6 Electronic Supplemental Information

Supplementary Videos: Note that all simulation ran with random initial positions up to time $t = 4000$ $(200,000$ timesteps$)$.

File	Description	Simulation Parameters
SI_Video1.mp4	Individual phenotype obtained from simulation of self-propelled particle model with proliferation disabled.	$\alpha = 0.07$, $P = 0.021$
SI_Video2.mp4	Branching phenotype obtained from simulation of self-propelled particle model with proliferation disabled.	$\alpha = 0.09$, $P = 0.011$
SI_Video3.mp4	Branching with clusters phenotype obtained from simulation of self-propelled particle model with proliferation disabled.	$\alpha = 0.25$, $P = 0.017$
SI_Video4.mp4	Clustered phenotype obtained from simulation of self-propelled particle model with proliferation disabled.	$\alpha = 0.23$, $P = 0.009$
SI_Video5.mp4	Branching phenotype obtained from simulation of self-propelled particle model with proliferation enabled.	$\alpha = 0.09$, $P = 0.009$
SI_Video6.mp4	Branching with clusters phenotype obtained from simulation of self-propelled particle model with proliferation enabled.	$\alpha = 0.19$, $P = 0.013$
SI_Video7.mp4	Clustered phenotype obtained from simulation of self-propelled particle model with proliferation enabled.	$\alpha = 0.23$, $P = 0.007$

Note S1: Computational of Persistence Diagrams and Wasserstein Distance

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.[59] For example, consider a set of 7 points at varying filtration, illustrated by gray disks with radius ε_1 centered at each point (**Fig. S1a**). As the radius increases to ε_2, certain gray disks overlap, indicating that the corresponding points should be connected by 7 edges (blue lines, dimension 0 homology) at this cutoff distance (**Fig. S1b**). Moreover, these connected edges form 1 closed loop around an empty region (orange circle, dimension 1 homology). A further increase in radius to ε_3 maintains connectivity of the 7 edges, but collapses the closed loop (**Fig. S1c**).

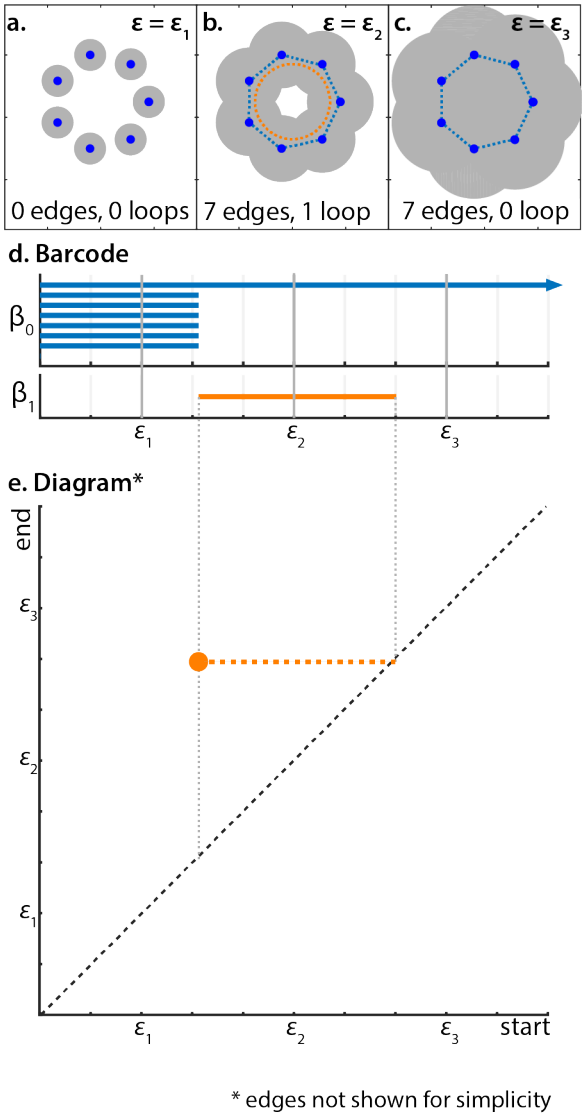

Fig. S1 Computation of persistence homology. (a,b,c) Visualization of connectivity between points at varying values of spatial parameter ε. Edges are colored in blue (dimension 0 homology), and loops are colored in orange (dimension 1 homology) (d) Corresponding topological barcode. (e) Persistence diagram showing loops only for simplicity.

The corresponding topological barcode uses horizontal bars to visualize the persistence of features that appear or disappear at varying ε intervals (**Figure S1d**). 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 appears ("start"), 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,). For example, at ε_1, there are seven distinct blue bars corresponding to the number of discrete points, which are not connected at this length scale. At ε_2, there is only one blue bar, since all points have linked together forming a connected component, and there is the appearance of one closed loop, indicated by an orange bar. At ε_3, there remains only one blue bar and no orange bar, since the closed loop has collapsed. The same information can also be organized in a persistence diagram (start ε values on x-axis and end ε values on y-axis), where the distance from the diagonal is indicative of the significance or importance of a topological feature (**Fig. S1e**). In particular, note that the closed loop is appreciably offset from the diagonal, signifying that it is a relatively stable topological structure for this particle

configuration. For simplicity, edges are not shown in the persistence diagram.

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 off-diagonal points in Y. Points close to 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[60]. Therefore, in practice, projections of off-diagonal points to the diagonal are exchanged between persistence diagrams before matches are obtained (**Fig. S2, a-c**). The Wasserstein distance, $W(X,Y)$ is then defined as the infimum over all possible bijections, s:

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

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

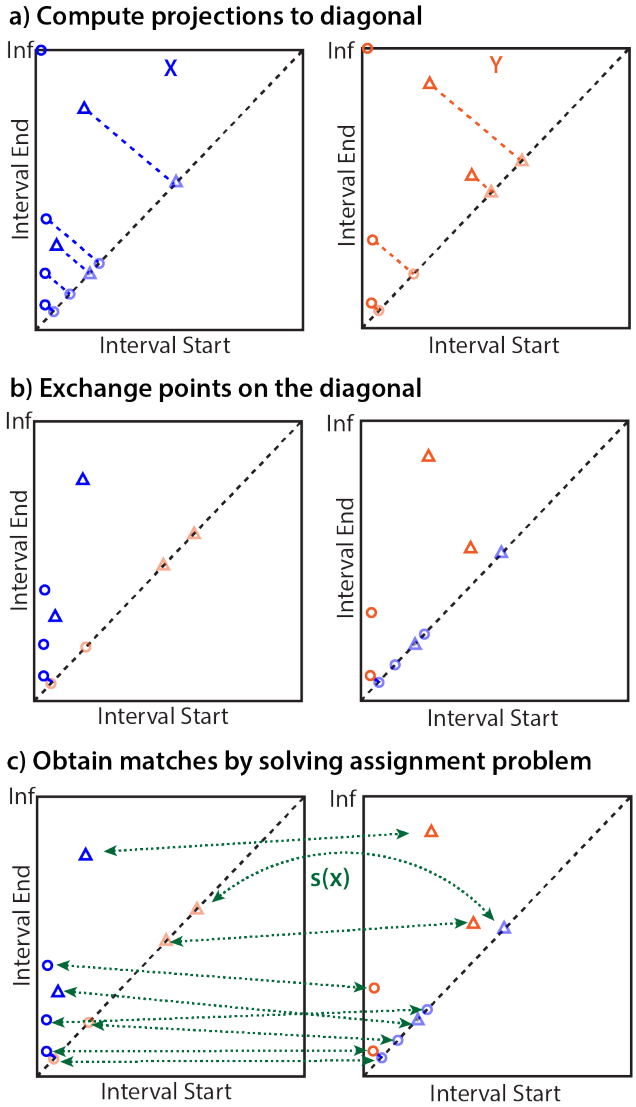

Fig. S2 Computation of Wasserstein distance. (a) Projections of off-diagonal points to the diagonal are computed. Circles represent edges or connected components and triangles represent loops. Point at $(0,\infty)$ representing 1 connected component for high values of spatial parameter ε is not considered. (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.

