






# Fish shoals resemble a stochastic excitable system driven by environmental perturbations

Received: 23 February 2022

Accepted: 12 December 2022

Published online: 6 February 2023

 Check for updates

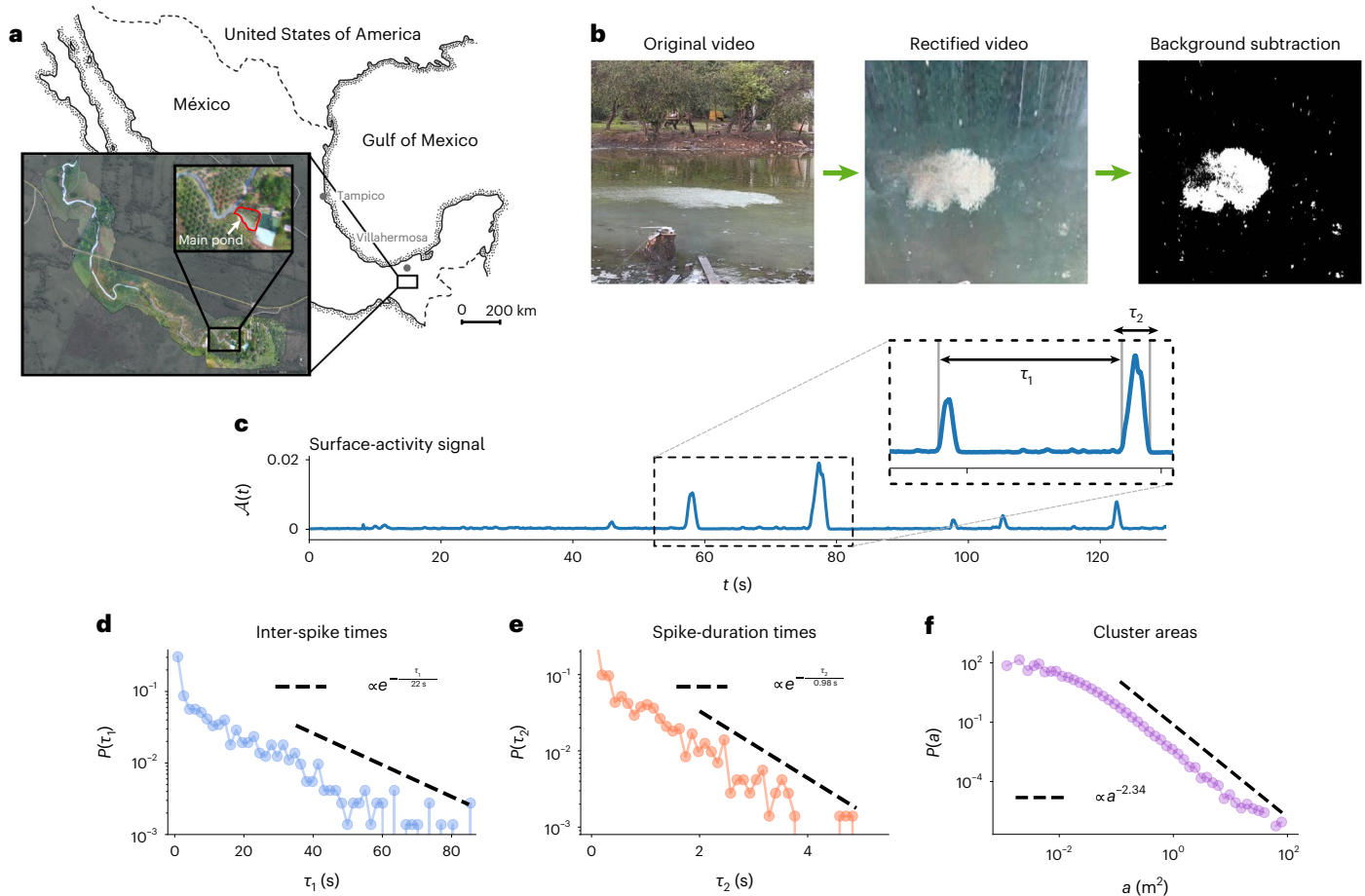
Luis Gómez-Nava <sup>1,2</sup>, Robert T. Lange<sup>2,3</sup>, Pascal P. Klamser <sup>1,4</sup>, Juliane Lukas<sup>5,6</sup>, Lenin Arias-Rodriguez<sup>7</sup>, David Bierbach<sup>2,5,6</sup>, Jens Krause<sup>2,5,6</sup>, Henning Sprekeler <sup>2,3,4</sup> & Pawel Romanczuk <sup>1,2,4</sup> 

