Electronic Supplemental Information for "Reinforcement Learning reveals fundamental limits on the mixing of active particles"

Dominik Schildknecht[1], Anastasia N. Popova[2], Jack Stellwagen[3], and Matt Thomson[1]

[1] Biology and Biological Engineering, California Institute of Technology, Pasadena CA, USA
[2] Applied and Computational Mathematics, California Institute of Technology, Pasadena CA, USA
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh PA, USA

A Tabulated list of the parameters

While all relevant parameters are listed in the main text, we want to list all parameters used in the simulation. In particular, both the parameter set for the environment and the RL hyperparameters are listed in Appendices A.1 and A.2, respectively.

A.1 Environment

Here, we list the parameters for the simulation framework. It should be noted that several features are deactivated. In particular, there is no Brownian motion and no fluid integration taking place for this paper's purpose.

Parameter Name	Value	Notes
sweep_experiment	False	
mixing_experiment	True	
run_id	0	Only relevant if sweep_experiment is true.
savefreq_fig	1000000	Never stores any figures.
savefreq_data_dump	100000	Never store data dumps.
use_interpolated_fluid_velocities	True	Irrelevant because fluids are deactivated by Rdrag= 0
dt	0.05	
T	5	
particle_density	6.0	Leading to 96 particles in a 4×4 area.
MAKE_VIDEO	False	
SAVEFIG	False	
const_particle_density	False	
measure_one_timestep_correlator	False	
periodic_boundary	True	
activation_fn_type	activation_matrix	
L	2	Simulation area is in both directions from $-L$ to L.
m_init	1.0	
activation_decay_rate	10.0	
spring_cutoff	1.5	
spring_k	3.0	
spring_k_rep	3.0	Spring constant k for the repulsive interaction is chosen as strong as for the attractive interaction.
spring_r0	0	
LJ_eps	0	
brownian_motion_delta	0	
mu	10.0	
Rdrag	0	
drag_factor	1	
spring_lower_cutoff	0.015	
n_part	96	

A.2 Reinforcement learning framework: `rl-lib`

Here, we list the hyperparameters chosen for the RL framework. In particular, we used the Proximal Policy Optimization as implemented in `rl-lib` in conjunction with Population Based Training implemented in `ray`[1]. It should be noted that while some of the parameters are initially constrained to an interval, mutations caused by PBT can go beyond the initial interval. Here, $\mathcal{U}([a, b])$ denotes the uniform distribution over the interval $[a, b]$, and $\mathcal{I}(a, b)$ is a random integer between a and b (inclusive).

Parameter Name	Value
`time_attr`	`time_total_s`
`metric`	`episode_reward_mean`
`mode`	`max`
`perturbation_interval`	120
`hyperparam_mutations: lambda`	$\mathcal{U}([0.8, 1])$
`hyperparam_mutations: clip_param`	$\mathcal{U}([0.01, 0.7])$
`hyperparam_mutations: lr`	$[1 \cdot 10^{-2}, 5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}]$
`hyperparam_mutations: num_sgd_iter`	$\mathcal{I}(1, 30)$
`hyperparam_mutations: sgd_minibatch_size`	$\mathcal{I}(128, 16384)$
`hyperparam_mutations: train_batch_size`	$\mathcal{I}(2000, 60000)$

A.3 Computing resources used

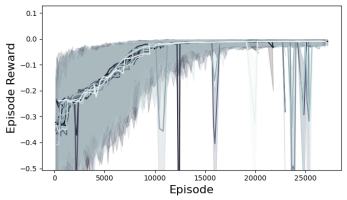
Here, we list the amount of computing resources used for our simulations. There were two sets of simulations, one to determine the strategies, the second one to extract the update matrix M. For both of these, $\{a, b, c, \cdots\}$ means that the experiment was repeated for all values in this set. If multiple sets are listed, simulations were performed for all different combinations. All other numbers are for a single combination. For example, every combination of a specific α and interaction set used 16 CPU cores. The simulation parameters for the strategy screen (see Sec. 4) are:

Parameter Name	Value
α	$\{0, \frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{6}{8}, \frac{7}{8}, 1\}$
(attractive interactions allowed, repulsive interactions allowed)	{(`True`, `True`), (`True`, `False`), (`False`, `True`) }
Number of CPU cores	16
Number of GPUs	1
Training time	2 or 3 days, depending if the 2 days run converged completely