Note S2: System Parameterization based on Peclet Number and Attraction / Polarity

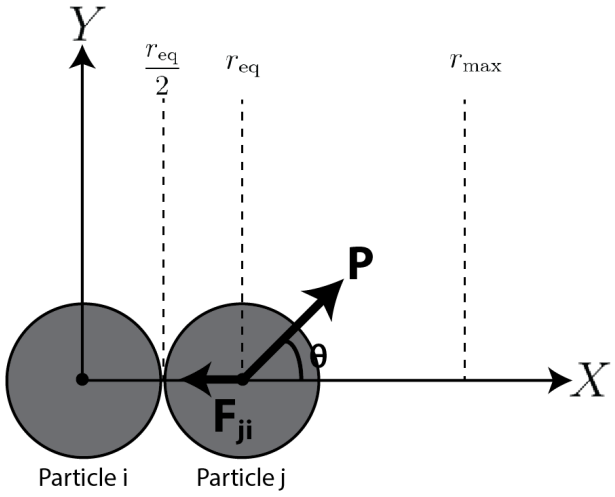

Fig. S3 Schematic for particle-particle interaction in the agent-based model. A particle j is at equilibrium with particle i. A polarization force P is applied to pull particle j to distance $r_{\rm max}$ against adhesion force F_{ij}.

The system is parameterized using two nondimensional variables corresponding to particle self-propulsion and adhesion. First, the Peclet number is defined in terms of the self-propulsion speed P/η, the particle radius r_{eq}, and the characteristic reorientation time τ_p. The particle radius is given by the equilibrium separation distance where the interaction potential is a minimum, that is:

$$|\mathbf{F}_{ij}| = -\left|\frac{dU}{dr}\right| = \alpha\left|\frac{1}{4L_R}e^{-r/L_A} - \frac{1}{L_A}e^{-r/L_A}\right| \quad 0 \le r \le r_{\rm max} \tag{5}$$

$$|\mathbf{F}_{ij}| = 0 \implies r_{\rm eq} = \frac{L_A L_R}{L_A - L_R}\ln\left(\frac{L_A}{4L_R}\right) \approx 1 \tag{6}$$

Thus,

$$\mathrm{Pe} = \frac{\frac{P\tau_P}{\eta}}{r_{\rm eq}} \approx 50P \tag{7}$$

where $\eta = 1$ and $\tau_P = 2500\Delta t = 50$ is the characteristic time to repolarization. The Peclet number ranges between 0.4 and 1.3 for polarization values in our simulation.

Second a non-dimensional adhesion (scaled by self-propulsion), A, is defined that compares the relative strengths of adhesion and self-propulsion. The energetic cost ΔU of breaking a particle-particle bond is determined by moving the particle by a distance Δr from r_{eq} with r_{max}, beyond which the interaction energy is 0. Thus,

$$U_{\rm eq} = U(r_{\rm eq}) = -\alpha\left(\frac{4L_R}{L_A}\right)^{\frac{L_R}{L_A-L_R}} + \frac{\alpha}{4}\left(\frac{4L_R}{L_A}\right)^{\frac{L_A}{L_A-L_R}} \approx -0.897\alpha \tag{8}$$

$$U(r_{\rm max}) = -\alpha e^{-r_{\rm max}/L_A} + \frac{\alpha}{4}e^{-r_{\rm max}/L_R} \approx -0.886\alpha \tag{9}$$

We note that at any given time, the self-propulsion force P is randomly oriented, thus we must average over all possible orientations θ. When the self-propulsion force of the jth particle is oriented away from the ith particle, the effective force acting against the particle-particle bond is given by $P\cos(\theta)$ (i.e. $-\pi/2 < \theta < \pi/2$). However, when the self-propulsive force of the jth particle is oriented towards the ith particle, the effective force is 0. (i.e. $-\pi < \theta < -\pi/2$; $\pi/2 < \theta < \pi$). Thus,

$$\mathrm{A} = \frac{\Delta U}{\Delta r\int_{-\pi}^{\pi} P\cos(\theta)d\theta} = \frac{(U(r_{\rm max}) - U(r_{\rm eq}))\pi}{P(r_{\rm max} - r_{\rm eq})} \approx \frac{0.011\alpha\pi}{P} \tag{10}$$

A ranges between 0.05 and 2.1 for α and P values in our simulation.

Note S3: Runtime performance measurements for Wasserstein distance computation

We computed pairwise Wasserstein distances between 121 simulations corresponding to 11 polarization values (linearly spaced between 0.005 and 0.025) and 11 adhesion values (linearly spaced between 0.05 and 0.25). Computations corresponding to the two lowest polarization values ($P = 0.005, P = 0.007$) were later discarded from the phase diagram since they appeared to be kinetically trapped. Since the resulting distance matrix is symmetric, the Wasserstein distances were computed for the upper triangular part, including the diagonal, consisting of $121 \times (121+1)/2 = 7381$ entries. Parallel computation was performed on 10 CPUs (Intel Core i7-7800X, 3.5 GHz) with the i^{th} core responsible for computing entries in columns $12 \times (i-1)+1$ to $12 \times (i)$. Therefore, core 1 computed entries in columns $1-12$ (78 values), core 2 computed entries in columns $13-24$ (222 values), and so on. Core 10 computed entries in columns $109-121$ (1495 values). In general, the i^{th} core was responsible for computing $144 \times i - 66$ values, for values of i ranging from 1 to 9. Note that the 10^{th} core computed values in 13 columns, whereas all other cores computed values in 12 columns. While this is not the most efficient allocation of work (in an optimal configuration, work would be divided equally among the processors), it is easy to implement and produces results within a reasonable time frame (see Table S1).

Simulation	Barcode Dimension	# Cores	Time (mins)
Without proliferation	0	10	11
Without proliferation	1	10	5
With proliferation	0	10	36
With proliferation	1	10	13

Table S1 Time taken to compute pairwise Wasserstein distances.

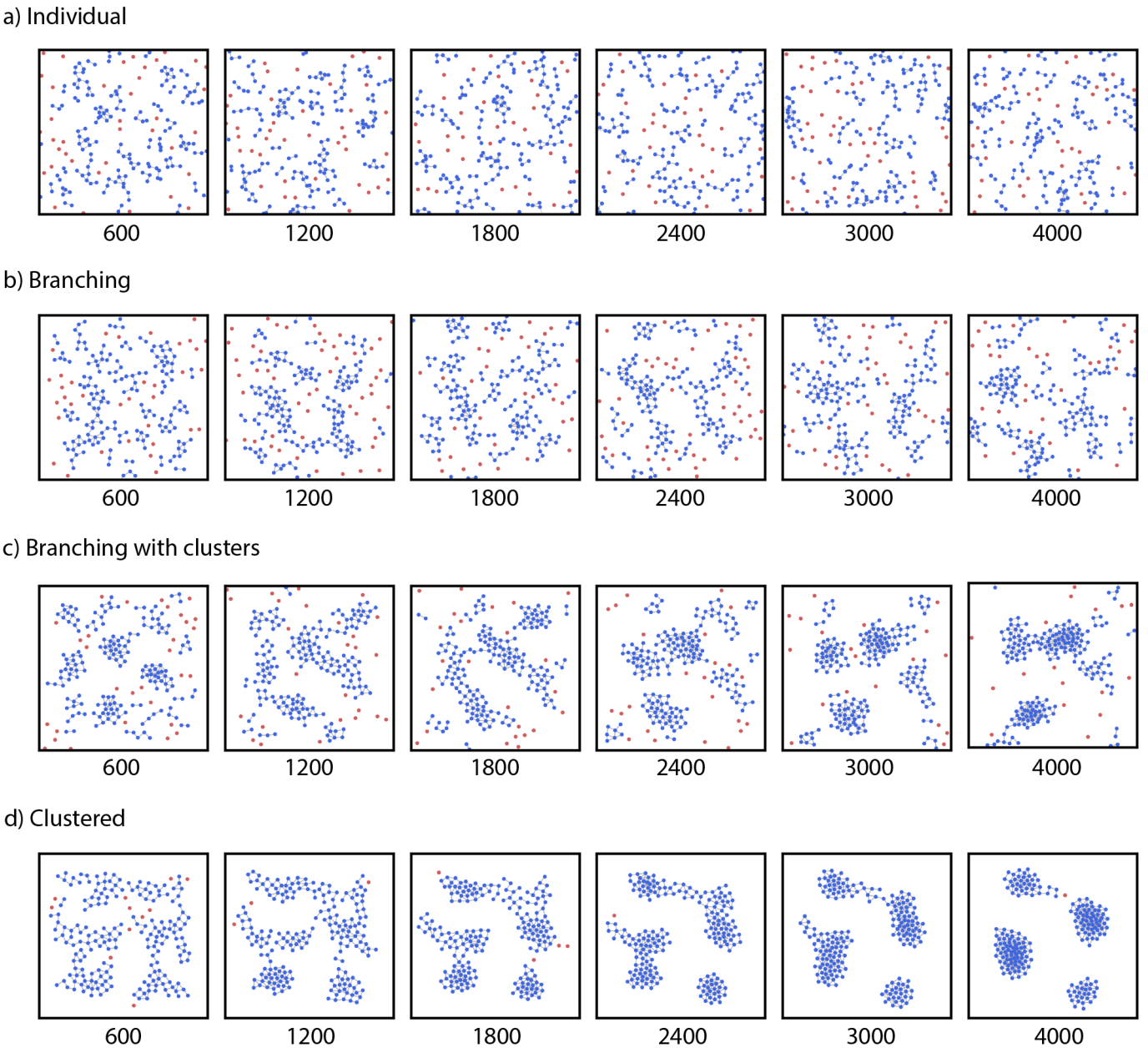

Fig. S4 Non-proliferating particle model results in individual, branching, branching with clusters and clustered phenotypes with varying adhesion and self-propulsion force. Representative snapshots at time intervals of 600 (every 30,000 timesteps, except for the last snapshot) of individual phase with $\alpha = 0.07, P = 0.021$ (a), branching phase with $\alpha = 0.09, P = 0.011$ (b), branching with clusters phase with $\alpha = 0.25, P = 0.017$ (c), and clustered phase with $\alpha = 0.23, P = 0.009$ (d), with all simulations starting with random initial positions. Particle with one or more neighbors are plotted in black, 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.