Groups of animals can perform highly coordinated collective behaviours that confer benefits to the participating individuals by facilitating social information exchange and protection from predators<sup>1</sup>. Some of these characteristics could arise when groups operate at critical points between two structurally and functionally different states, leading to maximal responsiveness to external stimuli and effective propagation of information<sup>2,3</sup>. It has been proposed that animal groups constitute examples of self-organized systems at criticality<sup>2,3</sup>; however, direct empirical evidence of this hypothesis—in particular in the wild—is mostly absent. Here we show that highly conspicuous, repetitive and rhythmic collective dive cascades produced by many thousands of freshwater fish under high predation risk resemble a stochastic excitable system driven by environmental perturbations. Together with the results of an agent-based model of the system, this suggests that these fish shoals might operate at a critical point between a state of high individual diving activity and low overall diving activity. We show that the best fitting model, which is located at a critical point, allows information about external perturbations—such as predator attacks—to propagate most effectively through the shoal. Our results suggest that criticality might be a plausible principle of distributed information processing in large animal collectives.

Collective systems in biology—such as neuronal networks or large animal groups—are able to perform efficient collective information processing, although each unit or agent often only has access to local information<sup>1</sup>. It has been proposed that this feature of collective biological systems is due to systems operating at the boundary between dynamical regimes of different types, at critical points. At such critical

points, the system's macroscopic behaviour undergoes a qualitative change—that is, a phase transition<sup>4,2</sup>—and the system as a whole becomes highly sensitive to external perturbations, facilitating efficient collective information processing<sup>2,5,6</sup>. Critical or slightly sub-critical behaviour has been reported for neuronal systems<sup>5,7–13</sup>, but evidence for (near)-criticality in other collective information-processing systems,

<sup>1</sup>Institute for Theoretical Biology, Department of Biology, Faculty of Life Sciences, Humboldt Universität zu Berlin, Berlin, Germany. <sup>2</sup>Science of Intelligence, Research Cluster of Excellence, Berlin, Germany. <sup>3</sup>Berlin Institute of Technology, Berlin, Germany. <sup>4</sup>Bernstein Center for Computational Neuroscience, Berlin, Germany. <sup>5</sup>Thaer Institute, Faculty of Life Sciences, Humboldt-Universität zu Berlin, Berlin, Germany. <sup>6</sup>Department of Biology and Ecology of Fishes, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany. <sup>7</sup>División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco, Villahermosa, Mexico. ✉e-mail: [pawel.romanczuk@hu-berlin.de](mailto:pawel.romanczuk@hu-berlin.de)



**Fig. 1 | Analysis of empirical data of the surface-wave activity.** **a**, Localization of the field site in Teapa, Tabasco. The main pond, where all the videos were taken, is marked in the inset of the map. **b**, Snapshots of the rectification and background subtraction processes performed on the empirical videos. Active pixels caused by the disturbance of the surface as fish dive down are depicted in white. **c**, Plot of one surface-activity signal  $\mathcal{A}(t)$  that is representative for all empirical videos. The inset serves to visualize the inter-spike time  $\tau_1$  and the spike-duration time  $\tau_2$ . **d, e**, Plots of the distributions of the characteristic times

$P(\tau_1)$  (**d**) and  $P(\tau_2)$  (**e**) obtained from the empirical data. Both distributions are consistent with gamma distributions, which have exponentially decaying tails (KS statistic,  $p = 0.013$  and  $p = 0.034$ , respectively), shown as black dashed lines. **f**, Plot of the distribution of the cluster areas  $P(a)$  obtained from the empirical data. The power-law fit (black dashed line) was estimated following ref.<sup>30</sup> (KS statistic,  $p = 0.044$ ). For spatial scales ranging from  $10^{-1} \text{ m}^2$  up to  $10^2 \text{ m}^2$ , the exponent is found to be  $\alpha_a \approx 2.3$ . See Methods for details on the goodness-of-fit for all empirical distributions.

including animal groups, remains sparse<sup>2,4,14</sup>. In addition, studies of collective behaviour of animal groups under natural conditions beyond a few hundreds of individuals and over extended periods of time beyond a few seconds<sup>4,15</sup> are missing. In this Letter we address this important gap by observing large groups of fish comprising hundreds of thousands of individuals<sup>16</sup> over hours and in their natural ecosystem. By analysing this unique dataset, we provide evidence that giant fish shoals under high predation risk resemble a stochastic excitable system driven by environmental perturbations. By combining these empirical observations with computational modelling, we suggest that the fish system appears to operate close to criticality and that this could result in a near-optimal propagation of environmental cues, something that would confer benefits at the individual level in the corresponding ecosystem.