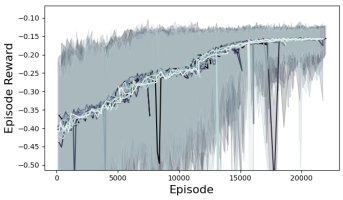
The simulation parameters for the eigenvalue screen are:

Parameter Name	Value
α	$\frac{1}{2}$
(attractive interactions allowed, repulsive interactions allowed)	{(`True`, `True`), (`True`, `False`), (`False`, `True`) }
Number of CPU cores	16
Number of GPUs	1
Training time	3 days

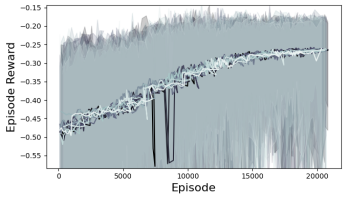
B Rewards during training



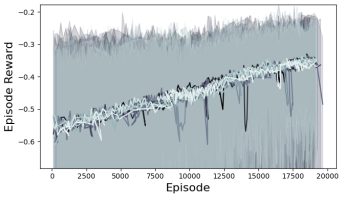
(a) $\alpha = 0$



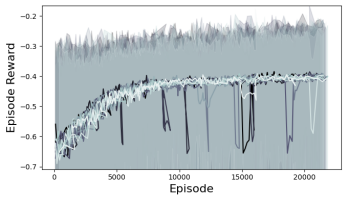
(b) $\alpha = \frac{1}{8}$



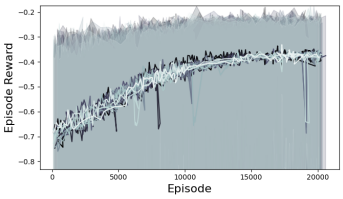
(c) $\alpha = \frac{1}{4}$



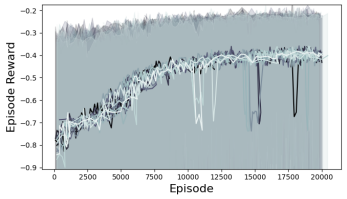
(d) $\alpha = \frac{3}{8}$



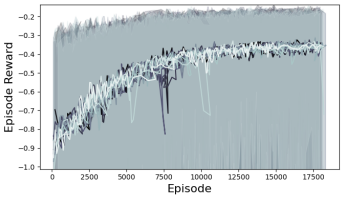
(e) $\alpha = \frac{1}{2}$



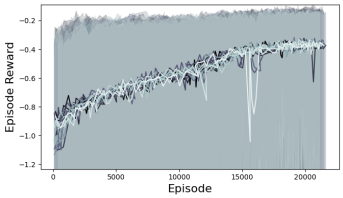
(f) $\alpha = \frac{5}{8}$



(g) $\alpha = \frac{3}{4}$

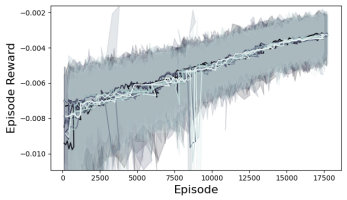


(h) $\alpha = \frac{7}{8}$

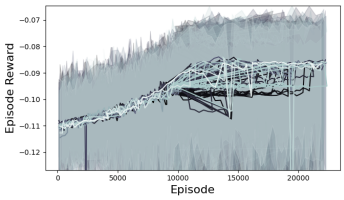


(i) $\alpha = 1$

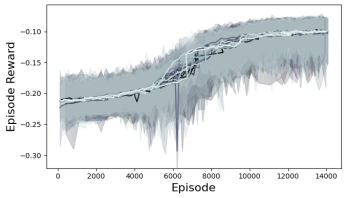
Supplementary Figure 1: Rewards during training for attractive interactions only. Solid lines are average episode rewards of the 16 PBT agents, whereas the error bands denote minimal and maximal episode rewards during each batch. It should be noted that panel (i) is the same as Fig. 2(a).



