \rangle \approx 4$ (**Fig. 6a,ii; b**). Finally, for weak polarity force $\mathbf{P}$ and strong adhesion strength α, all particles were associated with clusters of compact morphology and $4 < \langle n \rangle < 5$ (**Fig. 6a,iv; b**). We further verified the variation in $\langle n \rangle$ by running simulations with different initial conditions but identical parameters for polarity force and adhesion strength (**Fig. S2b**). Due to contact inhibition of proliferation, the particle dynamics and population size approached some steady state at the completion of the simulations, where particles either remained in a dynamic equilibrium within a spanning network (**Fig. S4a,b**) or as isolated clusters (**Fig. S4c**), well before the completion of the simulation ($t = 150,000\Delta t$). Nevertheless, it should be noted that population size varied from 160 - 360 particles across varying parameter values, with larger total numbers of particles at high polarity and low adhesion, and decreasing particle numbers with decreasing polarity and increasing adhesion, as more clusters formed.

We again computed pairwise Wasserstein distance between the persistence diagrams of all 121 simulations with varying polarity force $\mathbf{P}$ and adhesion strength α. Hierarchical clustering of Wasserstein distance grouped simulations by spanning and clustered phases along the diagonal (**Fig. S5**). Interestingly, the spanning grouping was further divided into two spanning subgroups that alternated with the "spanning with clusters" subgroups by the hierarchical clustering algorithm (complete linkage method, implemented in R `hclust` function) (**Fig. S5**). Unexpectedly, these two spanning subgroups showed increased similarity with off-diagonal entries. Moreover, the first "spanning with clusters" subgroup (from the top left) showed high similarity with the second spanning subgroup, as well as the purely clustered group (**Fig. S5**). Lastly, one simulation ($\alpha = 0.19, \|\mathbf{P}\| = 0.023$) was misclassified as spanning with clusters. Mapping these classifications back to the phase diagram shows good agreement with the phases defined by the nearest neighbor order parameter (**Fig. 6c**). Indeed, the top left, top right, and bottom left regions

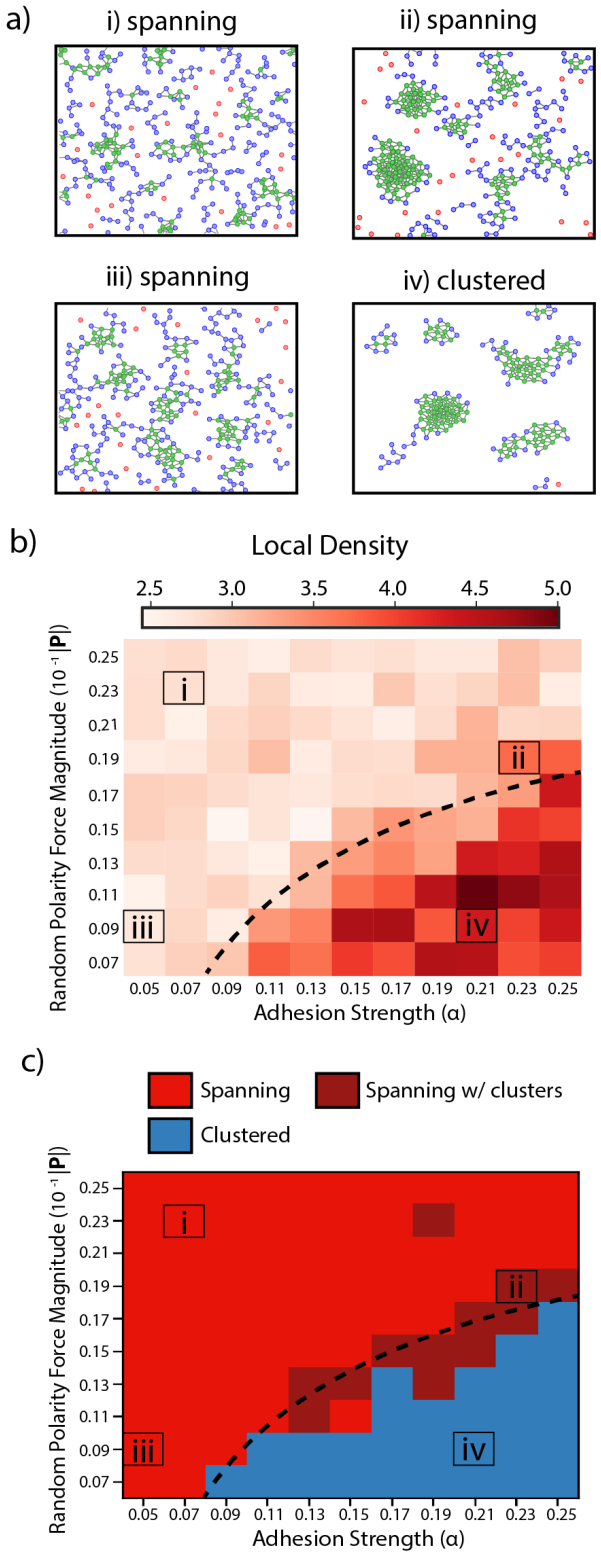

Figure 6: **Spanning and clustered phases exhibit distinct topological structure at constant population size.** (a) Snapshots of final configurations observed in simulations of the self-propelled particles for various adhesion and polarization values. (b) Comparison of spanning and clustered phases based on counting the ensemble averaged number of nearest neighbors within $r = 1$. (c) Comparison of spanning and clustered phases classified by topological data analysis.

were classified as spanning, the bottom right was classified as clustered, and some transition region of "spanning with clusters" classified between them. These results show for the first time that TDA can perform unsupervised classification when population size varies significantly, showing quantitatively similar results as spanning and clustered phases defined by some predetermined order parameter.

4.4 Classifying Experimentally Measured Epithelial Cells after Varying Biochemical Treatments

As a case study, we sought to classify our recent experimental measurements of mammary epithelial cells (MCF-10A) that transition from individuals to clusters when cultured in "assay" media with reduced concentrations of epidermal growth factor (EGF, 0.075 ng/mL).[18] We previous showed that these cells in assay media exhibited slower proliferation and migration over 60 h, organizing over time into clusters with extended branching (fractal-like) architectures, analogous to diffusion-limited aggregation of non-living colloidal particles. These branching conformations were more pronounced after treatment with 4-hydroxytamoxifen (OHT), which activated EMT through an inducible Snail-estrogen receptor construct to drive leader cell formation,[6] relative to a DMSO control with more morphologically compact clusters. In comparison, cells cultured in "growth" media with considerably higher concentrations of EGF (20 ng/mL) remained highly migratory as individuals, before eventually proliferating over 60 h to fill the field of view as a confluent monolayer. In combination with varying initial cell densities, these experimental measurements represent a more challenging test set for TDA-based classification.

The cell positions were defined from the centroid of fluorescent nuclei (i.e. mCherry-H2B), which were detected as described previously.[18] Persistence homology and pairwise Wasserstein distances were computed using the same methodology described above for analyzing simulation data. Hierarchical clustering (using complete linkage method, implemented in `hclust` function in R) was employed to group together similar experiments based on Wasserstein distance.

Experimental conditions cultured with growth media (20 ng/mL EGF) were typically confluent monolayers with high cell density after 60 h, and were consistently grouped together by hierarchical clustering (**Fig. 7a**). Moreover, replicate experiments with comparable biochemical treatments and initial cell densities were also grouped together, indicating their high similarity. Interestingly, the OHT-treated conditions with growth media and lower initial cell density (500 cells/well) were classified separately from the other growth media conditions, and appeared individual (**Fig. 7b**). This is consistent with the effect of this biochemical treatment, since OHT-treatment to induce Snail and EMT results in enhanced motility, slower proliferation, and downregulated cell-cell junctions, particularly at lower initial cell densities.

In comparison, experimental conditions cultured with assay media (0.075 ng/mL EGF) exhibited lower cell densities after 60 h, and were also grouped together by hierarchical clustering. DMSO-treated cells at lower initial cell density (500 cells/well) typically organized into morphologically compact clusters that were spatially well separated (**Fig. 7c**). In comparison, OHT-treated cells were grouped together and displayed spanning, dendritic architectures at both initial cell densities (500, 1000 cells/well), consistent with our previous results (**Fig. 7d**).[18] Finally, DMSO-treated cells at higher initial cell densities (1000 cells/well) also formed spanning, dendritic architectures (**Fig. 7d**). It should be noted that this analysis is based on the cell nuclear positions only, whereas the cell morphology in the experiments was highly elongated. Thus, cells could connect together into spanning networks over longer distances than a typical epithelial cell length.