Our results are based on an investigation of the collective diving behaviour of sulphur molly (*Poecilia sulphuraria*) shoals. These fish are endemic to a sulfidic stream system near the city of Teapa in southern Mexico<sup>16,17</sup> (Fig. 1a and Methods). Among several adaptations to toxic concentrations of hydrogen sulfide ( $\text{H}_2\text{S}$ ) as well as hypoxic conditions in their habitats<sup>17–20</sup>, these fish spend substantial parts of their time at the water surface, using the thin but oxygen-rich air–water interface to perform aquatic surface respiration<sup>21,22</sup> (Supplementary Video 1). Typically, these fish aggregate in quasi-two-dimensional (2D) shoals of up

to 3,000 individuals per square metre at the surface<sup>16</sup>, which suggests that the area covered in our recordings contained several hundred thousands of individuals. These surface-dwelling, high-density fish shoals attract various bird species as aerial predators at high frequencies<sup>16,22,23</sup>, and fish react to the attacking birds by exhibiting a highly synchronized collective diving behaviour. As the fish touch the water surface with their tails when initiating a dive, a characteristic and easily traceable water disturbance (visual splash—a wave) may spread through the shoal in an avalanche-like manner<sup>16,24</sup> that resembles the ‘Mexican waves’ observed in football stadiums<sup>25</sup>. These large-scale repeated waves that appear after bird attacks function as a deterrent behaviour that leads to fewer and less successful attacks by birds<sup>16</sup>.

A key observation is that the synchronized collective diving behaviour, hereafter referred to as ‘surface-wave activity’, also occurs spontaneously in the absence of bird attacks, as seen in Supplementary Video 2. This spontaneous and stochastic surface-wave activity may be viewed as analogous to the resting-state activity in neuronal systems<sup>26,27</sup>. The system thus provides an optimal set-up for empirically testing the criticality hypothesis and investigating the benefits and potential trade-offs of critical behaviour in the wild<sup>28</sup>. Due to the high predatory bird activity in the system<sup>22,24</sup>, we hypothesized that these fish shoals would probably benefit from being in a constant state of

alertness, operating at a critical point between two phases, one of low and another of high surface-wave activity. If so, the prediction would be to find intermittent surface-wave activity that can propagate through the whole system, giving rise to waves of different sizes ranging from small ones, involving a few individuals, to large ones involving most individuals in the shoals, which would follow a power-law distribution.

To test this hypothesis, we acquired videos, over multiple days, of the spontaneous and stochastic surface-wave activity in the absence of bird attacks, driven only by small-scale perturbations induced by external factors or uncertainty in individual perception (refs. <sup>22 and 24</sup> and Methods). Using a custom computer vision processing pipeline, we binarized the original videos into ‘active’ pixels that represent the diving of fish (state  $\mathcal{D}$ ) and ‘non-active’ pixels corresponding to fish during surface and underwater states  $\mathcal{S}$  and  $\mathcal{U}$  (see Fig. 1b and Methods for details). An example of the resulting processed videos is provided as Supplementary Video 3.

We first analysed the empirical surface-activity signal  $\mathcal{A}(t)$ , defined as the fraction of active pixels in the video, as a proxy for the number of fish diving at a given moment in time. We observed peaks of activity (‘spikes’), corresponding to waves spreading through the system, separated by long periods of low activity corresponding to small-scale, non-propagating surface activity (Fig. 1c, inset). Both the probability distribution of inter-spike time intervals  $\tau_1$  and the probability distribution of spike-duration times  $\tau_2$  have exponentially decaying tails (Fig. 1d,e and Extended Data Fig. 1). The mean time interval between spikes was significantly longer than their duration ( $\langle \tau_1 \rangle = 12.84 \pm 8.35$  s and  $\langle \tau_2 \rangle = 0.68 \pm 0.41$  s, Wilcoxon one-sided signed-rank test,  $W = 1,523,502.5$ ,  $p < 0.001$ ), indicating that the propagation time of the waves is much smaller than the time between successive waves.

In a second step, we analysed how single waves spread through the shoals. To this end, we defined activity clusters as the number of active pixels connected in time and space corresponding to a single wave (as defined in ref. <sup>29</sup>; see Methods for details). We found that the empirical cluster size distribution is consistent with a power-law distribution with an exponent of  $\alpha_c = 2.3$  (Kolmogorov–Smirnov (KS) statistic<sup>30</sup>,  $p = 0.044$ ; Fig. 1f and Extended Data Fig. 2). This exponent is robust across several decades, ranging from  $10^{-1}$  m<sup>2</sup> to  $10^2$  m<sup>2</sup>. The existence of a wave size distribution that is consistent with a power law suggests that the system operates at criticality<sup>12</sup>.