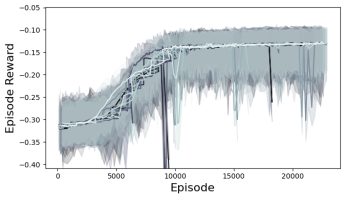
(a) $\alpha = 0$



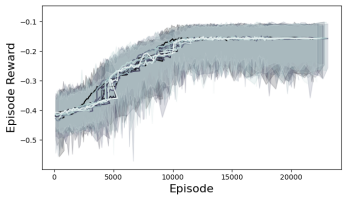
(b) $\alpha = \frac{1}{8}$



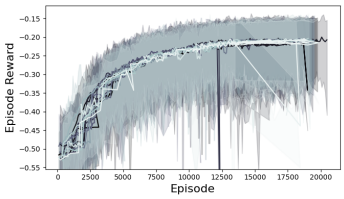
(c) $\alpha = \frac{1}{4}$



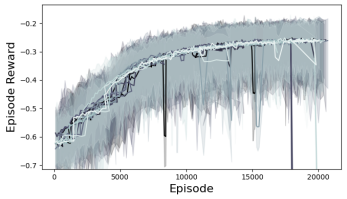
(d) $\alpha = \frac{3}{8}$



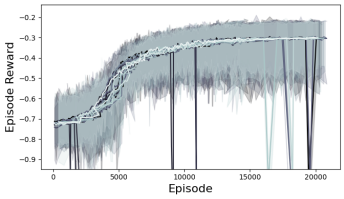
(e) $\alpha = \frac{1}{2}$



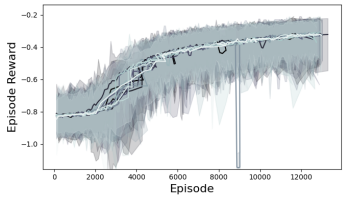
(f) $\alpha = \frac{5}{8}$



(g) $\alpha = \frac{3}{4}$

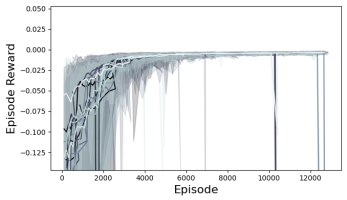


(h) $\alpha = \frac{7}{8}$

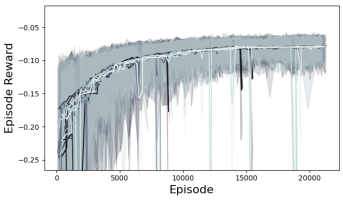


(i) $\alpha = 1$

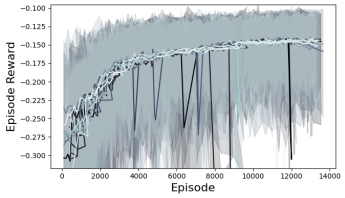
Supplementary Figure 2: Rewards during training for repulsive interactions only. Solid lines are average episode rewards of the 16 PBT agents, whereas the error bands denote minimal and maximal episode rewards during each batch.



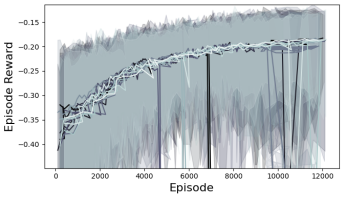
(a) $\alpha = 0$



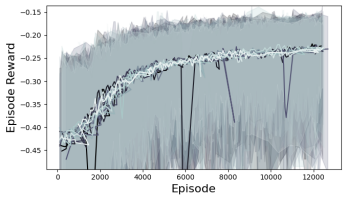
(b) $\alpha = \frac{1}{8}$



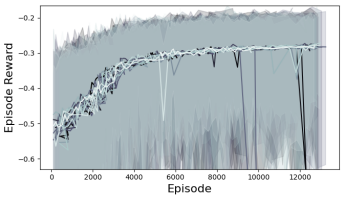
(c) $\alpha = \frac{1}{4}$



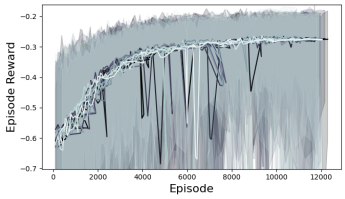
(d) $\alpha = \frac{3}{8}$



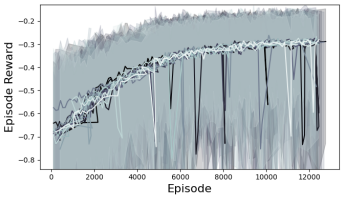
(e) $\alpha = \frac{1}{2}$



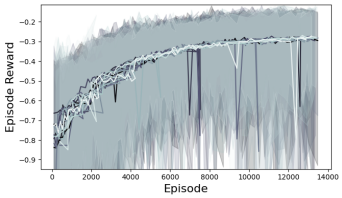
(f) $\alpha = \frac{5}{8}$



(g) $\alpha = \frac{3}{4}$



(h) $\alpha = \frac{7}{8}$



(i) $\alpha = 1$

Supplementary Figure 3: Rewards during training for combined interactions. Solid lines are average episode rewards of the 16 PBT agents, whereas the error bands denote minimal and maximal episode rewards during each batch.