As a more challenging test of TDA, we considered a comparison of experimental conditions cultured with assay media (0.075 ng/mL EGF) relative to treatment with gefitinib (500 nM), which inhibits downstream signaling of the EGFR pathway.[45] Our previous experiments showed that gefitinib treatment results in qualitatively similar spanning configurations, albeit with slightly faster proliferation relative to assay media. Hierarchical clustering grouped experimental conditions by assay media or by gefitinib treatment, respectively (**Fig. S6**). In assay media, cells typically organized as sparse spanning networks or clusters with elongated branches of single-file cells (**Fig. S6a**). One OHT and gefitnib treated condition (500 cells/well) was grouped with the other assay media conditions, but appeared more consistent with these sparse spanning network morphologies by visual inspection. In comparison, gefinib treatment also resulted in spanning networks, but the branches were many cells wide (**Fig. S6b**). One OHT and assay media condition (1,000 cells/well) was grouped in with the other gefitinib treated conditions, and also appeared consistent with these wider spanning networks by visual inspection. Thus, hierarchical clustering with TDA is able to distinguish spanning networks with differing morphology due to different biochemical treatments.

4.5 Spatial Connectivity of Individual and Clustered Phases can be Quantified by Random Particle Removal

Our results with self-propelled particle models and experimental data show empirically that TDA can accurately classify individual, spanning, and clustered particle configurations at varying population size. In order to gain further mechanistic insight into how Wasserstein distances compare different topological structures, we systematically varied the particle number for a given configuration. For example, a particle configuration representing an individual phase initially had 200 particles (**Fig. 8a**). One random particle was then removed, and the Wasserstein distance of this new configuration was computed relative to the initial particle configuration. This process was iterated repeatedly until 180 particles were removed, leaving 20 particles present. Moreover, this process of particle removal was also implemented for spanning and clustered configurations with 200 initial particles (**Fig. 8a**). It should be noted that these particle configurations remained visually quite similar even after up to 125 particles have been removed, and were roughly recognizable as individual, spanning, and clustered phases.

Quantitatively, the Wasserstein distance for the individual phase increased with the number of particles removed, saturating at ≈ 3.5 for 150 particles removed (**Fig. 8b**). In comparison, the spanning phase saturated at at ≈ 2.5, while the clustered phase saturated at ≈ 1. These three curves were well separated, suggesting that they could be used as an alternative classifier for these different phases. To investigate this possibility, particle removal was applied to simulations at different timepoints. All simulations were initialized so that particles were randomly positioned, but rejection sampling was used to ensure that particles were not placed too close together. As a

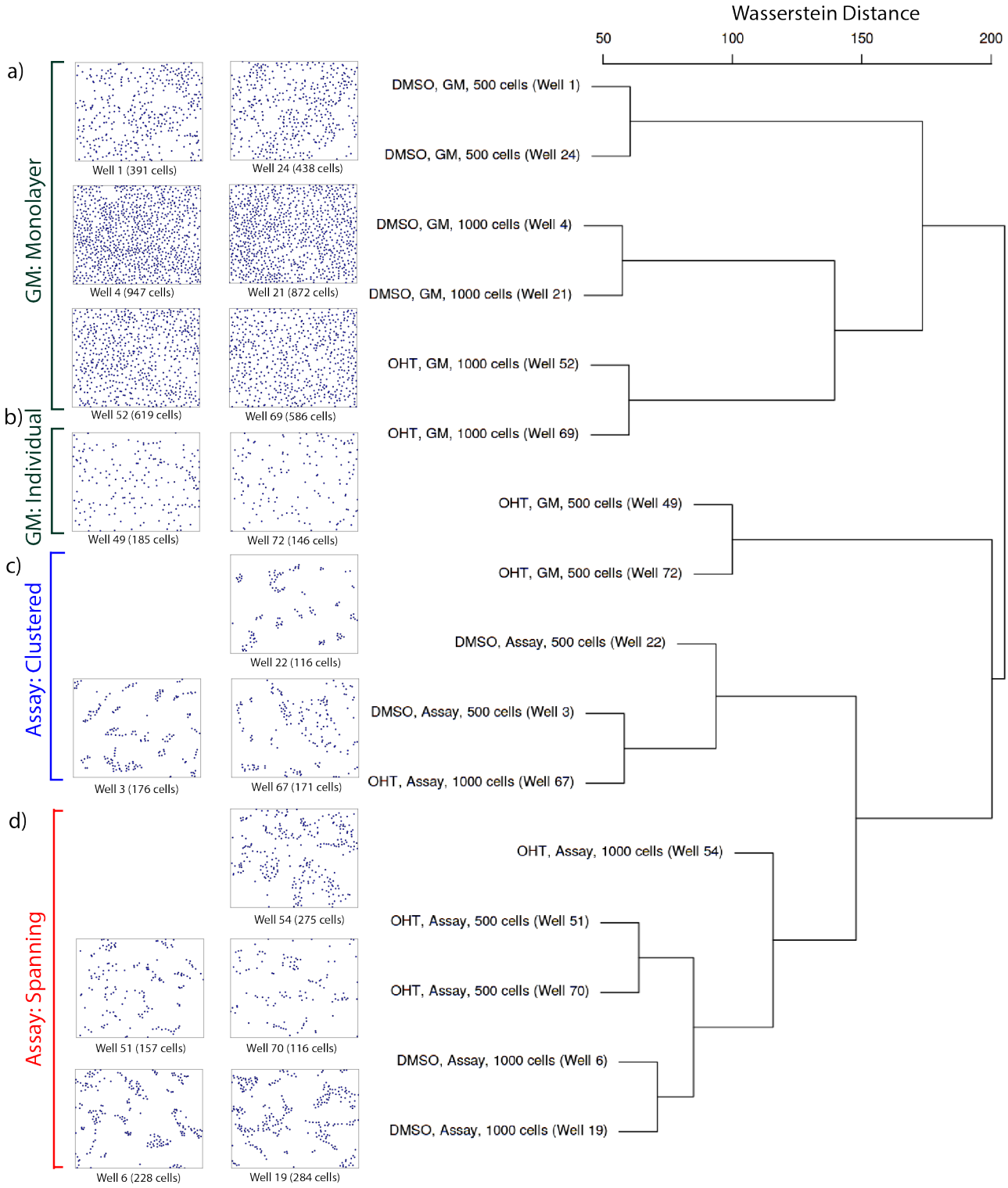

Figure 7: **Classification of experimental conditions based on pairwise Wasserstein distance groups similar experimental conditions (e.g. cell density, biochemical treatment)**. "DMSO" treatment corresponds to an epithelial phenotype, "OHT" treatment corresponds to an induced EMT phenotype, "GM" correspnds to growth media with 20 ng/mL EGF, "Assay" corresponds to assay media with 0.075 ng/mL EGF. Cells were seeded at initial densities of 500 or 1000 cells per well.

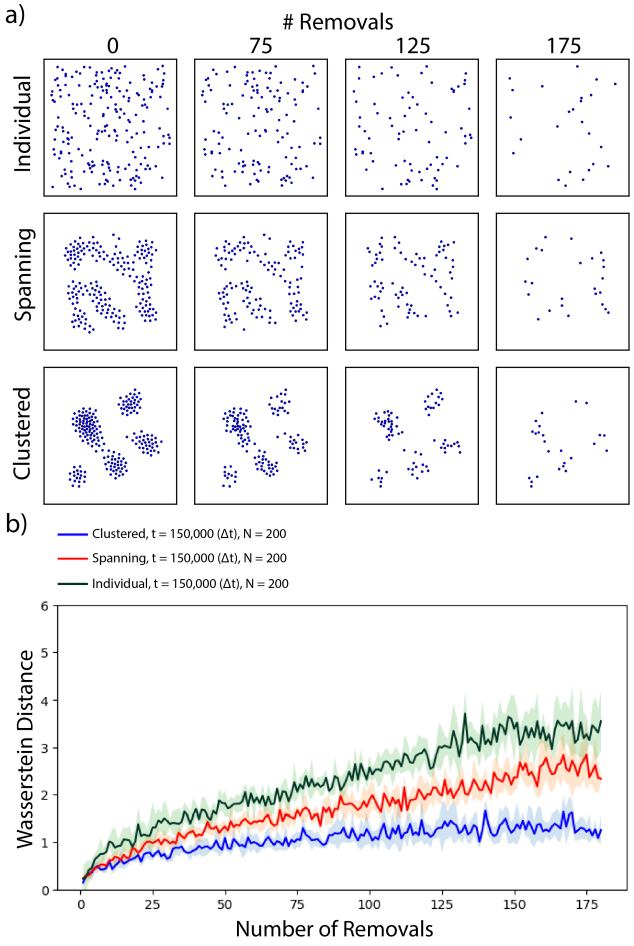

Figure 8: **Spatial connectivity of individual and clustered phases at constant population size can be quantified by random particle removal.** (a) Snapshots for individual, spanning and clustered configurations with random particles removed. (b) Wasserstein distance was computed by randomly removing points from the final simulation state (at $t = 150,000(\Delta t)$) and comparing to the configuration without any removals, consisiting of 200 particles. Colored lines indicate mean Wasserstein distance and shaded region indicates standard deviation for 5 replicates.