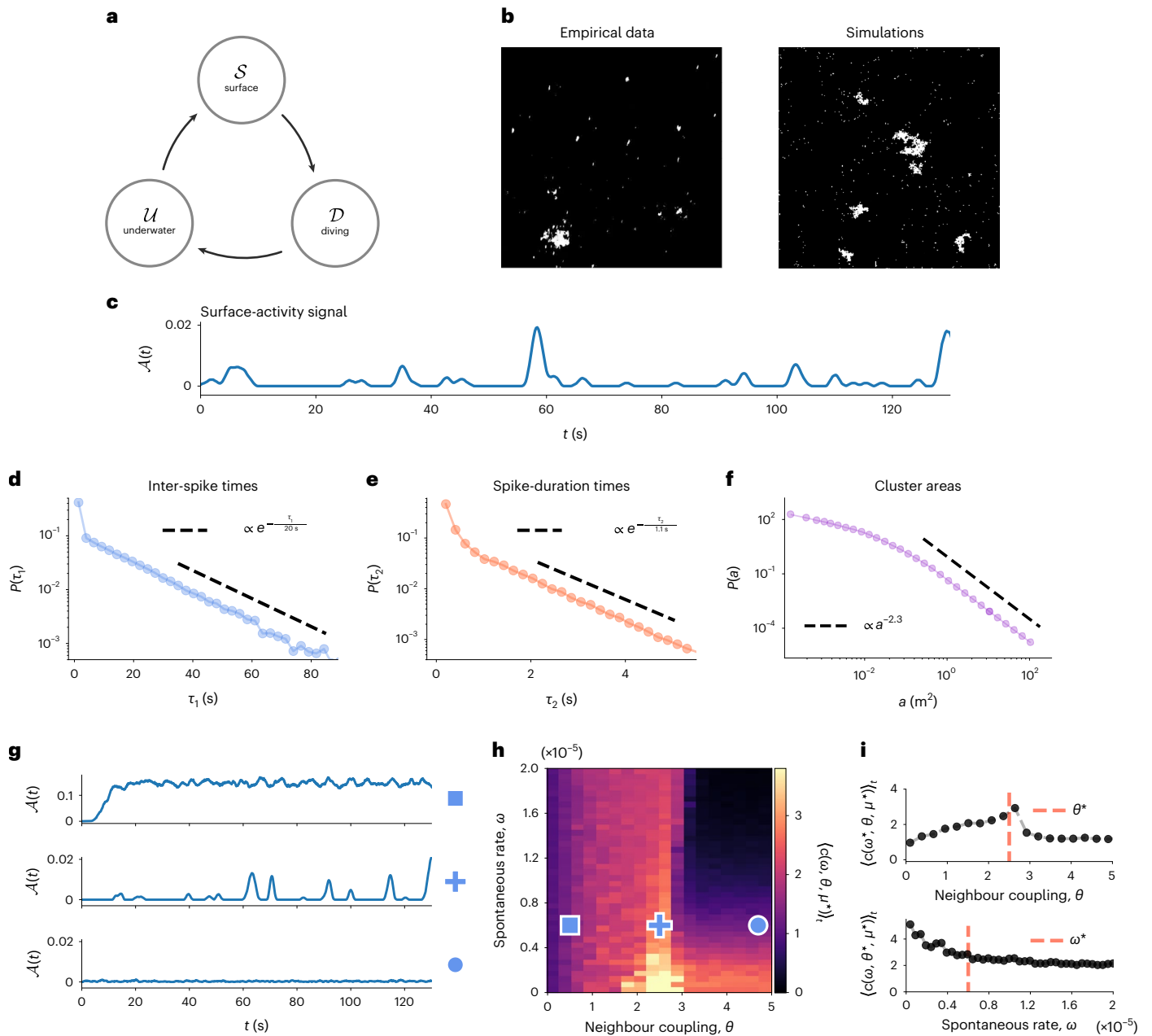
However, a power-law distribution in empirical data is not sufficient, alone, to conclude that a system is at criticality<sup>12</sup>. Thus, to provide further evidence that the observed surface-wave activity is at the edge between two different dynamical regimes, we devised a generic model for the spatiotemporal propagation of the surface-activity waves. In this way we can simulate the emergent collective dynamics and identify parameters best fitting the observed macroscopic behaviour. Our prediction here is that if the fish shoals are operating at a critical point, we will find that the parameters that best fit the data are indeed located at a phase transition (critical point). One way to quantify this is to calculate the average correlation of fluctuations between neighbouring cells, which peaks at the phase transition<sup>6,12</sup>.