C Detailed description of the strategies

This part of the supplemental information will provide a more detailed description of the emergent strategies we observed during our training. In particular, in Appendix C.1, we show a curated time series for each of the emergent strategies, and in Appendix C.2, we show a more quantitative comparison between the different emergent strategies.

C.1 Curated Time Series

C.1.1 Attractive Interactions

For attractive interactions the following strategies emerged:

- "Collapse all": This strategy activates large parts of the system. The agents try to collapse the system as quickly as possible, and typically not much dynamics is happening after $t = 2.5$ (i.e., time-step 50 out of $N_t = 100$) because most of the system has already collapsed into individual clusters that are separated more than the interaction cutoff R_c. As such, this strategy focuses on speed of collapse over everything else, and often leaves multiple clusters behind as an artifact.

- "Collapse all, careful": Similar to the "Collapse all" strategy, the goal of this strategy is to collapse the system into one dense cluster. However, in contrast, this strategy does so slower and more methodical. This strategy typically leads to dynamics taking more than half of the simulation time: As an example consider Suppl. Fig. 4b, where the system is still fairly spread out in the third panel, after 2/3 of the simulation time. In particular, we believe the strategy does so in order to avoid dense clusters of a single color. As an additional feature of this strategy, by leaving the system more time to react, the cluster formation is typically more thorough, and this strategy often succeeds in collapsing the system into a single cluster, in contrast to the "collapse all" strategy.

- "Collapse some": This strategy is an evolution of the "Collapse all, careful"-strategy where no longer all particles are collected, but only the ones in "bad bins". In particular, this strategy will collapse some of the cells into clusters, achieving good mixing scores in these bins, however, leaving other parts of the system untouched in order to benefit from the homogeneity of the initial state.

- "Activate little": Finally, in the activate little, the system activates very few bins, and if it does so, rarely neighboring bins. This strategy leads to a possible contraction within a single bin, however, rarely over multiple bins. Due to the binning procedure for the rewards, a collapse within a single bin does not affect the rewards, and as such, can be performed by the network without changing the reward.

Curated examples for these strategies can be found in Suppl. Fig. 4.

C.1.2 Repulsive Interactions

For repulsive interactions the following strategies emerged:

- "Repulsive spreading": This strategy activates most of the system. Using this strategy, the agents achieve additional homogenization on top of the random initial configuration. Indeed, as we can observe in Suppl. Fig. 5a, the initial random configuration exhibits some clustering just by random chance. However, activating most of the system repulsively will achieve a decent separation between the particles, and the system in the last panel is better homogenized than the initial configuration was.

- "Activate one side": This strategy uses repulsive interactions on one side of the system very often and rarely on the other side of the system. This strategy will keep one tag of particles inactive, where the others will spread out. Due to the periodic boundary conditions and $N_g = 4$, all left bins are boundary bins to all right bins. As such, if the left bins can be emptied and all the particles can be put into the right bins, an optimal mixing can be achieved in the boundary region, while the homogenization is not strongly punishing, as only half of the system is empty. It should be noted that this strategy might be an artifact of a small number of bins and that for finer binning, where most cells are not boundary cells, this strategy might no longer be optimal for any α.

Curated examples for these strategies can be found in Suppl. Fig. 5.

C.1.3 Combined Interactions

For combined interactions the following strategies emerged:

- "Collapse all": Similar to the collapse all strategy using attractive interactions exclusively, this strategy tries to collapse as much of the system to a single bin as quickly as possible. Repulsive interactions are sometimes used to accelerate the process.

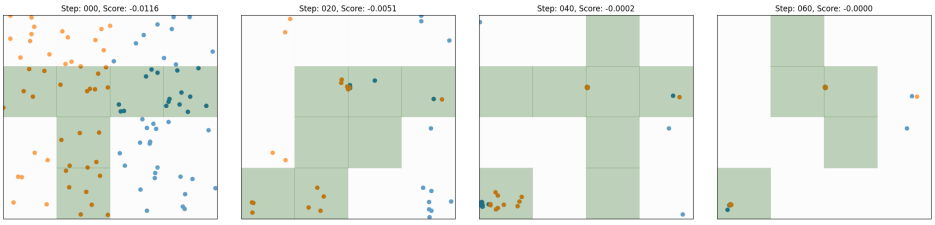
- "Oscillation w/ collapse": This strategy similarly starts by contracting most of the system to a dense cluster. However, it will then apply a repulsive pulse to these particles. The particles then subsequently spread out (cf. Suppl. Fig. 6b, panels 2 to 3). Because the particles are clustered densely, the interactions are very strong once they become repulsive, and the dynamics are quick. As such, this strategy achieves relatively rapid dynamics and mixing. However, it does not necessarily focus on homogenizing the system eventually, as one can observe in the rather collapsed state in the last panel of Suppl. Fig. 6b.