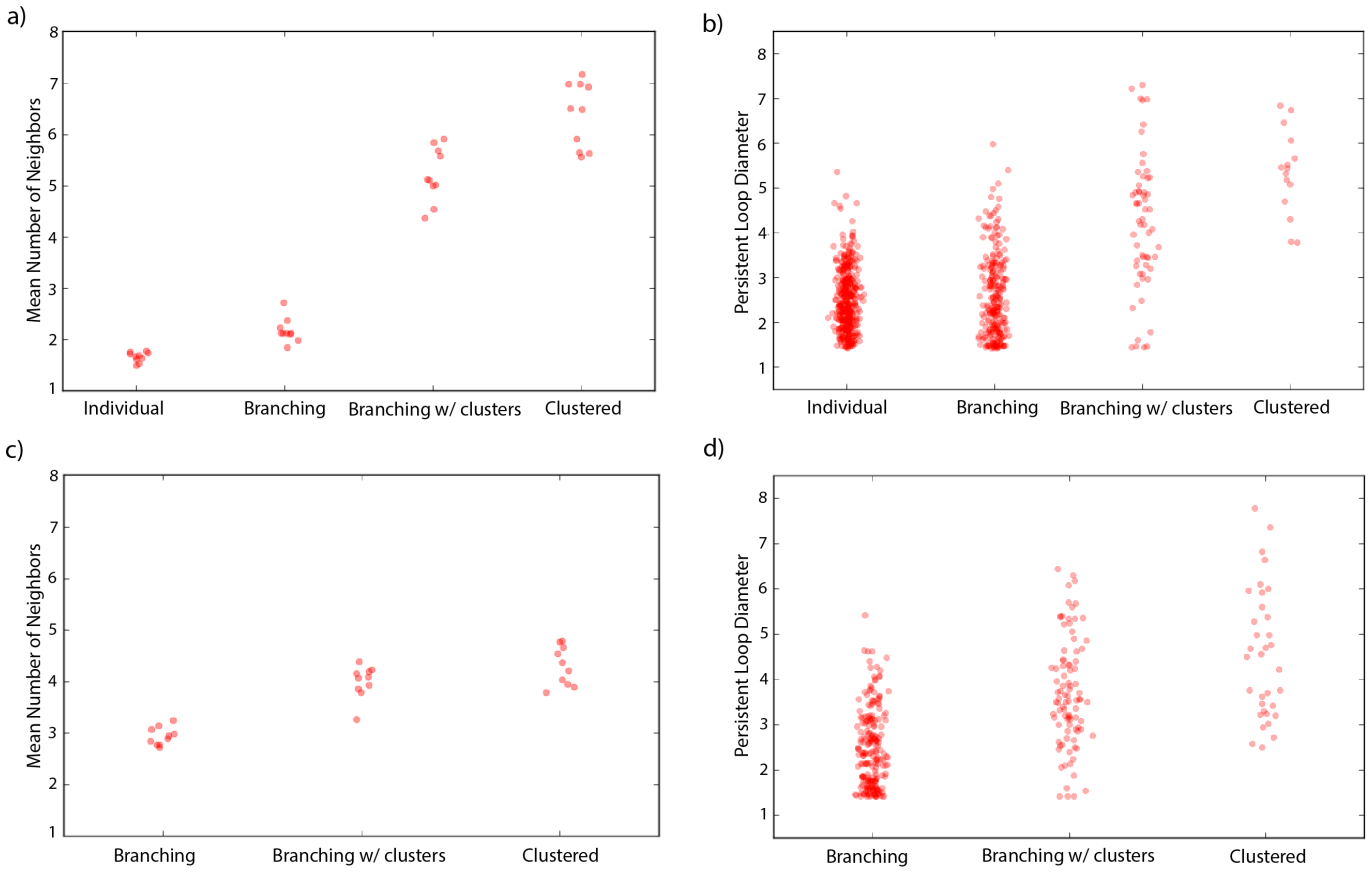

Fig. S5 Comparison of local density (nearest neighbor count) and persistent loop diameter associated with distinct phases. In the top row, data acquired from individual, branching, branching with clusters and clustered simulations performed 10 times each with random initialization and parameters $(\alpha = 0.07, P = 0.021)$, $(\alpha = 0.09, P = 0.011)$, $(\alpha = 0.25, P = 0.017)$ and $(\alpha = 0.23, P = 0.009)$ respectively. In the bottom row, data acquired from branching, branching with clusters and clustered simulations performed 10 times each with random initialization and parameters $(\alpha = 0.09, P = 0.009)$, $(\alpha = 0.19, P = 0.013)$ and $(\alpha = 0.23, P = 0.007)$ respectively. (a) Comparison of nearest neighbor count for individual, branching, and clustered phases at constant population size. (b) Comparison of characteristic loop diameter for individual, branching, and clustered phases at constant population size. (c) Comparison of nearest neighbor counts for branching and clustered phases in proliferating populations. (d) Comparison of characteristic loop diameter for individual, branching, and clustered phases in proliferating populations.