We modelled the spontaneous surface-wave activity in a similar manner to Farkas and others<sup>25</sup>. The surface is subdivided into spatial cells, each representing the mean dynamics of a subset of fish. Each cell can be in one of three states, corresponding to the stereotypical behaviours  $\mathcal{S}$  (surface),  $\mathcal{D}$  (dive) or  $\mathcal{U}$  (underwater), as seen in the scheme in Fig. 2a. Cells enter the diving state stochastically, either spontaneously by environmental noise or driven by neighbouring cells<sup>31</sup>, making the propagation of diving events possible. The model has five free parameters: two time constants that control how long the cells remain in the diving and underwater state, two parameters that control the stochasticity of spontaneous events (parameters  $\omega$  and  $\mu$ ) and—crucially—a coupling parameter  $\theta$ , which represents an activation threshold and controls how many neighbouring cells need to be activated to prompt a given cell to become active. We first matched the

time constants to previously published experimental data<sup>16</sup> before performing a systematic parameter search to identify the values  $\omega^*$ ,  $\theta^*$  and  $\mu^*$  that best fit the surface-wave activity of the empirical system (Extended Data Fig. 3). The fit optimized two quantities, one quantifying the temporal dynamics of the system and one characterizing the local spatial dynamics (see Methods and Extended Data Figs. 4 and 5 for details). The best fitting model reproduces the empirical data, both qualitatively (snapshots of the dynamics and surface-activity signal are shown in Fig. 2b,c, respectively) and quantitatively in terms of the distributions of the characteristic times  $P(\tau_1)$  and  $P(\tau_2)$  (Fig. 2d,e), as well as the cluster size distribution  $P(a)$  (Fig. 2f). We note that in the analysis of the simulation results we used time series of the same length as were available from experiments (Methods). Thus, the exponential tails in the inter-spike distributions are probably induced by the relatively short observation windows due to experimental constraints.

By varying the spontaneous and coupling parameters  $\omega$  and  $\theta$  while fixing the other parameters to their best fitting values, we find that the ones that best fit the data are indeed located at a phase transition (critical point), where the system behaviour undergoes a fundamental change mainly with respect to variations in the coupling parameter  $\theta$ . For stronger coupling (lower  $\theta$ ), the system exhibits high levels of persistent activity, effectively spanning the whole system. For weaker coupling (higher  $\theta$ ), we observe only small non-propagating, purely noise-driven activations (Fig. 2g and Supplementary Video 4). As predicted, the model shows a maximal average correlation at a specific value of the coupling parameter  $\theta$  and remains largely unaffected by the spontaneous transition rate  $\omega$ , as seen in Fig. 2h–i (see Methods for details). This means that variation of the social interactions among individuals is responsible for moving the system across the activity phases. To further support that the observed transition in the computational model is indeed a critical one, we computed the susceptibility of the surface-activity signal  $\mathcal{A}(t)$  as a function of system size ( $L \in [25, 50, 100, 200, 400, 800]$ ). We found that the susceptibility increases as a function of system size in a way that is consistent with a power law with exponent  $\approx 1.7$  (Extended Data Fig. 6). As a complement to this numerical analysis of the critical transition in the model, we computed the susceptibility in the empirical videos using different window sizes. We found that, in the empirical data, the susceptibility of the activity also increases with increasing window size (Extended Data Fig. 7). The exponents vary across measurement days—potentially due to environmental factors such as temperature, or lighting conditions—but they are consistent with the exponent in the model for at least three measurement days. Thus, the combination of these numerical and empirical results suggests that the system operates in close vicinity to a critical point. Details on the analysis of the critical transition are provided in Methods (see section Analysis of the critical transition).

To investigate the possibility that the individual fish in shoals at the critical point could benefit from optimal sensitivity to environmental stimuli<sup>32</sup>, as well as efficient information propagation across large distances<sup>33</sup>, we combined two different computational approaches. First, using the numerical model, we tested the sensitivity of the system when exposed to external perturbations of different intensities, analogous to ref. <sup>34</sup>. At random points in time we activated a number  $\mathcal{J}$  of cells in a perturbation zone (PZ in Fig. 3a) always located in the centre of the system, and observed its response in a sampling zone (SZ in Fig. 3a), which consisted of a larger set of cells excluding the cells of the PZ. The sizes of the PZ and SZ were selected such that the size of the SZ is one order of magnitude larger than the PZ. We characterized the ability of the system to detect a perturbation of a given intensity  $\mathcal{J}$  by calculating the surface-activity signal of the cells in the SZ ( $\mathcal{A}_{SZ}(t)$ ) and then computing its time average. We then implemented receiver operating characteristic (ROC) analysis to quantify the effect of the perturbations on the system. To obtain a binary classification, we implemented numerical simulations with no perturbation (true negative (TN) case) and compared them with simulations