- "Oscillation w/o collapse": This strategy is similar to the "Oscillation w/ collapse" strategy as it also uses attractive interactions to bring particles closer together to increase their mutual interaction and then repulsive interactions to spread them out. However, this strategy is more careful never

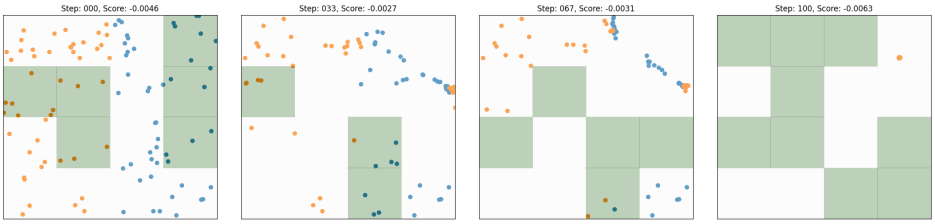
to facilitate a complete collapse. This difference is the main separating characteristic between this strategy and the "oscillation w/ collapse" strategy. It should be noted that because this strategy never collapses particles to dense clusters, the subsequent repulsive interactions are weaker. Hence, the mixing tends to be slower and requires more oscillations. However, this strategy typically also achieves fairly decent homogenization towards the end of the simulations.

- "Attractive-repulsive spreading": This strategy mostly resembles the "repulsive spreading" strategy for repulsive-interactions-only systems. Indeed, the agents use predominantly repulsive interactions to further homogenize the system from the initial configuration.

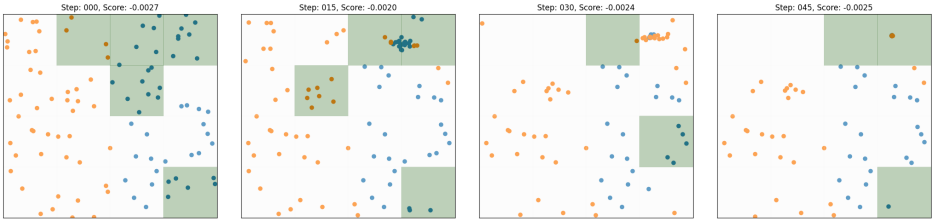
Curated examples for these strategies can be found in Suppl. Fig. 6.