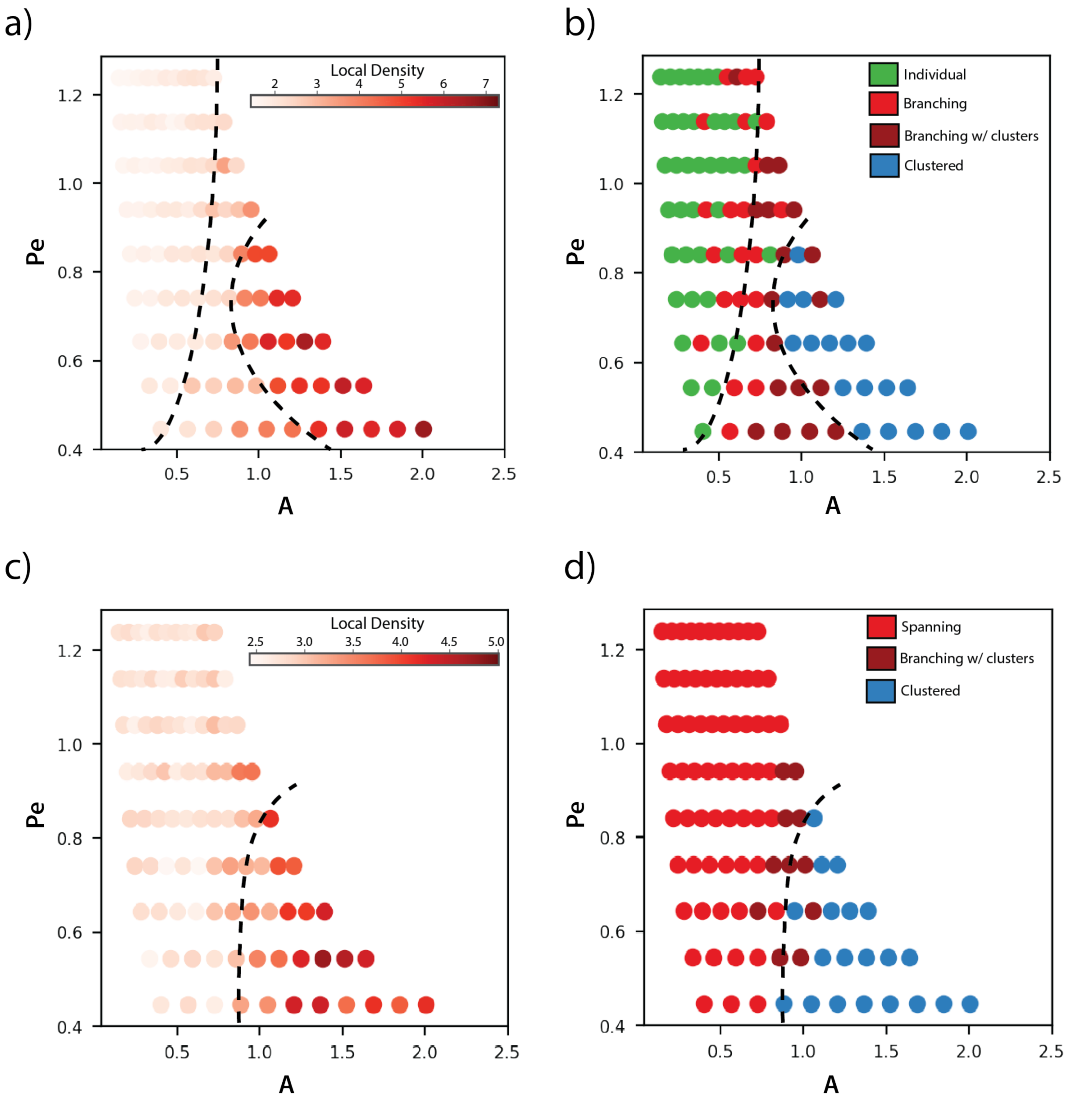

Fig. S6 Phase transitions parameterized by dimensionless parameters for self-propulsion (Pe) and adhesion (A). (a) Local density (nearest neighbor count) for a constant population size. (b) TDA classification of individual, branching, and clustered phases at constant population size. (c) Local density (nearest neighbor count) for a proliferating population. (d) TDA classification of branching and clustered phases for a proliferating population.

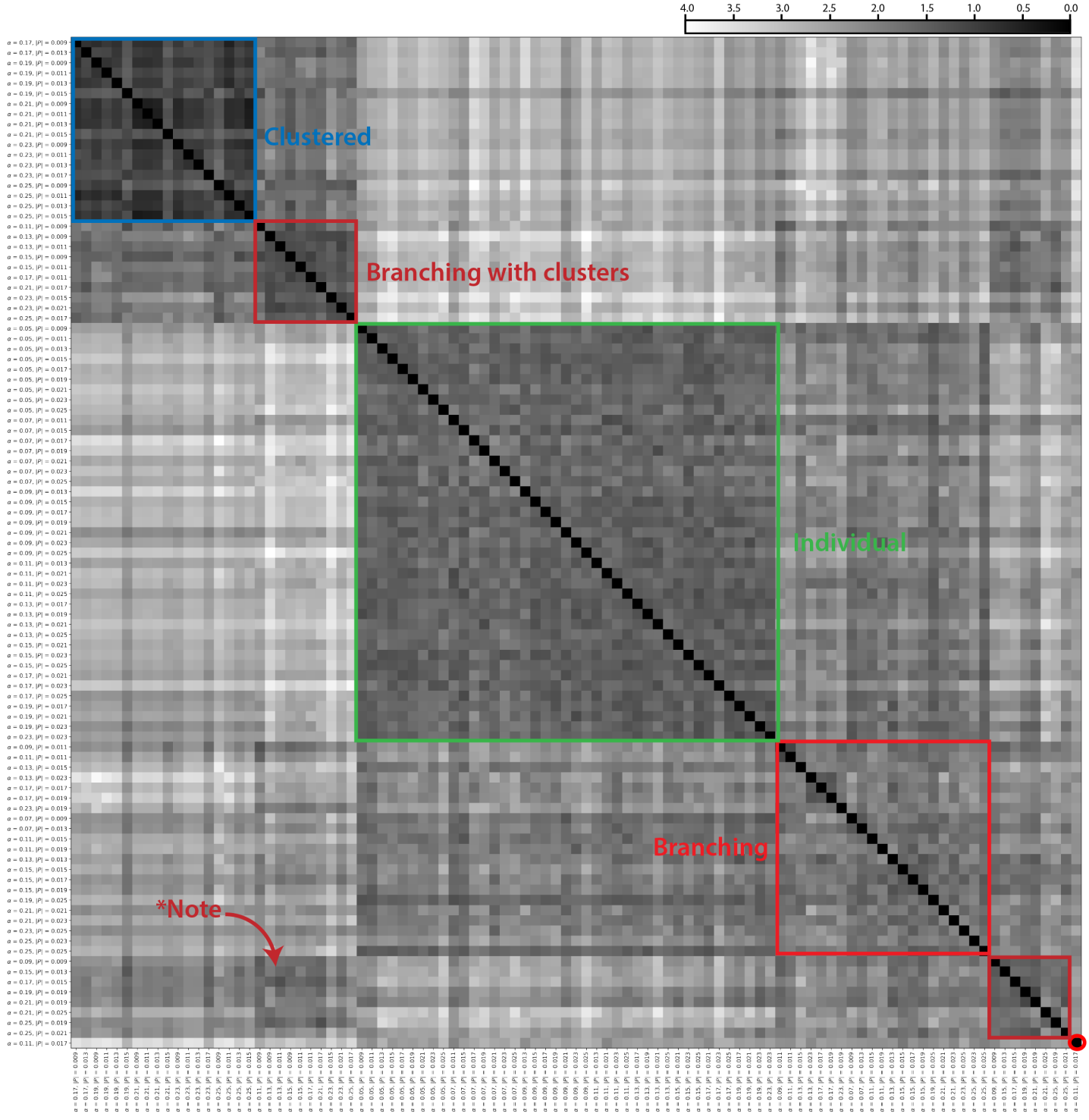

Fig. S7 Pairwise Wasserstein distances between all simulations without proliferation, as well as varying adhesion and self-propulsion force. Hierarchical clustering groups simulations into individual, branching, branching with clusters, and clustered categories corresponding to distinct phases along the diagonal. Note that branching with clusters phase exhibits similarity with both branching and clustered phases.