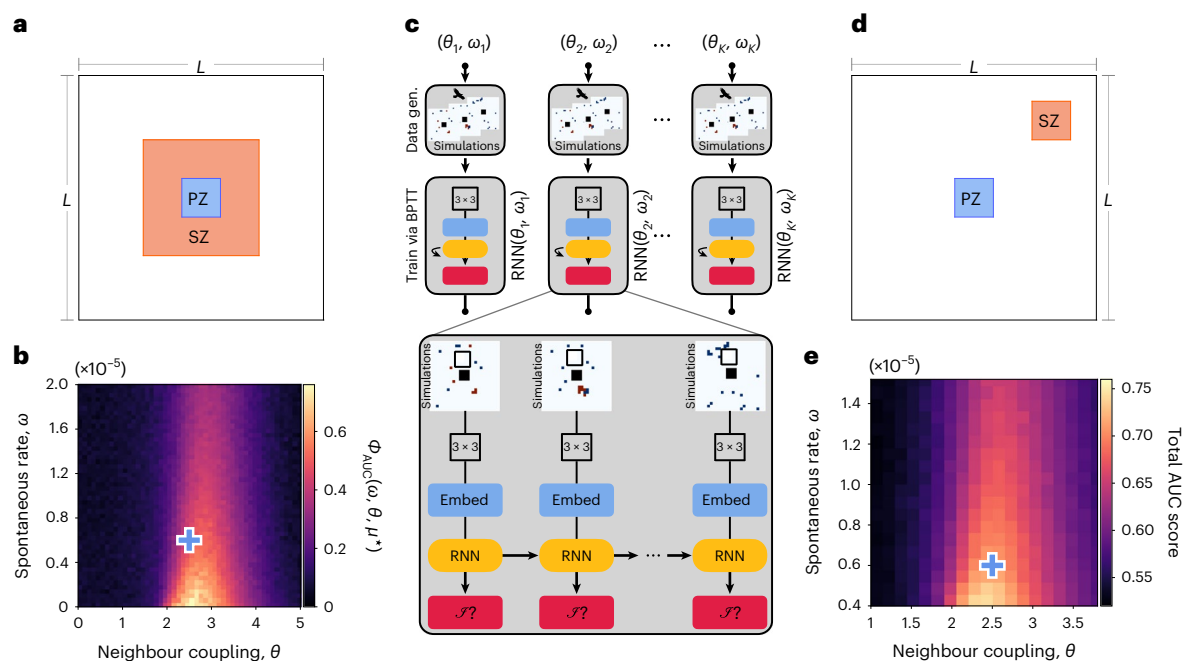


**Fig. 2 | Computational model of surface activity.** **a**, We categorized the behaviour of individual fish into three fundamental states: swimming near the surface for respiration ( $S$ ), fast diving ( $D$ ) and underwater hovering with subsequent slow resurfacing ( $U$ )<sup>24</sup>. **b**, Comparison between snapshots obtained from the empirical videos and the numerical simulations. **c**, Surface-activity signal  $\mathcal{A}(t)$  computed using the model shown in **a**. **d, e**, Characteristic time distributions for the inter-spike times  $\tau_1$  (**d**) and the spike-duration times  $\tau_2$  (**e**). **f**, Distribution of cluster areas  $P(a)$ , which is statistically consistent with a power

law (black dashed line) with exponent  $\alpha_a = 2.3$ . **g**, Plots of the surface-activity signal  $\mathcal{A}(t)$  for three different sets of parameters, highlighted with different markers in **h**. **h**, Average neighbour correlation function  $\langle c(\omega, \theta, \mu^*) \rangle_t$ , showing a maximum at the critical region. The blue cross indicates the location of  $(\omega^*, \theta^*)$ . **i**, Plots of  $\langle c(\omega^*, \theta, \mu^*) \rangle_t$  and  $\langle c(\omega, \theta^*, \mu^*) \rangle_t$  to visualize the maximum when changing variables  $\theta$  and  $\omega$ . The plots in **b–f** were generated using numerical simulations and the parameters  $\omega^*$ ,  $\theta^*$  and  $\mu^*$ . For details, see the section Model implementation in Methods.

perturbed with the five different intensities (true positive (TP) cases). For each perturbation intensity  $\mathcal{I}$ , we computed the area under the curve (AUC) of the false positive (FP) versus TP curve. We found that the average sensitivity  $\Phi_{\text{AUC}}(\omega, \theta, \mu^*)$  is maximal at the critical boundary, as seen in Fig. 3b (see Methods and Extended Data Fig. 8 for details). Again, the parameters that best fit the empirical observations (highlighted with a blue cross in Fig. 3b) are close to this boundary, suggesting that the experimentally observed shoals could maximize their sensitivity to external perturbations.