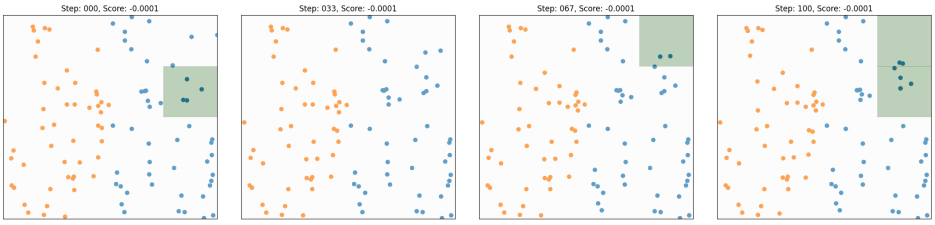
(a) collapse all



(b) collapse all, careful

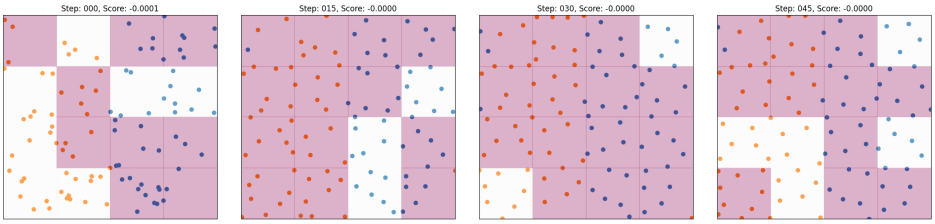


(c) collapse some

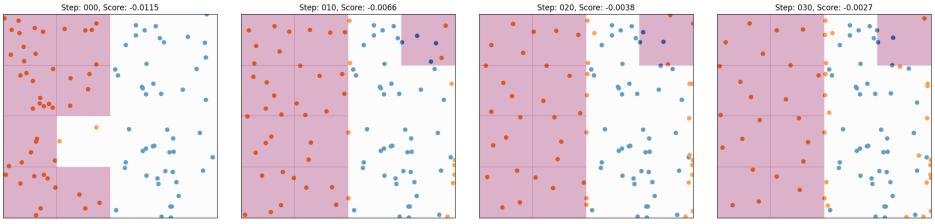


(d) activate little

Supplementary Figure 4: Curated examples for all strategies emerging in attractive-interactions-only simulations. It should be noted that Suppl. Fig. 4a is the same time series as depicted in the main manuscript in Fig. 2(b).

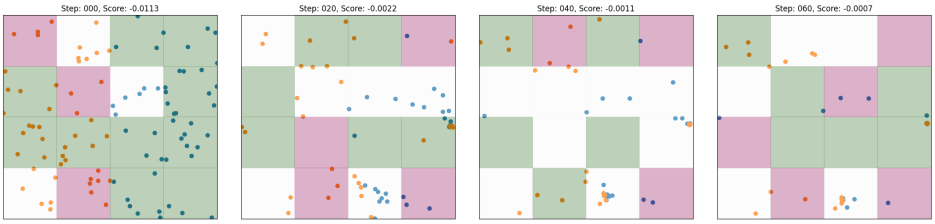


(a) repulsive spreading

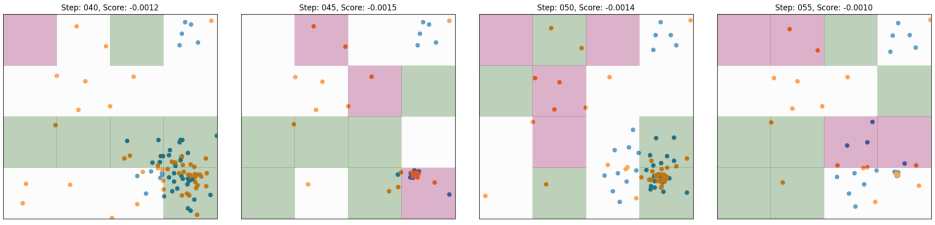


(b) activate one side

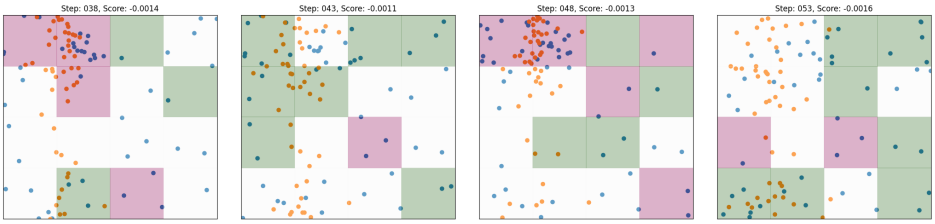
Supplementary Figure 5: Curated examples for all strategies emerging in repulsive-interactions-only simulations. It should be noted that Suppl. Fig. 5b is the same time series as depicted in the main manuscript in Fig. 3(a).



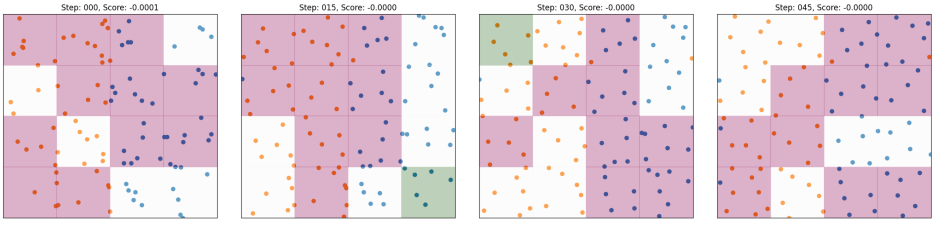
(a) collapse all (with both interactions allowed)