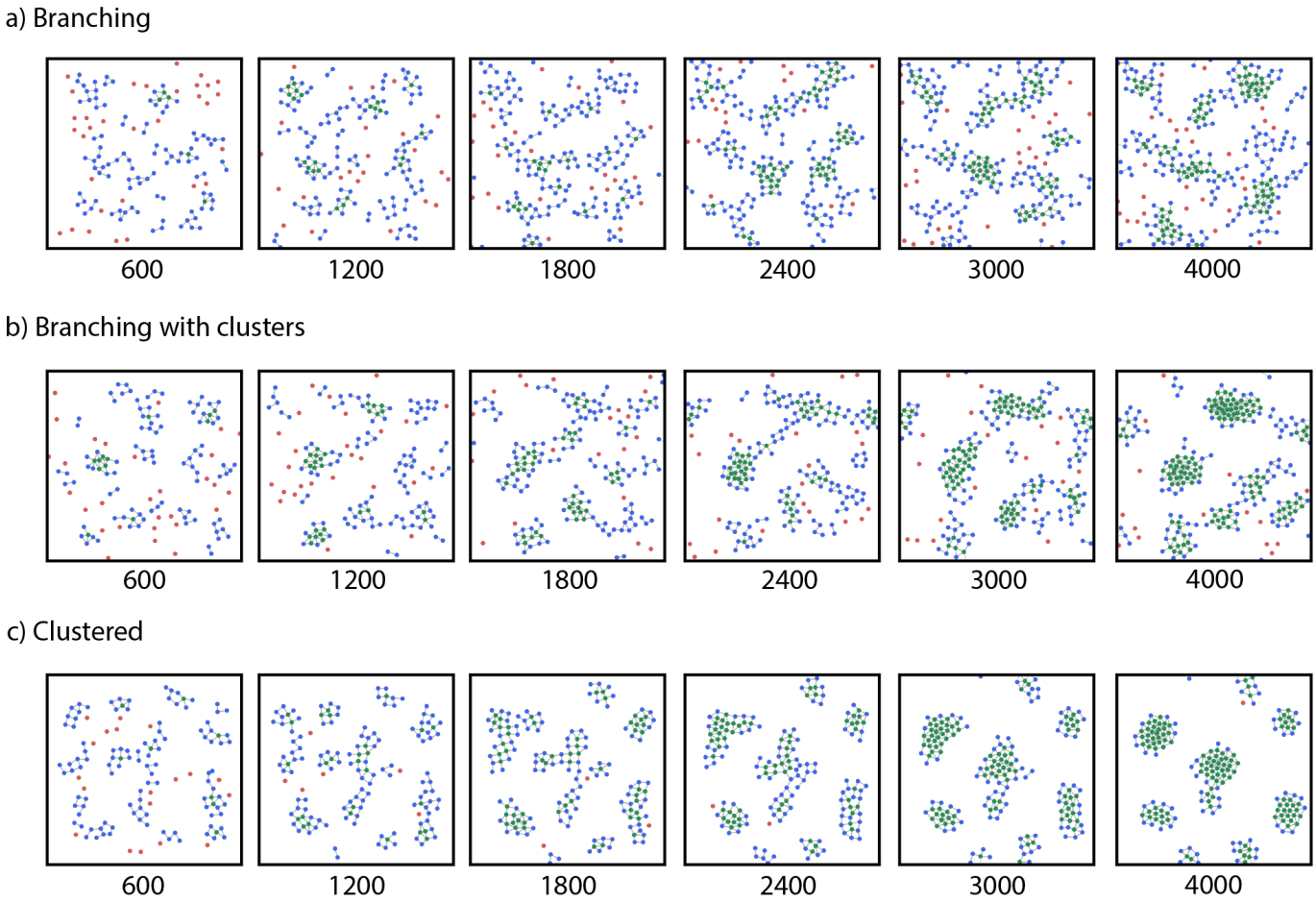

Fig. S8 Self-propelled, proliferating particle model results in branching, branching with clusters and clustered phenotypes with varying adhesion and self-propulsion force. Representative snapshots at time intervals of 600 (every 30,000 timesteps, except for the last snapshot) of branching phase with $\alpha = 0.09, P = 0.009$ (a), branching with clusters phase with $\alpha = 0.19, P = 0.013$ (b), and clustered phase with $\alpha = 0.23, P = 0.007$ (c), with all simulations starting with random initial positions. Particle with one or more neighbors are plotted in black, with a "bond" drawn between any two cells within radial distance 1.0. Particles with four or more neighbors that cannot proliferate due to contact inhibition of proliferation are shown in green.

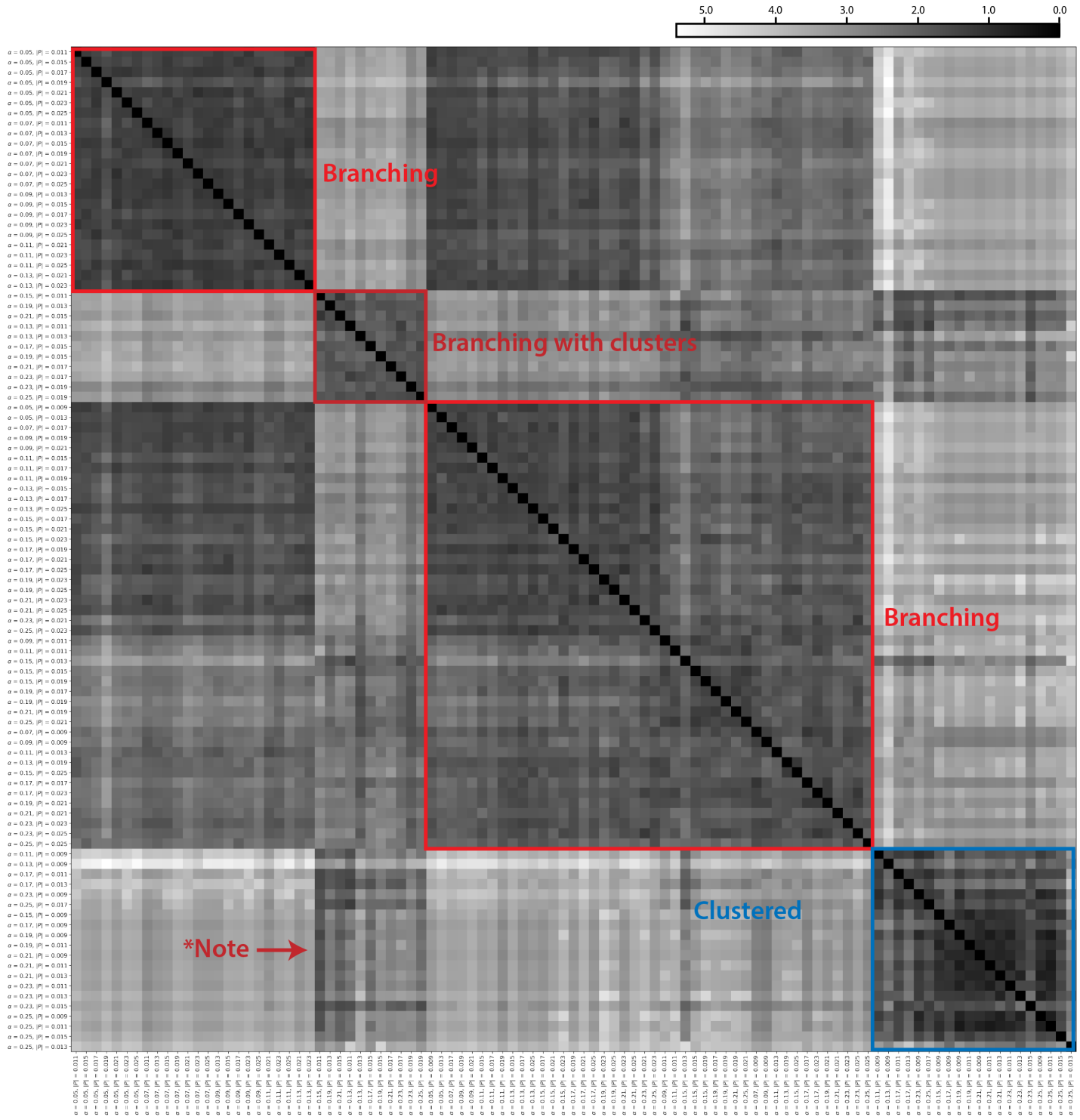

Fig. S9 Pairwise Wasserstein distances between all simulations with proliferation, as well as varying adhesion and self-propulsion force. Hierarchical clustering groups branching, branching with clusters, and clustered phases along the diagonal. Note that branching with clusters phase exhibits similarity with both branching and clustered phases.