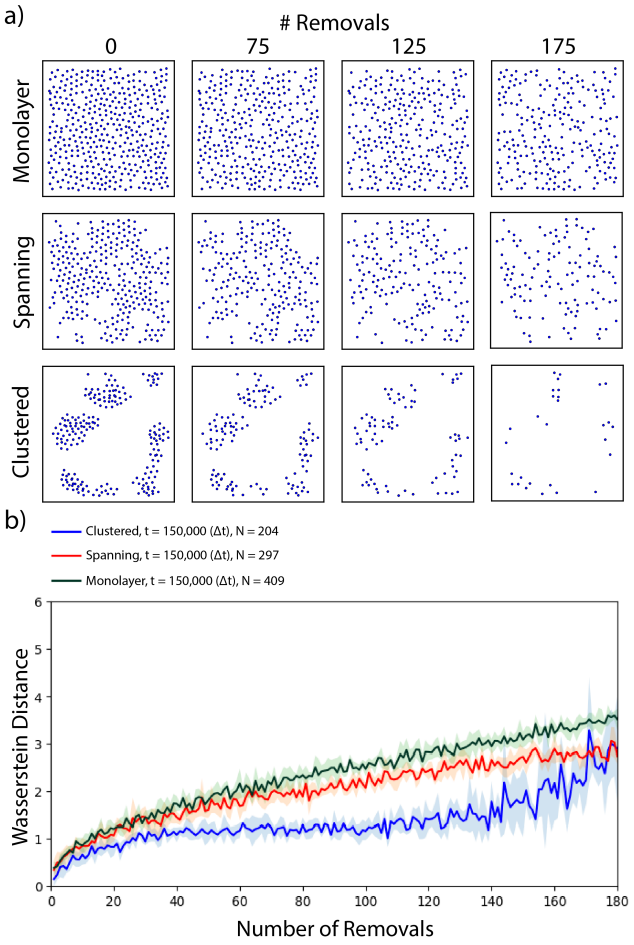

Figure 9: **Spatial connectivity of spanning and clustered phases at varying population size can be quantified by random particle removal.** (a) Snapshots for monolayer, spanning and clustered configurations with random particles removed. (b) Wasserstein distance was computed by randomly removing points from the final simulation state (at $t = 150,000(\Delta t)$) and comparing to the configuration without any removals. Colored lines indicate mean Wasserstein distance and shaded region indicates standard deviation for 5 replicates.

consequence, all simulations exhibit similar removal curves at $t = 0\Delta t$ (**Fig. S7a-c**). However, for the individual simulation configuration, these removal curves rapidly converged to similar scaling from $t = 30,000\Delta t - 150,000\Delta t$, showing that the simulation maintained similar topological structure (equivalent to random placement of particles) over time (**Fig. S7a**). In comparison, for the spanning simulation configuration, these removal curves were more consistent with an individual phase at earlier times $t = 30,000\Delta t - 60,000\Delta t$, but converged towards lower values representing a spanning phase at later times $t = 90,000\Delta t - 150,000\Delta t$ (**Fig. S7b**). Finally, for the clustered simulation, these removal curves again transitioned through an individual phase at earlier times $t = 30,000\Delta t - 60,000\Delta t$, but converged towards the lowest values representing a clustered phase at later times $t = 90,000\Delta t - 150,000\Delta t$ (**Fig. S7c**).

For proliferating particles, random particle removal from monolayer ($n = 409$), spanning ($n = 297$), and clustered ($n = 204$) configurations also resulted in recognizably similar topological structures, up to at least 125 particles removed (**Fig. 9a**). Quantitatively, monolayer and spanning phases also saturated at Wasserstein distances of ≈ 3.5 and ≈ 2.5, respectively (**Fig. 9b**), similar to the previous non-proliferative case (**Fig. 8b**). Nevertheless, the clustered phase saturated at ≈ 1 after about 50 particle removals, but then exhibited a non-monotonic increase at about 140 particle removals (**Fig. 9b**). Based on the snapshots of particle configurations at varying times, this sharp increase can be attributed to the complete removal of a multi-particle cluster, which dramatically affects the topological structure relative to the initial configuration. This result suggests that the similarity between two particle configurations (i.e. persistent homology) will be bounded by the number of particle removals needed to erase the smallest multi-particle cluster initially present.

Finally, random particle removal was applied to the experimental data on epithelial cells.[18] Nuclei positions were normalized to fit inside a $[-10, 10] \times [-10, 10]$ box to maintain consistency in spatial scaling with simulations. For representative individual, spanning, and clustered phases with 200 cells in the field of view, particles remained in recognizable configurations for at least 125 particle removals (**Fig. S8a**). Based on the particle removal curves, individual phases again saturated at a Wasserstein distance of ≈ 3.5 (**Fig. S8b**), similar to the particle removal curves calculated from self-propelled particle simulations (**Fig. 9b**). Nevertheless, the particle removal curves for spanning and clustered phases were not distinguishable, since they both saturated at f ≈ 2.5 (**Fig. S8b**). Based on visual inspection, there are regions of the clustered phases that could plausibly be connected spatially, suggesting that this particle configuration was approaching a percolation threshold. Based on this geometry, it is plausible that the topological structure could not be distinguished from a fully percolating structure based on random particle removal.

5 Discussion and Conclusion

This classification approach based on topological barcodes is robust to noise, which is particularly useful for active matter systems that are far from thermodynamic equilibrium and are driven by energy dissipation towards various states of (dis)order.[1] Although phase transitions can occur in active matter that are analogous to those in equilibrium soft matter systems, it may not be obvious *a priori* how to select the most appropriate order parameters that would capture this transition. Indeed, active matter systems may exhibit distinct self-organized phases and transitions that do not occur in an analogous soft matter system. It should be noted that our approach only considers the topological barcode at the completion of the simulation, and does not consider dynamics. Nevertheless, temporally varying topological barcodes have been previously demonstrated elsewhere (at constant particle number),[37–39] and could be implemented to provide additional insights into particle dynamics. Indeed, TDA could enable efficient sampling of time-series data to identify events of interest across varying simulation parameters. In the future, we envision that TDA could be generalized across different types of propulsion mechanisms and interparticle interactions to infer unifying principles for self-organization.[46]

The minimal model of epithelial cells as self-propelled particles also neglects many interesting biological mechanisms that also drive collective migration. For instance, this model does not consider cell shape changes,[47] which can affect cell-cell interactions as well as motility. Moreover, this model does not address the sensing or release of soluble biochemical signals, which can also function to recruit or repel cells through directed migration.[48] One crucial question is whether a population of cells can truly be treated as homogeneous, due to genetic and non-genetic heterogeneity that is manifested at the single cell level.[49] Indeed, mixtures of two different cell types can exhibit fascinating self-sorting behaviors, which would not be observed with either cell type alone.[9,15] Moreover, cells may alter their migration phenotype over time, such as a epithelial-mesenchymal transition from clustered epithelial cells to individual mesenchymal cells.[6] There is extensive interest in the emergence of "leader cells" that exhibit a partial EMT, allowing collective guidance of mechanically connected followers.[18,50,51] The application of TDA to elucidate biological heterogeneity in an experimental and computational context also represents a fruitful direction for further work.

Finally, we have shown that this classifier remains highly effective over considerable changes in population size, based on both a proliferating self-propelled particle model as well as our experimental data. We elucidate this by systematically varying population size with an initial configuration that is altered by random particle removal. Remarkably, the dependence of Wasserstein distance on particle removal exhibits a characteristic scaling for individual, spanning, and clustered phases. Based on this result, we suggest that the topological structure of clusters can be inferred from relatively sparse sampling of the associated particles. Put another way, TDA visualizes how the particles reside on some lower-dimensional manifold, which can often be determined from a reduced number. For individual particle configurations, we conjecture that TDA classification will remain robust until a threshold number of particles is removed that significantly perturbs the average interparticle spacing, likely at least half the particles. In comparison, for clustered phases, TDA classification is even more robust since a cluster can be defined based on only a few particles in close proximity. In this scenario, we argue that the classifier is bounded by the number of removals needed to destroy the smallest cluster present. It should be noted that this random particle removal also successfully identified spanning phases that are intermediate between individuals and clusters. Nevertheless, the classifier was less effective for distinguishing experimental cell positions for a spanning configuration relative to clusters with extended branches. TDA may have difficulty with this last scenario since the topology of a percolating network is only subtly different from that of branching

clusters near percolation. For these weakly connected structures, a less random particle removal strategy may be more effective to distinguish different phases, such as bootstrap percolation.[52]

In conclusion, we demonstrate that TDA can successfully classify spatial patterns of individuals and clusters in a robust and unbiased fashion. First, we investigate the emergence of individual and clustered phases for self-propelled particles that polarize in random directions and exhibit some attractive interaction, at long times when the particle dynamics approach steady-state. We show that at constant particle density, pairwise Wasserstein distance is sufficient to classify distinct individual, spanning, and clustered phases. Next, we show that this approach also holds for proliferating, self-propelled particles, which exhibit spanning and clustered phases with varying population size. We use TDA to classify patterns of epithelial cells after varying biochemical treatment with EGF and Snail induction through OHT, based on our recent experimental measurements.[18] Finally, we explore the dependence of Wasserstein distance as particle configurations are modified by random particle removed, suggesting why TDA is robust to changes in population size. Overall, this topological approach is generic and could be widely applicable to a variety of active biological systems at multiple scales where population size can change significantly.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank S.E. Leggett and Z.J. Neronha for acquiring and processing the experimental data used in this manuscript, which was previously published elsewhere.[18] This work was supported by the National Cancer Institute's Innovative Molecular Analysis Technologies (IMAT) Program (R21CA212932) and a Brown University Data Science Initative Seed Grant. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

References

[1] T. Vicsek and A. Zafeiris, *Phys Rep*, 2012, **517**, 71–140.