Second, we characterized the propagation of information through the system by classifying the presence of a perturbation based on the activity in a small SZ at varying location (Fig. 3d). This localized subset of cells aims to emulate the local perspective of small groups of fish in the system. We used a generic machine-learning-based classifier to probe the general ability of the system to identify the perturbations to the system. For this, we used the time-dependent stochastic activation patterns in the SZ from the simulated data. We computed the results in the parameter space defined by  $\theta$  and  $\omega$ , in particular the performance



**Fig. 3 | Information propagation in the critical regime.** **a**, Illustration of the PZ and SZ for sensitivity analysis. **b**, Heatmap showing the values of the parameter  $\Phi_{\text{AUC}}(\omega, \theta, \mu^*)$ . The blue cross denotes the location of the parameters  $(\omega^*, \theta^*)$ . **c**, Scheme to describe the training of the RNN for different parameters  $(\omega, \theta)$  to classify the event of a perturbation. The training was implemented using the

BPTT technique (Methods). **d**, Scheme to visualize the PZ and SZ for analysis of the propagation of information through the system. **e**, Heatmap showing the accuracy of the RNN. We find that the empirically fitted parameters  $(\omega^*, \theta^*)$  (highlighted with a blue cross) are located where the accuracy peaks.

at the set of parameters that best fit the experimental observations. We used a recurrent neural network (RNN; LSTM, long short-term memory)<sup>35</sup> to integrate the local activity information over time (Fig. 3c, Methods and Extended Data Table 1). For each parameter pair  $(\theta, \omega)$ , we trained a separate network, providing a binary prediction at every time step (Fig. 3c). This result, shown in Fig. 3e, is qualitatively similar to the sensitivity analysis (Fig. 3b), suggesting that individuals can detect the presence of a perturbation at a given distance optimally only when the system operates close to the critical region. Taken together, the above analysis demonstrates that, in the model, the critical state facilitates precise and robust propagation of information through the system, something that might also be present in the real-life system.

In summary, by combining empirical data with mathematical modelling, we have shown that the spatiotemporal collective dynamics of large shoals of sulphur mollies correspond to a noise-driven excitable system at criticality. Our estimated exponent for the cluster size distribution is consistent with those observed for the closely related self-organized critical forest fire model by Drossel and Schwab<sup>36</sup>. However, as noted by Grassberger<sup>37</sup>, corresponding estimations are difficult due to strong finite size effects. We further refrain from claims on the exponent belonging to a particular universality class. Although this is a fundamental concept in equilibrium statistical physics, the general relevance of universality for non-equilibrium phase transitions is under debate due to reports of non-universal, parameter-dependent scaling exponents<sup>38</sup>. A corresponding in-depth investigation of the scaling exponents and scaling relationship is beyond the scope of this work. Furthermore, using two different computational classifying methods, we have shown that operating at the critical point maximizes the discrimination ability of (local) environmental cues of different intensities, as well as the communication range between individuals in different locations of the system. In such natural systems, cue intensity scales with danger, as hunting birds often enter the water with large parts of their bodies, leading to high-intensity visual, acoustic and hydrodynamic perturbations, while overflying birds barely provide

visual cues<sup>24</sup>. Thus, information on the cue intensity is highly relevant for the fish in order to coordinate appropriate responses, including repeated diving for multiple minutes<sup>16</sup>. In addition, this information can be communicated across wide distances, allowing fish to take action even when not within the direct area of danger. Being at a critical point potentially allows the sulphur molly shoals to be in a constant state of alertness to environmental perturbations and information on the cue intensity that relates to the danger of these perturbations, which can then be passed most effectively through the shoals. This would suggest that there could be adaptive benefits of the critical state in terms of optimal information processing<sup>39</sup>. This wide range of empirical observations points to (self-organized) criticality as a possible general organizing principle of collective information processing<sup>2</sup>. Nevertheless, there are also important differences between our system and the other biological systems mentioned above. For example, in neuronal systems, the structure of the interaction network between individual elements changes on a much slower timescale than the dynamical behaviour of the network, so the networks may be assumed to be effectively constant. Such timescale separation is assumed essential for self-tuning of these systems to criticality<sup>10,40</sup>. For example, in starling flocks, signatures of criticality have been reported based on short-term observations of highly ordered flocks<sup>33</sup>, where network rearrangements are negligible<sup>2,41</sup>. For the sulphur molly shoals, the interaction networks between individuals are highly dynamic and change on timescales comparable to the observed behaviour. Thus, our results suggest that self-organized critical behaviour could be a robust feature of biological information processing not requiring such timescale separation, while at the same time pose new fundamental questions on the theoretical description and underlying mechanisms of self-organized adaptation towards criticality.