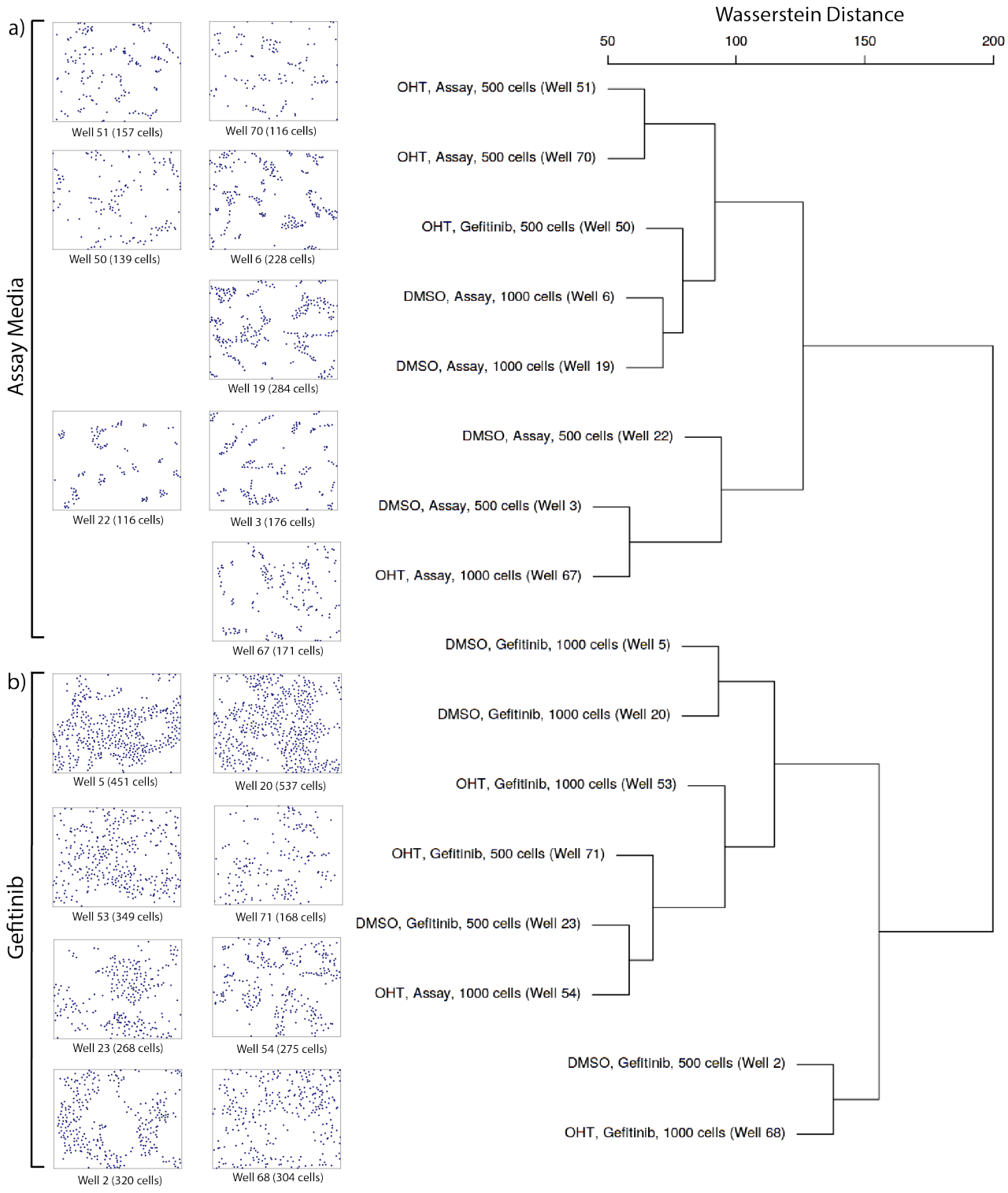

Fig. S10 Hierarchical clustering of pairwise Wasserstein distances between persistence diagrams of experimentally measured cell nuclei positions identifies distinct clustered and branching phases. Dendrogram obtained by running complete-linkage hierarchical clustering algorithm using Wasserstein distance groups experimental conditions based on assay media (a) and gefitinib treatment (b), then biochemical treatment with DMSO or OHT, as well as initial cell density.

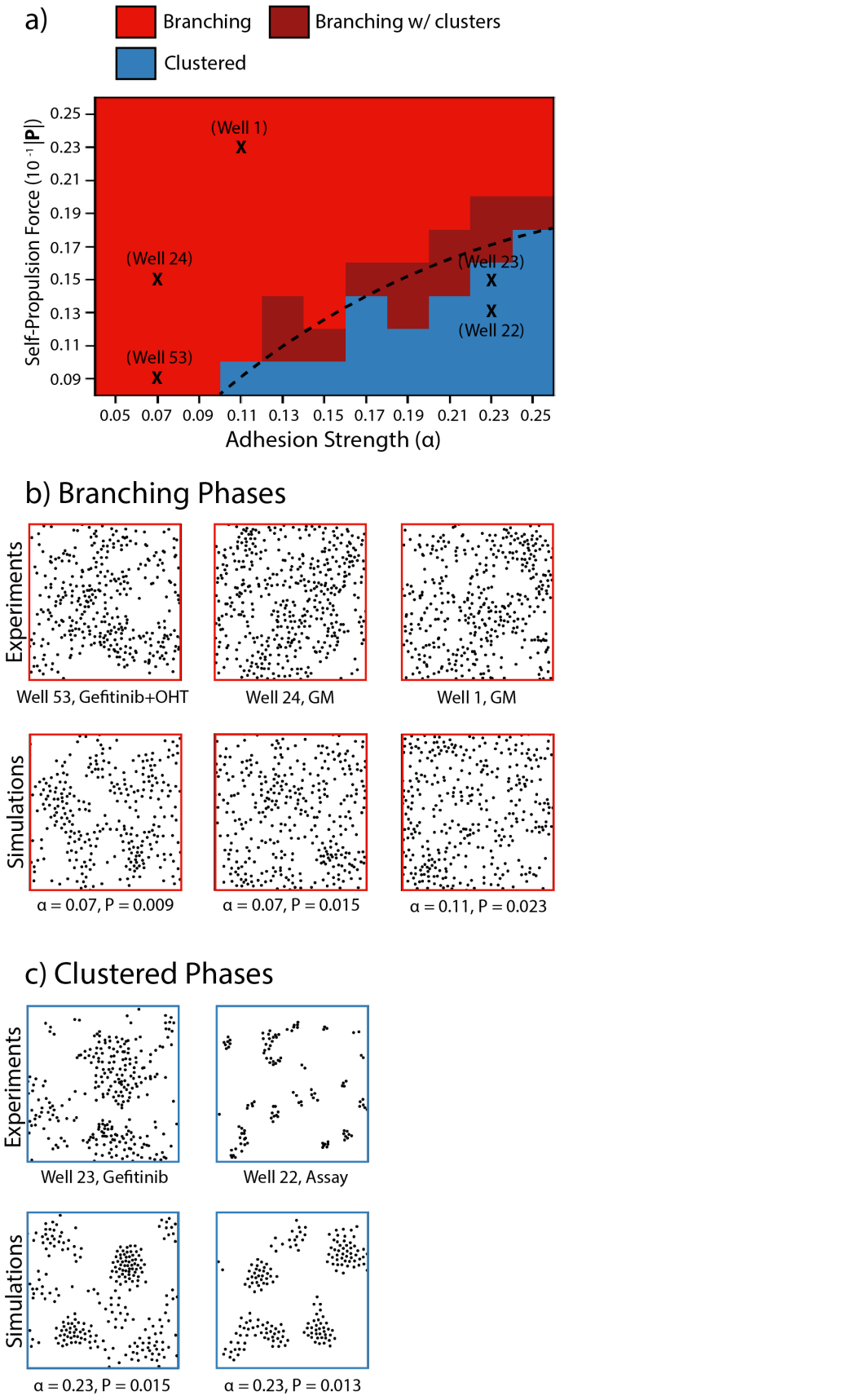

Fig. S11 Experiments compared to simulation results using Wasserstein distance. Pairwise Wasserstein distances are calculated between cell nuclei positions obtained from experiments (scaled to fit in $[-10, 10] \times [-10, 10]$ domain) and particle positions obtained from simulations. Experiments are matched to simulations by minimum distance. Experiments with Wasserstein distances exceeding 1.5 are not shown.