[2] W. Xi, T. B. Saw, D. Delacour, C. T. Lim and B. Ladoux, *Nat Rev Mater*, 2018, 1–22.

[3] J. de Rooij, A. Kerstens, G. Danuser, M. A. Schwartz and C. M. Waterman-Storer, *J Cell Biol*, 2005, **171**, 153–164.

[4] D. Loerke, Q. le Duc, I. Blonk, A. Kerstens, E. Spanjaard, M. Machacek, G. Danuser and J. de Rooij, *Sci Signal*, 2012, **5**, rs5.

[5] V. Maruthamuthu and M. L. Gardel, *Biophys J*, 2014, **107**, 555–563.

[6] I. Y. Wong, S. Javaid, E. A. Wong, S. Perk, D. A. Haber, M. Toner and D. Irimia, *Nature Mat*, 2014, **13**, 1063–1071.

[7] B. Szabó, G. J. Szöllösi, B. Gönci, Z. Jurányi, D. Selmeczi and T. Vicsek, *Phys. Rev. E*, 2006, **74**, 061908–5.

[8] T. E. Angelini, E. Hannezo, X. Trepat, M. Marquez, J. J. Fredberg and D. A. Weitz, *Proc. Natl. Acad. Sci. U.S.A.*, 2011, **108**, 4714–4719.

[9] E. Méhes, E. Mones, V. Németh and T. Vicsek, *PLoS ONE*, 2012, **7**, e31711–13.

[10] M. Suaris, J. A. Breaux, S. P. Zehnder and T. E. Angelini, *AIP Conf Proc*, 2013, **1518**, 536–540.

[11] J.-A. Park, J. H. Kim, D. Bi, J. A. Mitchel, N. T. Qazvini, K. Tantisira, C. Y. Park, M. McGill, S.-H. Kim, B. Gweon, J. Notbohm, R. Steward Jr, S. Burger, S. H. Randell, A. T. Kho, D. T. Tambe, C. Hardin, S. A. Shore, E. Israel, D. A. Weitz, D. J. Tschumperlin, E. P. Henske, S. T. Weiss, M. L. Manning, J. P. Butler, J. M. Drazen and J. J. Fredberg, *Nat Mater*, 2015, **14**, 1040–1048.

[12] D. Bi, J. H. Lopez, J. M. Schwarz and M. L. Manning, *Nat Phys*, 2015, **11**, 1074–1079.

[13] S. Garcia, E. Hannezo, J. Elgeti, J.-F. Joanny, P. Silberzan and N. S. Gov, *Proc Natl Acad Sci USA*, 2015, **112**, 15314–15319.

[14] D. Bi, X. Yang, M. C. Marchetti and M. L. Manning, *Phys. Rev. X*, 2016, **6**, 021011–13.

[15] M. Gamboa Castro, S. E. Leggett and I. Y. Wong, *Soft Matter*, 2016, **12**, 8327–8337.

[16] G. Duclos, C. Erlenkämper, J.-F. Joanny and P. Silberzan, *Nat Phys*, 2016, **13**, 58–62.

[17] L. Atia, D. Bi, Y. Sharma, J. A. Mitchel, B. Gweon, S. Koehler, S. J. DeCamp, B. Lan, J. H. Kim, R. Hirsch, A. F. Pegoraro, K. H. Lee, J. R. Starr, D. A. Weitz, A. C. Martin, J.-A. Park, J. P. Butler and J. J. Fredberg, *Nat Phys*, 2018, **14**, 613–620.

[18] S. E. Leggett, Z. J. Neronha, D. Bhaskar, J. Y. Sim, T. M. Perdikari and I. Y. Wong, *Proc Natl Acad Sci USA*, 2019, **116**, 17298–17306.

[19] J. H. Kim, A. F. Pegoraro, A. Das, S. A. Koehler, S. A. Ujwary, B. Lan, J. A. Mitchel, L. Atia, S. He, K. Wang, D. Bi, M. H. Zaman, J.-A. Park, J. P. Butler, K. H. Lee, J. R. Starr and J. J. Fredberg, *Biochem Biophys Res Commun*, 2020, **521**, 706 – 715.

[20] B. A. Camley and W.-J. Rappel, *J. Phys. D: Appl. Phys.*, 2017, **50**, 113002.

[21] J. M. Belmonte, G. L. Thomas, L. G. Brunnet, R. M. C. de Almeida and H. Chaté, *Phys. Rev. Lett.*, 2008, **100**, 248702.

[22] S. Henkes, Y. Fily and M. C. Marchetti, *Phys. Rev. E*, 2011, **84**, 040301.

[23] N. Sepulveda, L. Petitjean, O. Cochet, E. Grasland-Mongrain, P. Silberzan and V. Hakim, *PLOS Comp Biol*, 2013, **9**, 1–12.

[24] S. S. Soumya, A. Gupta, A. Cugno, L. Deseri, K. Dayal, D. Das, S. Sen and M. M. Inamdar, *PLOS Comp Bio*, 2015, **11**, 1–30.

[25] R. van Drongelen, A. Pal, C. P. Goodrich and T. Idema, *Phys. Rev. E*, 2015, **91**, 032706.

[26] K. Yeo, E. Lushi and P. M. Vlahovska, *Phys. Rev. Lett.*, 2015, **114**, 188301–5.

[27] A. Volkening and B. Sandstede, *Soc JR Interface*, 2015, **12**, 20150812–17.

[28] B. A. Camley, J. Zimmermann, H. Levine and W.-J. Rappel, *PLoS Comput Bio*, 2016, **12**, e1005008–28.

[29] B. A. Camley and W.-J. Rappel, *Proc. Natl. Acad. Sci. U.S.A.*, 2017, **114**, E10074–E10082.

[30] D. A. Matoz-Fernandez, E. Agoritsas, J.-L. Barrat, E. Bertin and K. Martens, *Phys. Rev. Lett.*, 2017, **118**, 158105.

[31] D. A. Matoz-Fernandez, K. Martens, R. Sknepnek, J. L. Barrat and S. Henkes, *Soft Matter*, 2017, **13**, 3205–3212.

[32] M. George, F. Bullo and O. Campàs, *Sci Rep*, 2017, **7**, 9720.

[33] D. R. McCusker, R. van Drongelen and T. Idema, *Europhys Lett*, 2019, **125**, 36001.

[34] M. R. McGuirl, A. Volkening and B. Sandstede, *Proc Natl Acad Sci USA*, 2020, **117**, 5113–5124.

[35] G. Carlsson, *Bull Amer Math Soc*, 2009, **46**, 255–308.

[36] R. Ghrist, *Bull Amer Math Soc*, 2008, **45**, 61–75.

[37] C. M. Topaz, L. Ziegelmeier and T. Halverson, *PLOS ONE*, 2015, **10**, e0126383.

[38] M. Ulmer, L. Ziegelmeier and C. M. Topaz, *PLOS ONE*, 2019, **14**, e0213679.

[39] D. Bhaskar, A. Manhart, J. Milzman, J. T. Nardini, K. M. Storey, C. M. Topaz and L. Ziegelmeier, *Chaos*, 2019, **29**, 123125.

[40] L. Speidel, H. A. Harrington, S. J. Chapman and M. A. Porter, *Phys. Rev. E*, 2018, **98**, 012318.

[41] A. I. McClatchey and A. S. Yap, *Curr Opin Cell Biol*, 2012, **24**, 685–694.

[42] G. Henselman and R. Ghrist, *arXiv: 1606.00199*, 2016.

[43] H. Edelsbrunner, *Computational Topology: An Introduction*, American Mathematical Society, 2009.