(b) oscillation w/ collapse



(c) oscillation w/o collapse



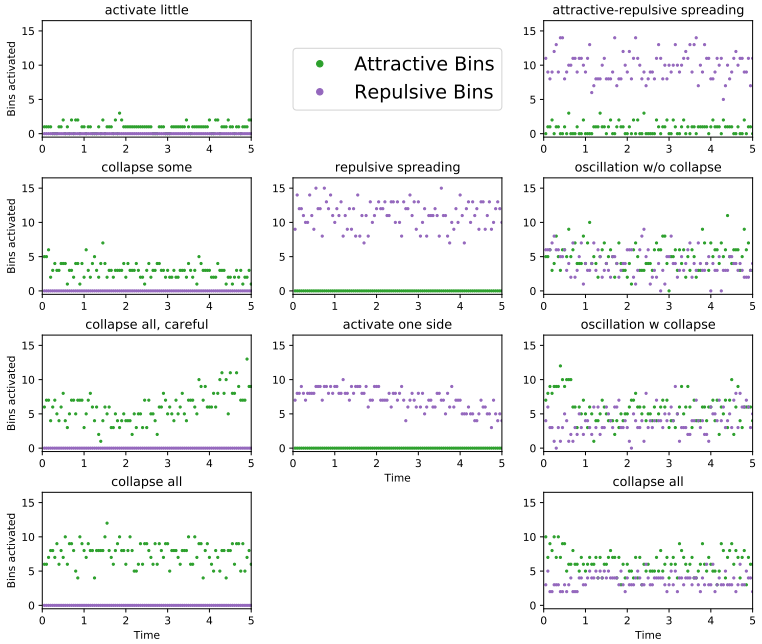
(d) attractive-repulsive spreading

Supplementary Figure 6: Curated examples for all strategies emerging if attractive and repulsive interactions are available. It should be noted that Suppl. Figs. 6b and 6c are the same time series as depicted in the main manuscript in Fig. 4(a) and (b), respectively.

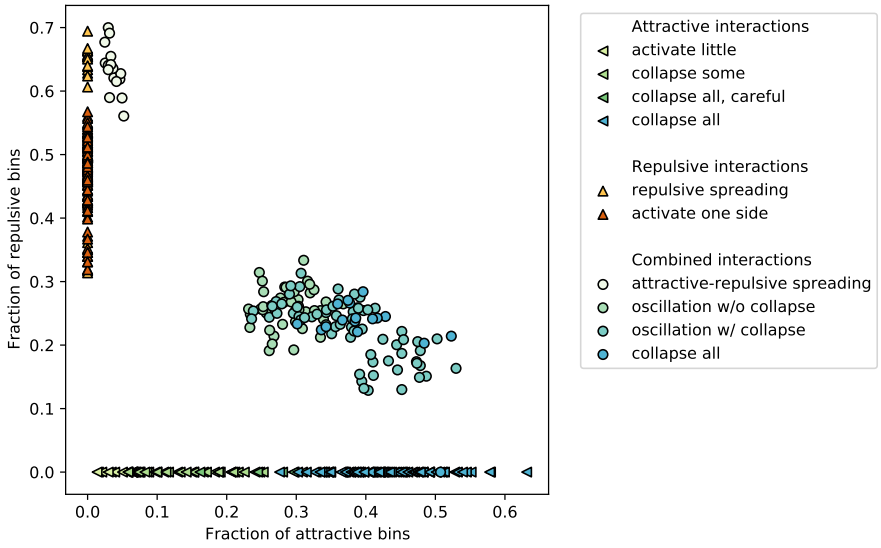
C.2 Quantitative Description

In this section, we present two figures: Suppl. Figs. 7 and 8. Both these figures demonstrate that while a complete classification in terms of raw data is challenging, a quantitative analysis supports the (mostly) smooth strategy evolution presented in the main manuscript.