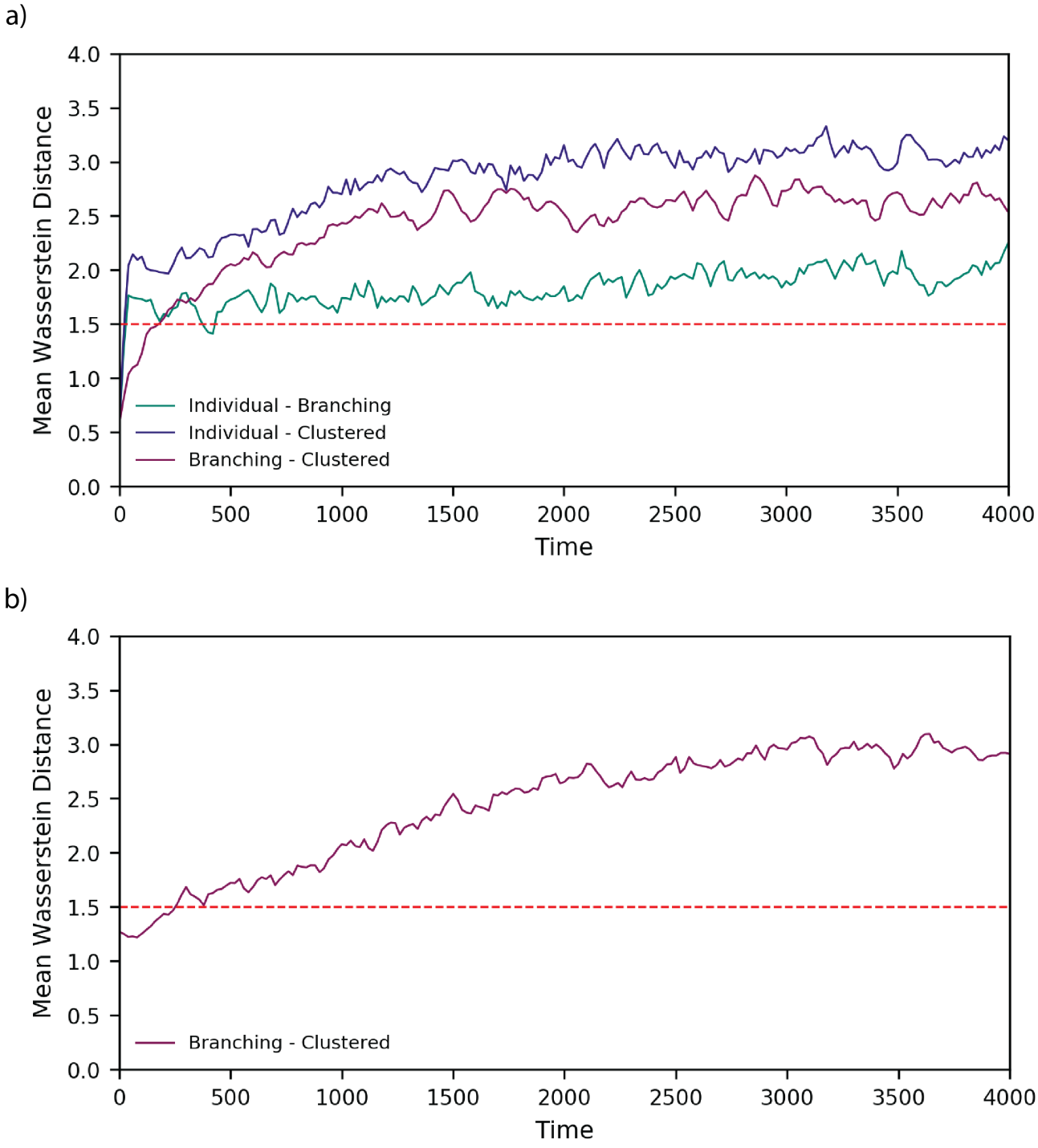

Fig. S12 Pairwise Wasserstein distances over time. (a) Pairwise Wasserstein distances comparing individual, branching, and clustered phases over time at fixed population size. (b) Comparison of branching, and clustered phases in a proliferating population. Dashed red line indicates threshold value above which the pair of simulations can be distinguished. Mean values computed using 10 random initializations for each simulation.

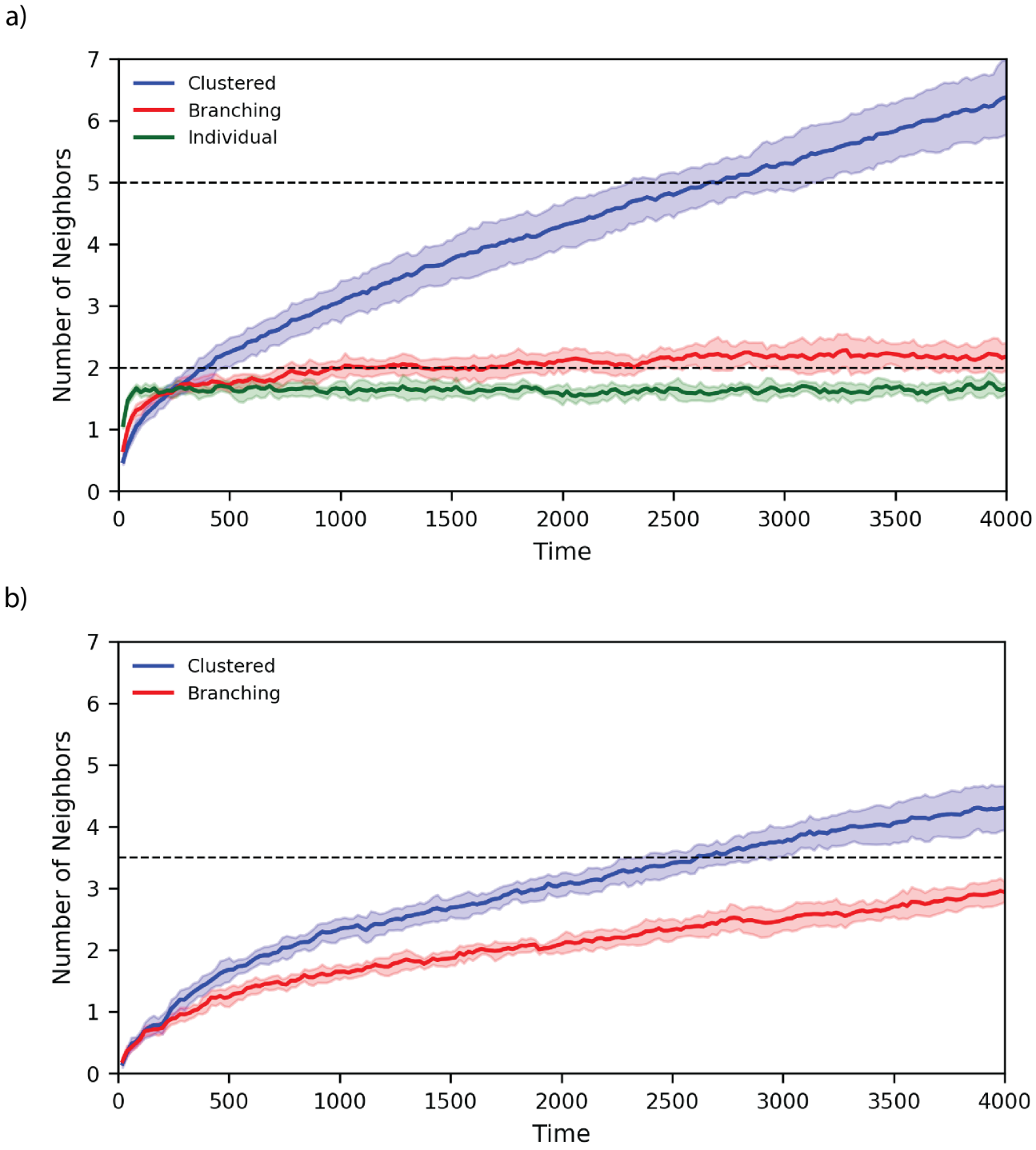

Fig. S13 Number of neighbors over time. (a) Local order parameter over time for individual, branching, and clustered phases at fixed population size. (b) Local order parameter over time for branching and clustered phases in a proliferating population. Dashed lines indicate threshold values to distinguish between individual, branching and clustered phases. Mean (solid bold line) and standard deviation (shaded region) computed using 10 random initializations for each simulation.

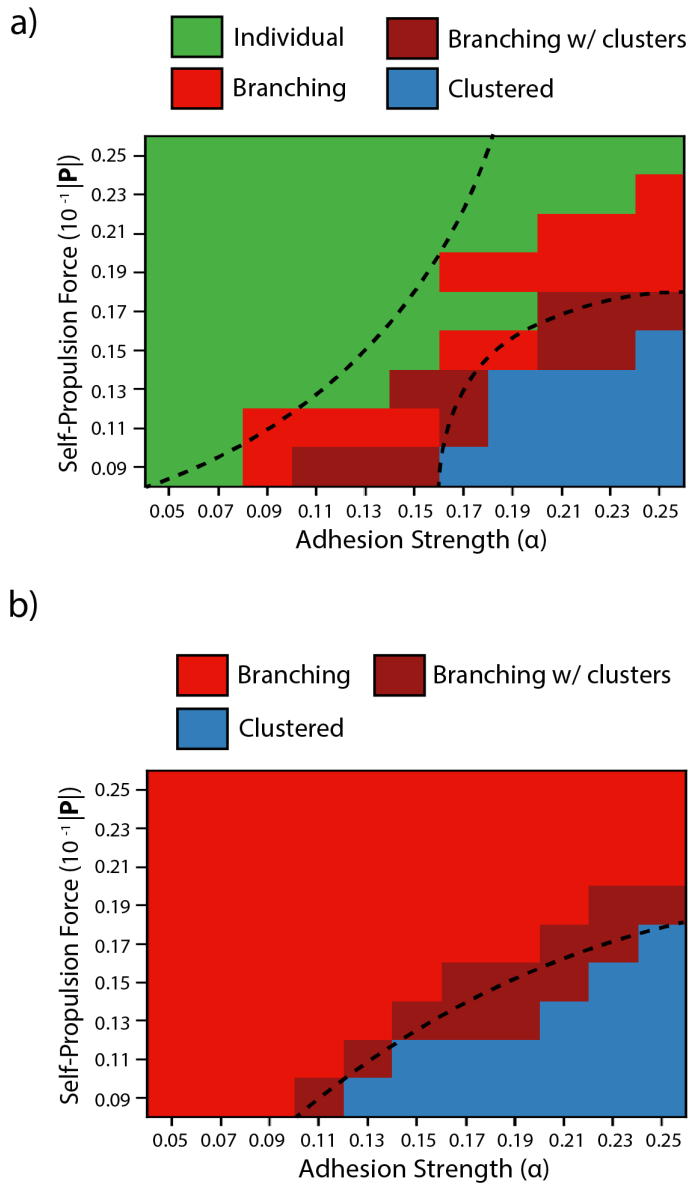

Fig. S14 Phase diagram classified using dimension 0 homology (connected components). (a) Simulations without proliferation are classified into 4 distinct phases, individual, branching, branching with clusters, and clustered. (b) Simulations with proliferation are classified into 3 distinct phases, branching, branching with clusters, and clustered. The dashed lines represent phase boundaries determined manually by inspecting particle positions. Classification results are biased by number of particles.