[44] J. Vidal, J. Budin and J. Tierny, *IEEE Transactions on Visualization and Computer Graphics*, 2020, **26**, 151–161.

[45] S. Barr, S. Thomson, E. Buck, S. Russo, F. Petti, I. Sujka-Kwok, A. Eyzaguirre, M. Rosenfeld-Franklin, N. W. Gibson, M. Miglarese, D. Epstein, K. K. Iwata and J. D. Haley, *Clin Exp Metastasis*, 2008, **25**, 685–693.

[46] F. Cichos, K. Gustavsson, B. Mehlig and G. Volpe, *Nat Mach Intell*, 2020, 1–10.

[47] S. E. Leggett, J. Y. Sim, J. E. Rubins, Z. J. Neronha, E. K. Williams and I. Y. Wong, *Integr Biol (Camb)*, 2016, **8**, 1133–1144.

[48] B. A. Camley, *J Phys Conden Matter*, 2018, **30**, 223001.

[49] S. J. Altschuler and L. F. Wu, *Cell*, 2010, **141**, 559–563.

[50] M. Reffay, M. C. Parrini, O. Cochet-Escartin, B. Ladoux, A. Buguin, S. Coscoy, F. Amblard, J. Camonis and P. Silberzan, *Nat Cell Biol*, 2014, **16**, 217–223.

[51] M. Vishwakarma, J. Di Russo, D. Probst, U. S. Schwarz, T. Das and J. P. Spatz, *Nat Commun*, 2018, **9**, 3469.

[52] J. Adler, *Physica A*, 1991, **171**, 453 – 470.

6 Electronic Supplemental Information

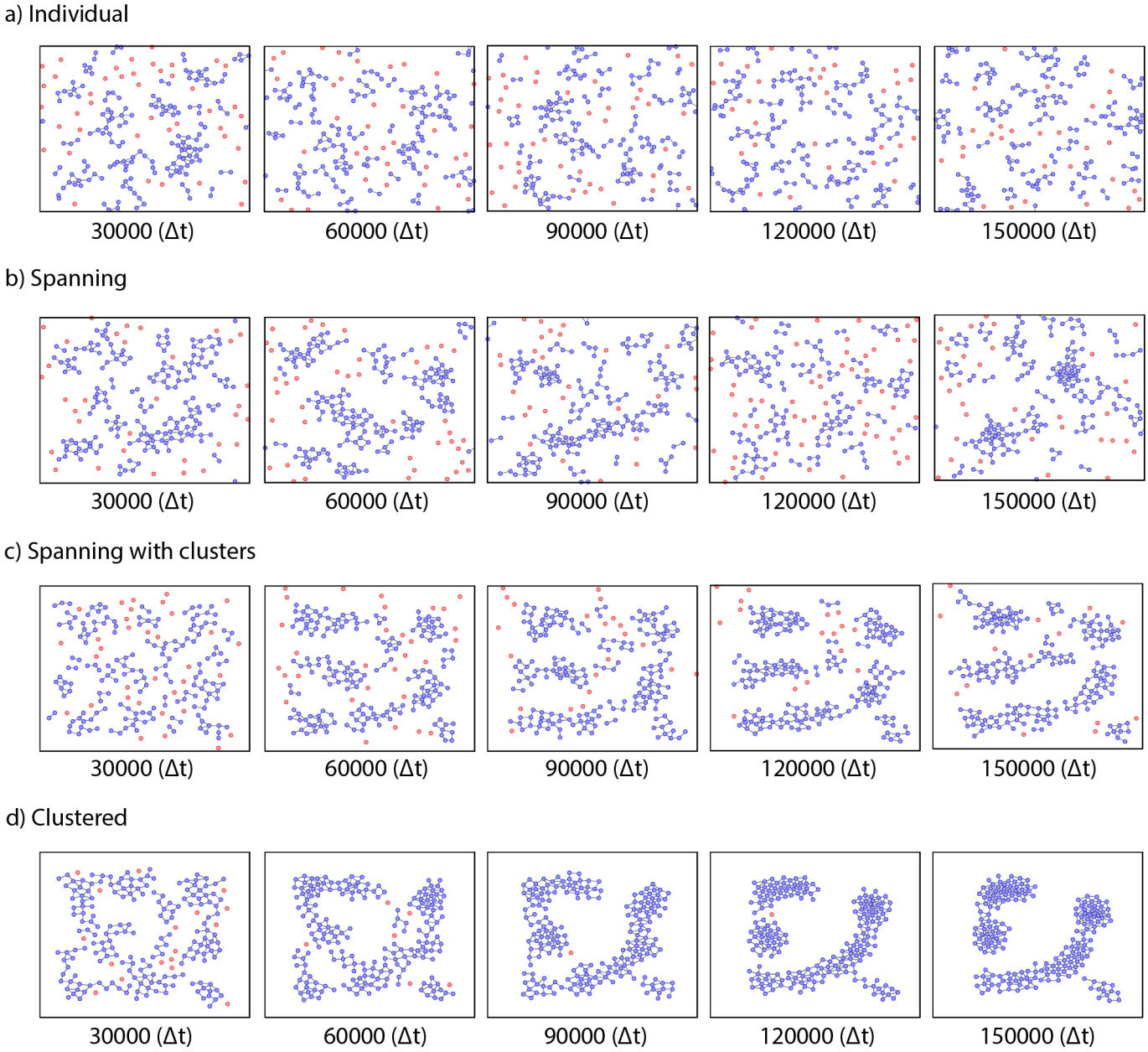

Figure S1: **Self-propelled, non-proliferating particle model results in individual, spanning, and clustered phenotypes with varying adhesion.** Representative snapshots every 30,000 timesteps of individual phases with $\alpha = 0.09, \|\mathbf{P}\| = 0.021$ (a), spanning phases with $\alpha = 0.21, \|\mathbf{P}\| = 0.021$ (b), spanning with clusters phase with $\alpha = 0.07, \|\mathbf{P}\| = 0.007$ (c), and clustered phase with $\alpha = 0.23, \|\mathbf{P}\| = 0.009$ (d), with all simulations starting with the same initial particle position. Particle with one or more neighbors are plotted in blue, with a "bond" drawn between any two cells within radial distance 1.0. Individual cells are shown in red. More individual cells are observed when random polarization dominates over adhesion force. Furthermore, the presence of clusters at low adhesion is transitory and cells are highly motile.

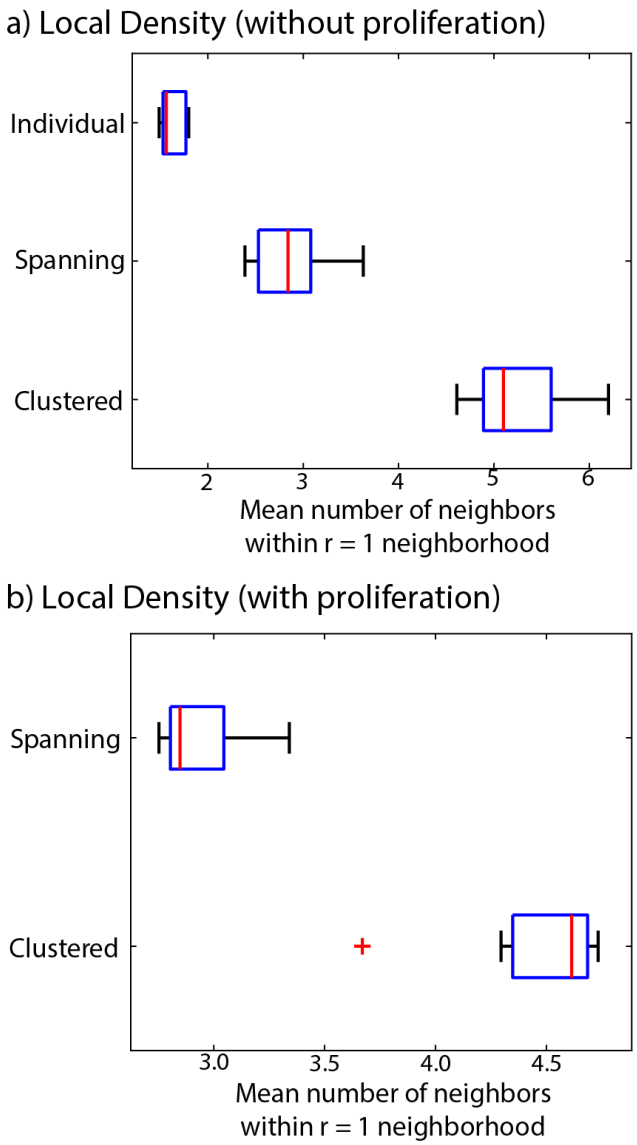

Figure S2: **Ensemble averaged nearest neighbor for individual, spanning and clustered phases.** (a) Statistical distribution of ensemble averaged nearest neighbor count $\langle n \rangle$ for non-proliferating self-propelled particles. Parameters for individual simulations: $\alpha = 0.07, \|\mathbf{P}\| = 0.021$. Parameters for spanning simulations: $\alpha = 0.09, \|\mathbf{P}\| = 0.009$. Parameters for clustered simulations: $\alpha = 0.23, \|\mathbf{P}\| = 0.007$. (b) Statistical distribution of ensemble averaged nearest neighbor count $\langle n \rangle$ for proliferating self-propelled particles. Parameters for spanning simulations: $\alpha = 0.09, \|\mathbf{P}\| = 0.009$. Parameters for clustered simulations: $\alpha = 0.23, \|\mathbf{P}\| = 0.007$. Statistics based on 10 replicates with different initial conditions, but identical parameters. For each boxplot, the central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers.