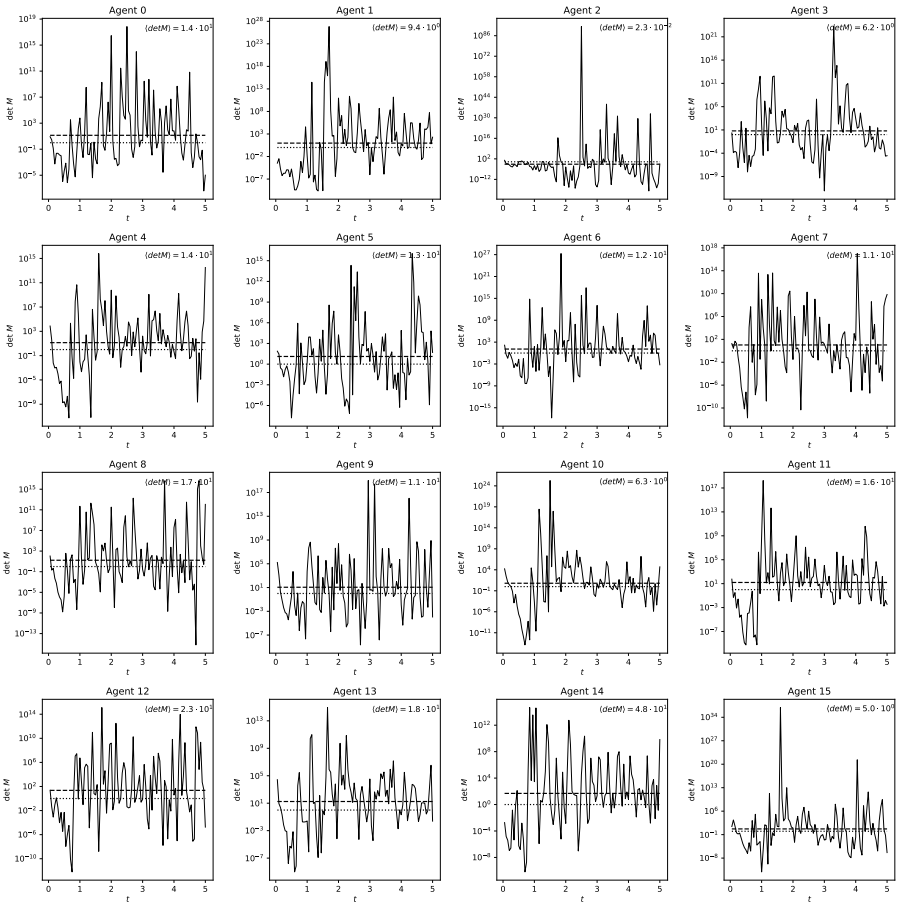
D Determinants as a function of time for trained agents



Supplementary Figure 7: In this figure, we show the number of activated bins (both attractive and repulsive) during the simulation time for the runs shown in the figures in Appendix C.1. It can be observed that the number of bins that are activated fluctuates heavily, but that several trends emerge: The left column shows the attractive only case, where going from the top down, the strategies focus more on the mixing reward. As such, the RL agent uses the interactions more heavily. Additionally, one can observe that the "collapse-all, careful"-strategy uses the attractive activation increasingly towards the end of the simulation, where the "collapse all"-strategy uses more of the activation already in the beginning. As another example, the right column shows combined interactions, where we can observe that the "attractive-repulsive spreading"-strategy uses mostly repulsive interactions, while the "collapse all"-strategy makes heavier use of the attractive attraction.



Supplementary Figure 8: In this figure, we show the fraction of attractive and repulsive activated bins for the last validation videos produced by all our agents (i.e., this data corresponds to the strategy diagrams Figs. 2c, 3b, 4c of the manuscript). In particular, for each of the videos, we used image analysis to extract how many bins were attractively activated, repulsively activated, or not activated during the course of the simulation. The simulations with a restricted interaction set are along the axes, because they could not activate the second type of interaction. In this figure, we can observe two features: Firstly, the strategy evolution is ordered, i.e., the order of the strategy in Figs. 2c, 3b, 4c of the manuscript corresponds to a direction in this plot, namely reducing α, i.e., going towards more homogeneous solutions moves strategies to a smaller fraction of attractive bins and a higher fraction of repulsive bins (to the top left corner of the plot). Secondly, the strategy evolution is mostly continuous, i.e., no large gaps emerge going from one strategy to the next. There are two exceptions to this continuous evolution, namely between "repulsive spreading" and "activate one side" where a small gap emerges, as well as a sizable gap between "attractive-repulsive spreading" and the other combined interactions strategies. This latter jump in strategies can also be observed in the strategy composition diagram for combined interactions shown in Fig. 4c.



Supplementary Figure 9: The value of the determinant during a simulation for $\alpha = 0.5$ with both interaction types for each of the trained agents. The data correspond to the same runs used to determine the eigenvalue histograms for Fig. 5 of the manuscript. In this figure, the determinant as a function of time is shown in a solid line, the dashed line represents the geometric mean of the data, which is also indicated in the top right corner of each panel by $\langle \det M \rangle$, and a dotted line indicates the value 1. It can be observed that the value of the determinant oscillates wildly indicating oscillatory strategies, however, the mean tends to be slightly larger than 1 as one would expect from our theoretical analysis because the linearized approach does not take the periodic boundary conditions into account, which provides an additional phase space limiting factor not taken into account by this linearized analysis.

Supplemental Material References

[1] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez and I. Stoica, *arXiv*, 2018.