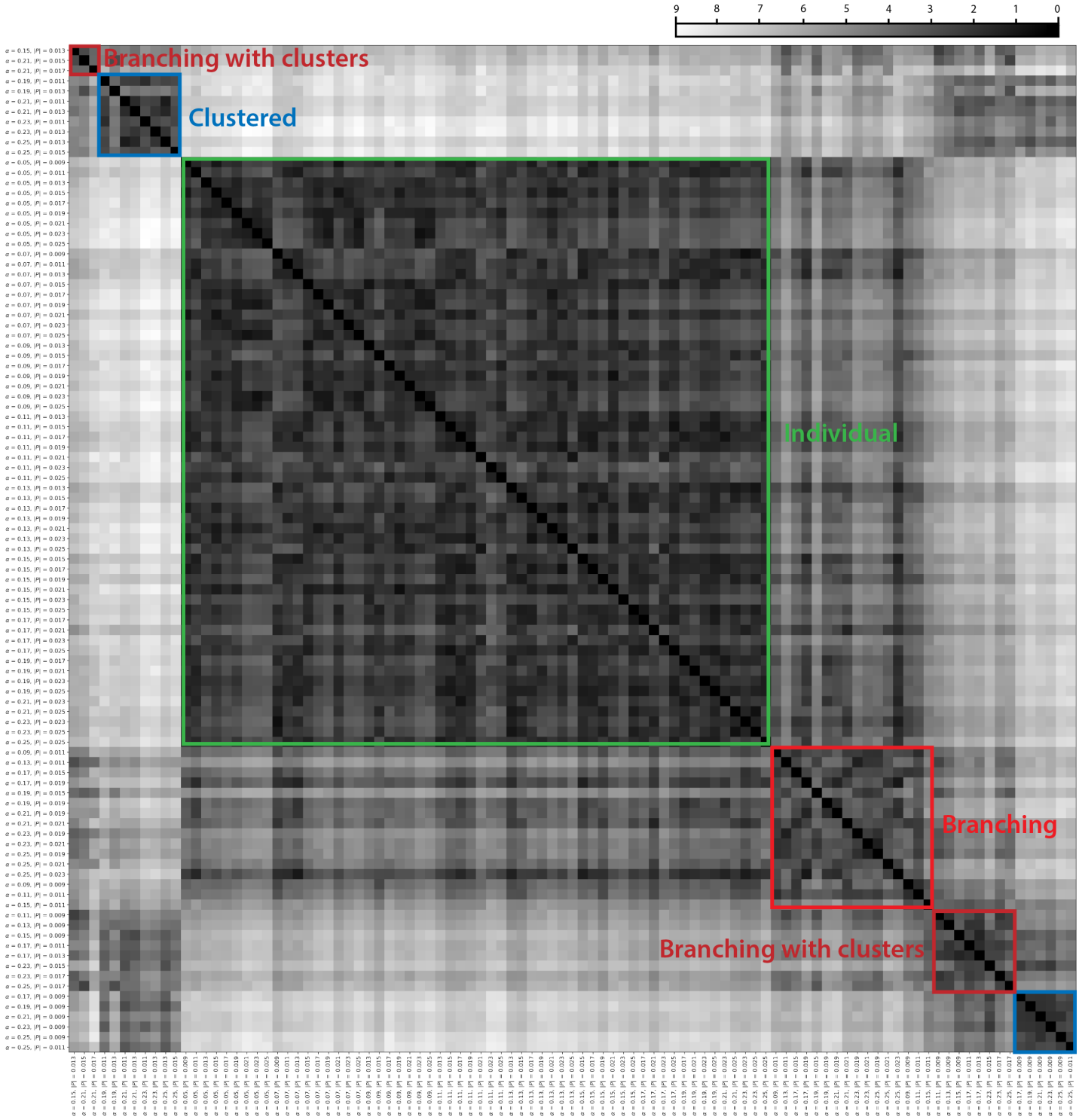

Fig. S15 Pairwise Wasserstein distances computed using dimension 0 homology (connected components) between all simulations without proliferation, as well as varying adhesion and self-propulsion force. Hierarchical clustering groups individual, branching, branching with clusters, and clustered corresponding to distinct phases along the diagonal. Note that branching with clusters phase exhibits similarity with both the branching and clustered phases.

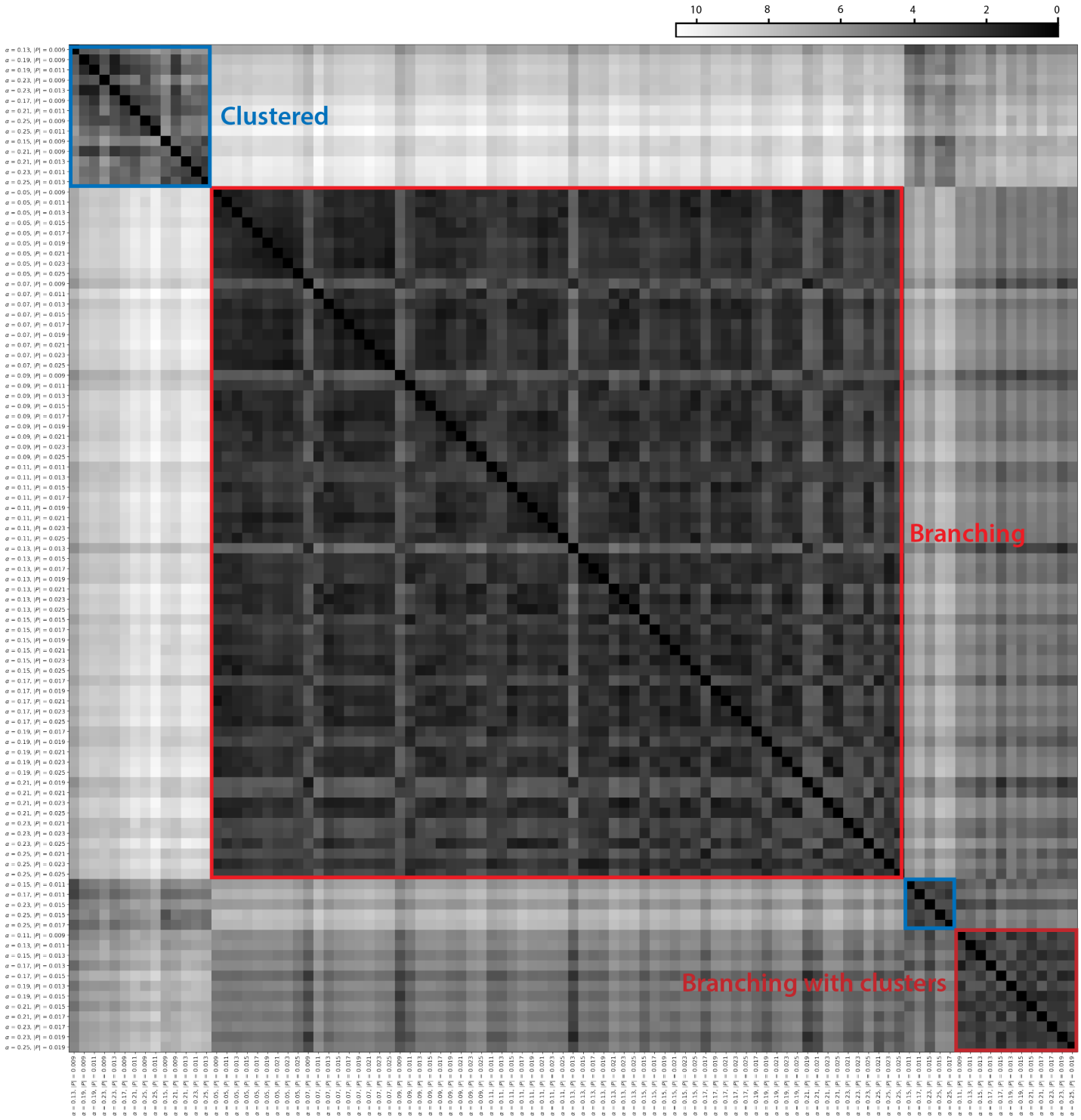

Fig. S16 Pairwise Wasserstein distances computed using dimension 0 homology (connected components) between all simulations with proliferation, as well as varying adhesion and self-propulsion force. Hierarchical clustering groups branching, branching with clusters, and clustered corresponding to distinct phases along the diagonal.