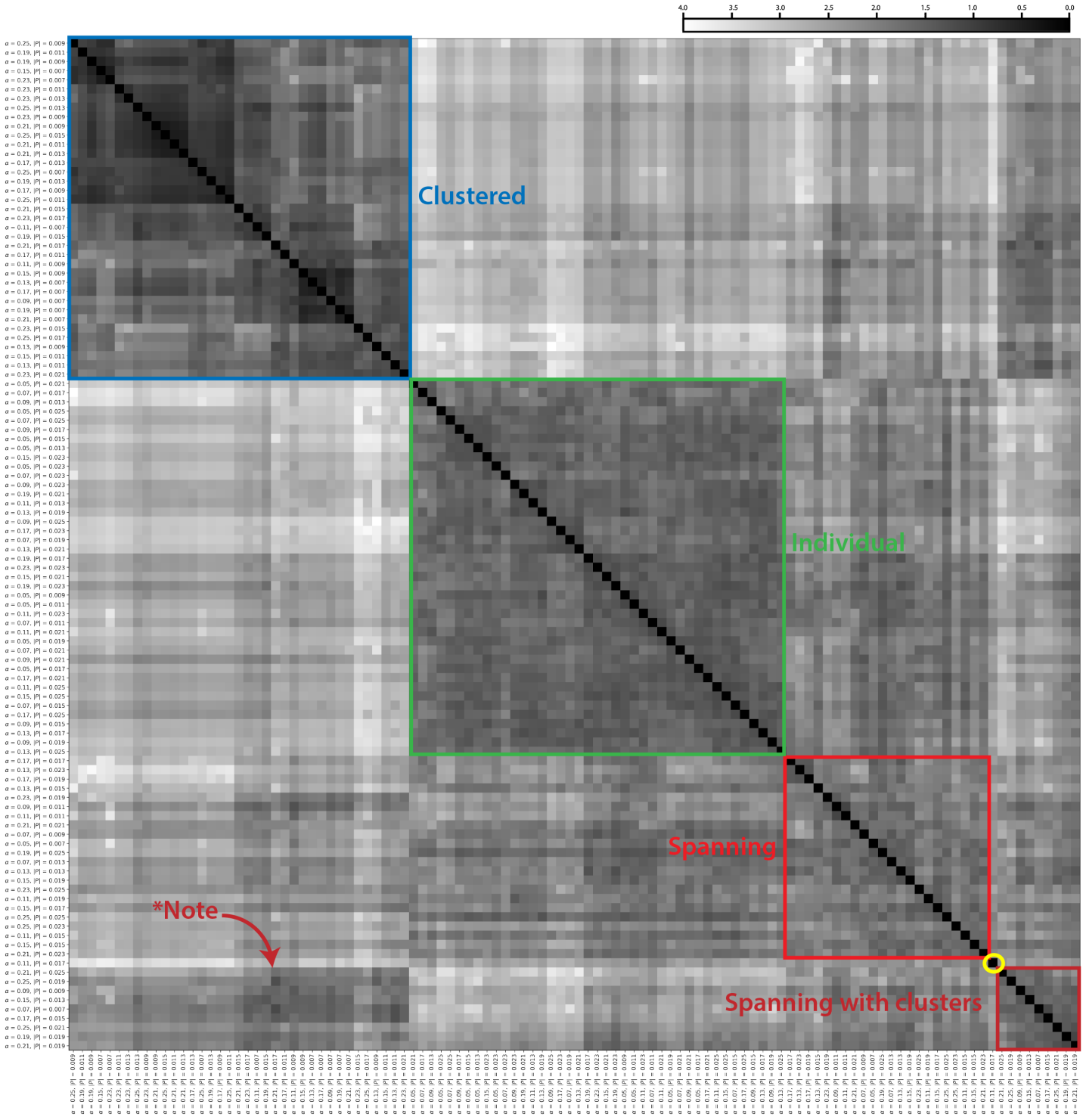

Figure S3: **Pairwise Wasserstein distances between all 121 simulations with no proliferation, as well as varying adhesion and polarization force.** Hierarchical clustering groups clustered, individual, spanning, and "spanning with clusters" phases along the diagonal. *Note that clustered phases exhibit some similarity with the spanning with clusters phase.

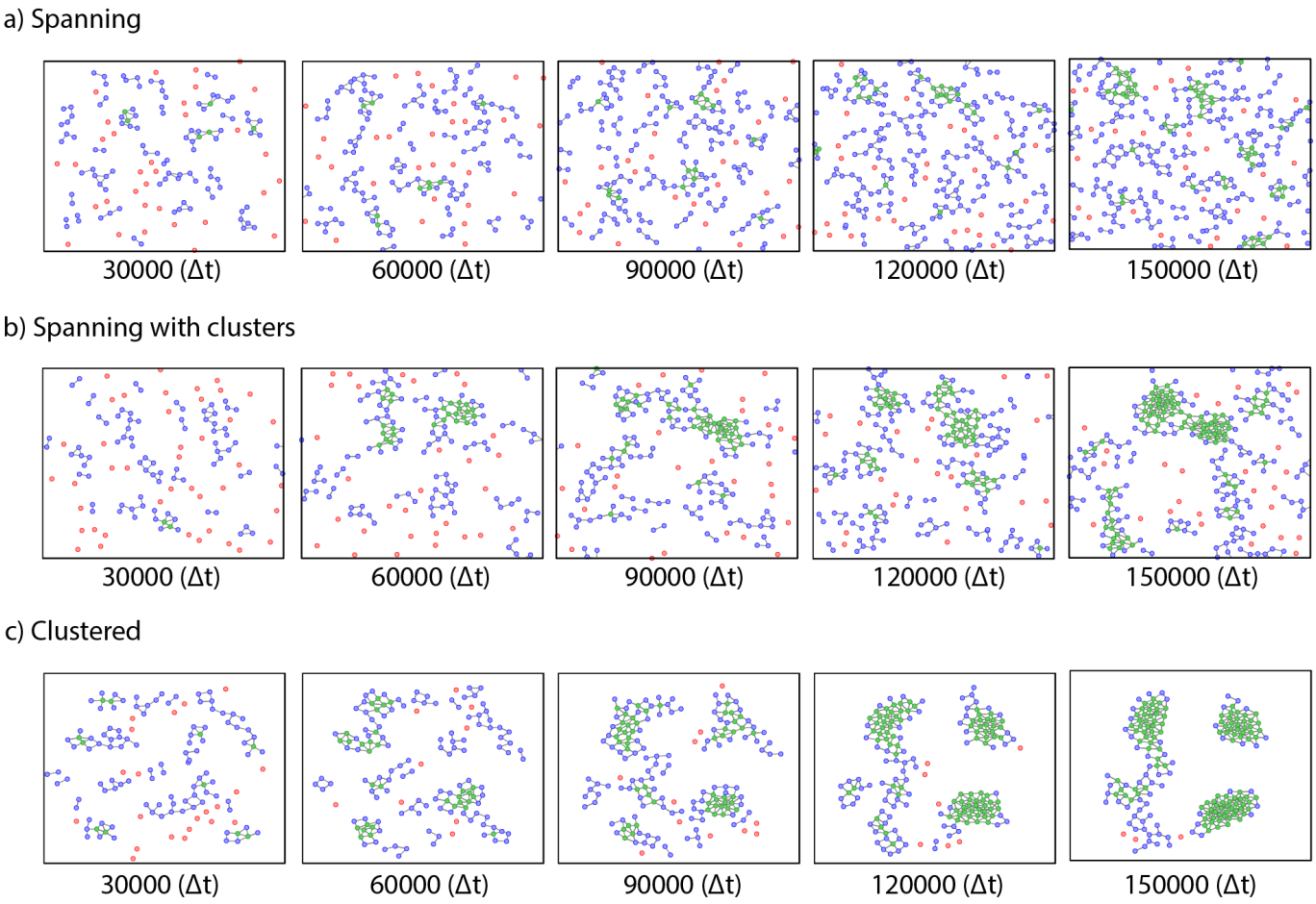

Figure S4: **Self-propelled, proliferating particle model results in spanning, and clustered phenotypes with varying adhesion.** Representative snapshots every 30,000 timesteps of spanning phases with $\alpha = 0.05, \|\mathbf{P}\| = 0.009$ (a), "spanning with clusters" phase with $\alpha = 0.23, \|\mathbf{P}\| = 0.019$ (b), and clustered phase with $\alpha = 0.21, \|\mathbf{P}\| = 0.009$ (c), with all simulations starting with the same initial particle position. Particle with one or more neighbors are plotted in blue, with a "bond" drawn between any two cells within radial distance 1.0. Cells with 4 or more neighbors that cannot proliferate due to contact inhibition of proliferation are shown in green. At low adhesion, cells continue to proliferate until a high cell density is reached, resulting eventually in the formation of a monolayer.

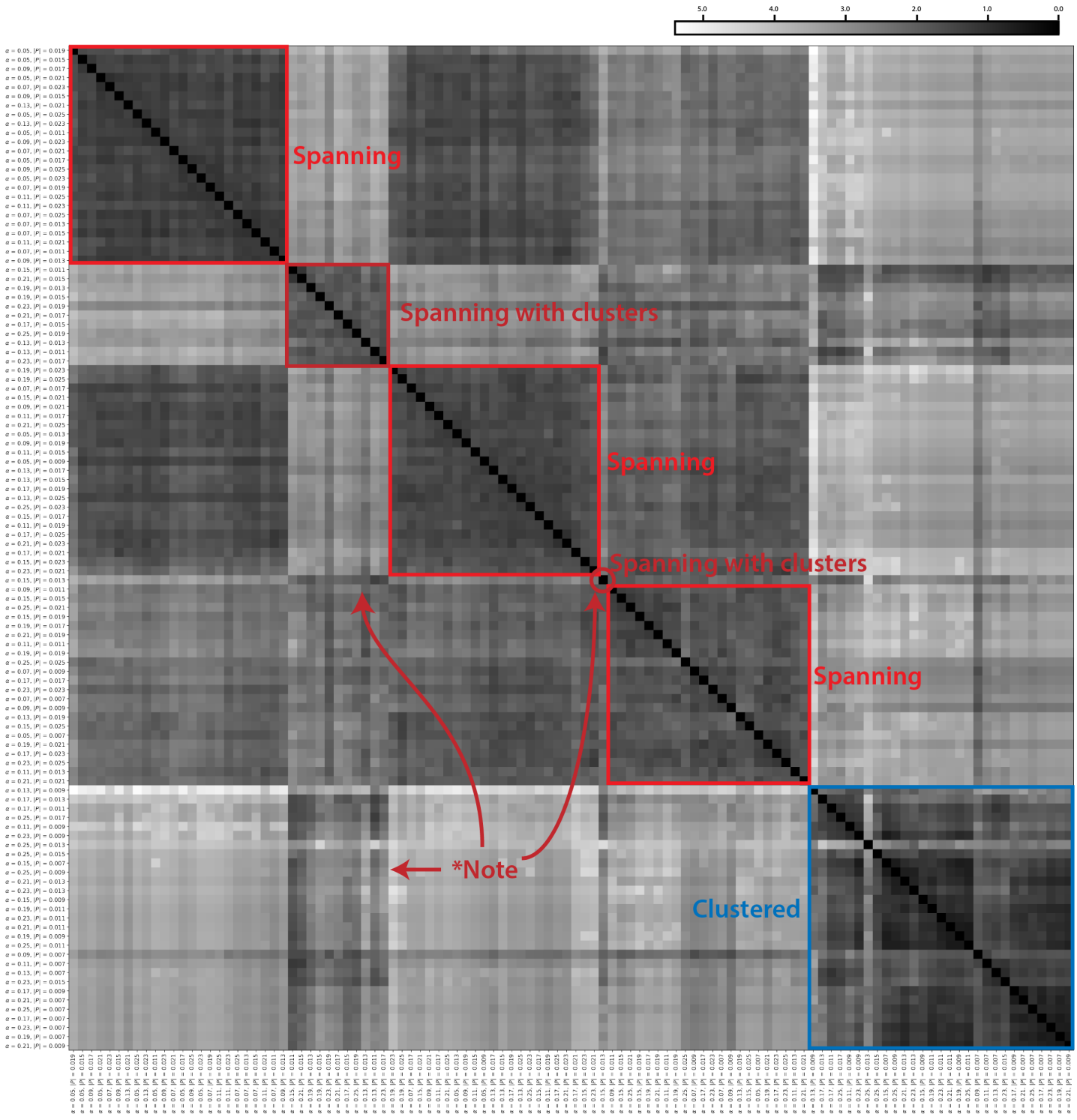

Figure S5: **Pairwise Wasserstein distances between all 121 simulations with proliferation, as well as varying adhesion and polarization force.** Hierarchical clustering groups spanning, "spanning with clusters", and clustered phases along the diagonal. *Note that spanning with clusters phases exhibit some similarity with the clustered phase, as well as one misclassified condition.

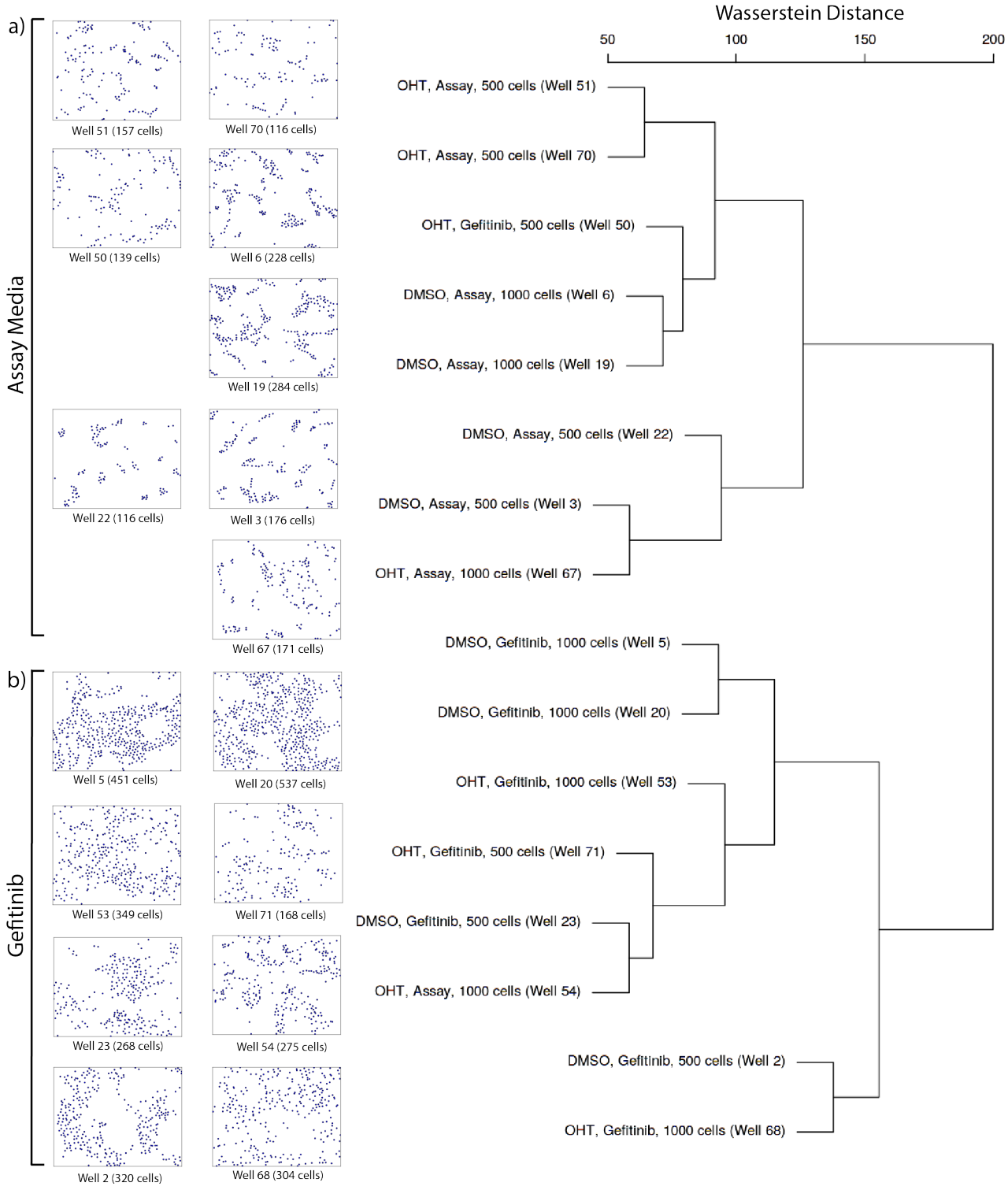

Figure S6: **Hierarchical clustering of pairwise Wasserstein distances between persistence diagrams of experimentally measured cell nuclei positions identifies distinct clustered and spanning phenotypes phases.** (b) Dendrogram obtained by running a hierarchical clustering algorithm using Wasserstein distance groups experimental conditions based on assay media and gefitinib treatment, then biochemical treatment with DMSO or OHT, as well as initial cell density.

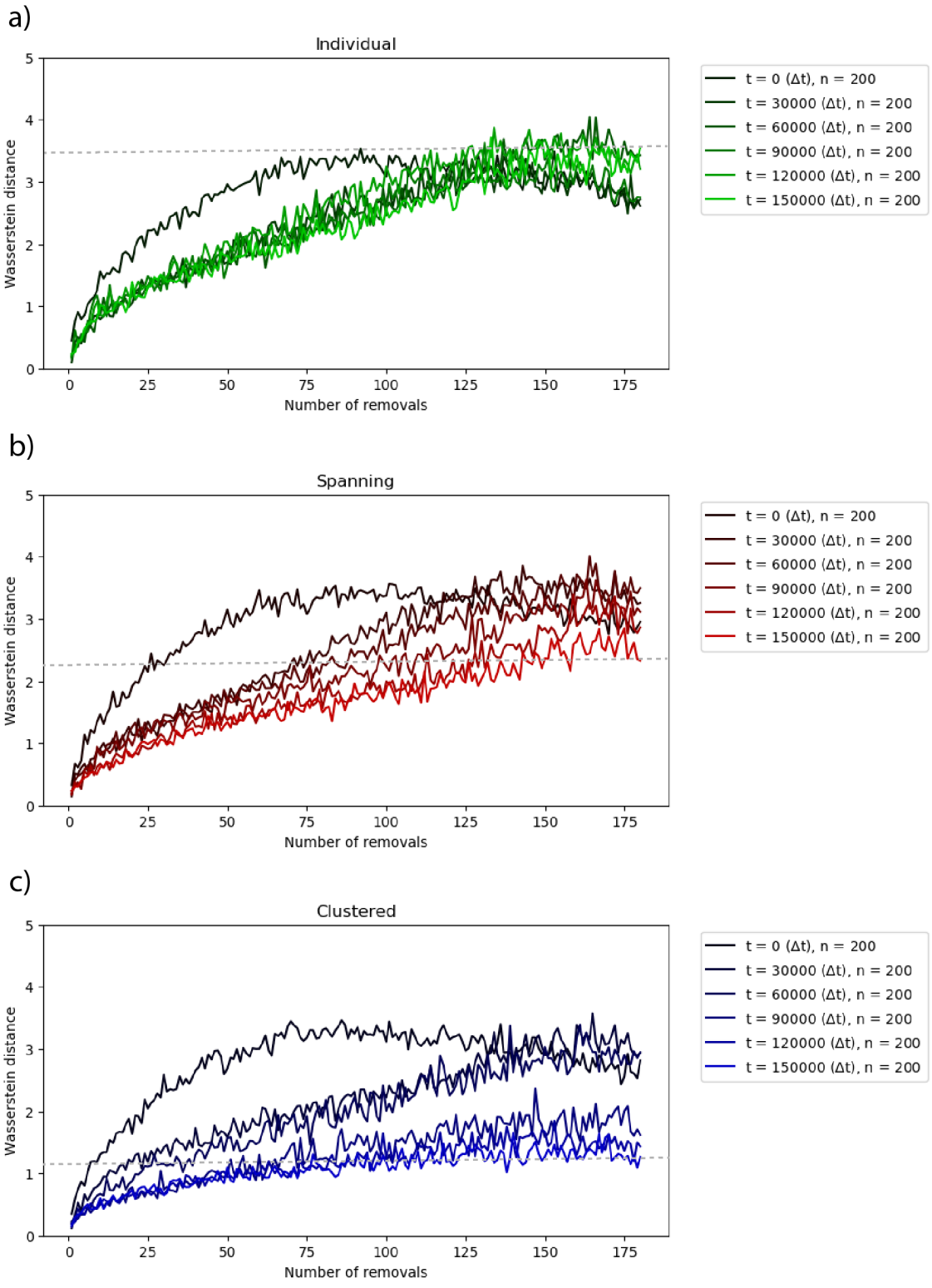

Figure S7: **Computation of Wasserstein distance with random point removal in simulations with non-proliferating self-propelled particles.** Mean Wasserstein distance is computed by randomly removing particles and comparing with reference (containing all particles) for simulations corresponding to individual, spanning and clustered (a-c) phenotypes respectively with fixed population size (n=200) and proliferation disabled. The mean is computed over 5 repetitions for each number of removals.

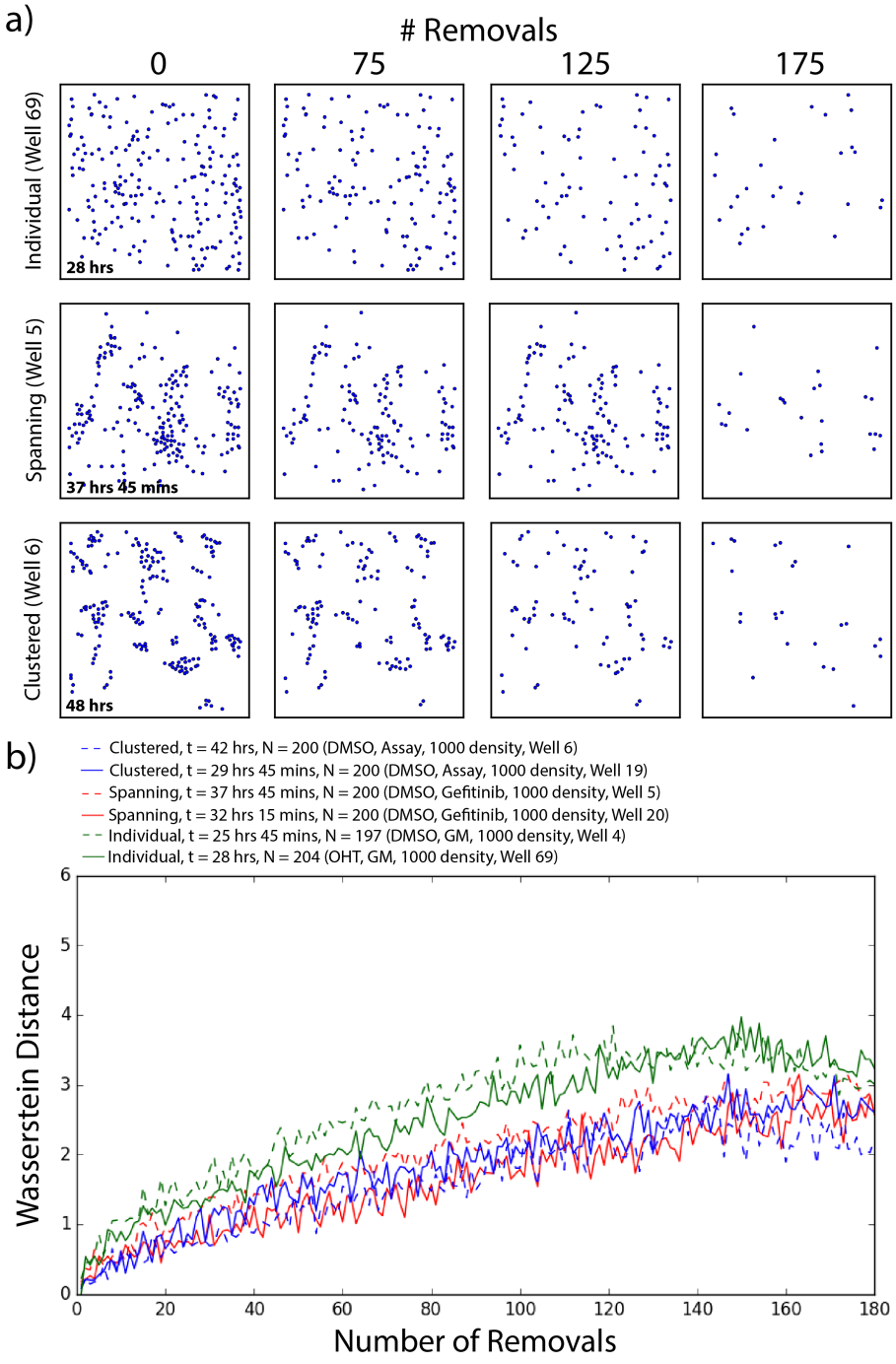

Figure S8: **Random particle removal in experimental conditions.** (a) Representative snapshots corresponding to individual, spanning and clustered phenotypes showing cell nuclei positions with random removals. (b) Mean Wasserstein distance is computed by randomly removing particles and comparing with reference (containing all particles). The mean is computed over 5 repetitions for each number of removals.