It remains to be studied how the sulphur mollies—and similar systems that show characteristics of criticality—mechanistically tune themselves towards critical points. Ultimately, criticality has to emerge from the adaptation of individual-level behaviour, such as effective

social interactions<sup>3,39,42</sup>. It has been shown that the density of sulphur molly shoals exhibits substantial variation over the course of a day<sup>22</sup>, which in turn modulates the effective strength of social interactions<sup>31,43</sup>. This suggests local density to be one core variable controlling the self-organization of living systems to criticality<sup>28</sup> and thus a variable to quantify over longer periods of time. However, individual behavioural parameters such as individual speed<sup>44</sup> or attention to conspecifics<sup>45</sup> have been shown to have strong effects on collective behaviour, which may yield alternative mechanisms for self-organization towards criticality through modulation of individual behaviour<sup>46</sup>. More generally, this relates also to the recently raised broader question on whether—and if so how—natural collectives and bioinspired artificial systems are capable of flexibly controlling their distance to criticality to modulate their information-processing capabilities depending on the environmental context<sup>28,46</sup>.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41567-022-01916-1>.

## References

- Krause, J. & Ruxton, G. D. *Living in Groups* (Oxford Univ. Press, 2002).
- Mora, T. & Bialek, W. Are biological systems poised at criticality? *J. Stat. Phys.* **144**, 268–302 (2011).
- Muñoz, M. A. Colloquium: criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* **90**, 031001 (2018).
- Bialek, W. et al. Social interactions dominate speed control in poising natural flocks near criticality. *Proc. Natl Acad. Sci. USA* **111**, 7212–7217 (2014).
- Hesse, J. & Gross, T. Self-organized criticality as a fundamental property of neural systems. *Front. Syst. Neurosci.* **8**, 166 (2014).
- Klamser, P. P. & Romanczuk, P. Collective predator evasion: putting the criticality hypothesis to the test. *PLoS Comput. Biol.* **17**, e1008832 (2021).
- Haldeman, C. & Beggs, J. M. Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Phys. Rev. Lett.* **94**, 058101 (2005).
- Levina, A., Herrmann, J. M. & Geisel, T. Dynamical synapses causing self-organized criticality in neural networks. *Nat. Phys.* **3**, 857–860 (2007).
- Beggs, J. M. The criticality hypothesis: how local cortical networks might optimize information processing. *Phil. Trans. R. Soc. A* **366**, 329–343 (2008).
- Meisel, C. & Gross, T. Adaptive self-organization in a realistic neural network model. *Phys. Rev. E* **80**, 061917 (2009).
- Friedman, N. et al. Universal critical dynamics in high resolution neuronal avalanche data. *Phys. Rev. Lett.* **108**, 208102 (2012).
- Beggs, J. & Timme, N. Being critical of criticality in the brain. *Front. Physiol.* **3**, 163 (2012).
- Priesemann, V. et al. Spike avalanches in vivo suggest a driven, slightly subcritical brain state. *Front. Syst. Neurosci.* **8**, 108 (2014).
- Bäuerle, T., Löffler, R. C. & Bechinger, C. Formation of stable and responsive collective states in suspensions of active colloids. *Nat. Commun.* **11**, 2547 (2020).
- Attanasi, A. et al. Finite-size scaling as a way to probe near-criticality in natural swarms. *Phys. Rev. Lett.* **113**, 238102 (2014).
- Doran, C. et al. Fish waves as emergent collective antipredator behavior. *Curr. Biol.* **32**, 708–714 (2021).
- Tobler, M. et al. Evolution in extreme environments: replicated phenotypic differentiation in livebearing fish inhabiting sulfidic springs. *Evolution* **65**, 2213–2228 (2011).
- Pfenninger, M. et al. Parallel evolution of *cox* genes in H<sub>2</sub>S-tolerant fish as key adaptation to a toxic environment. *Nat. Commun.* **5**, 3873 (2014).
- Tobler, M., Kelley, J. L., Plath, M. & Riesch, R. Extreme environments and the origins of biodiversity: adaptation and speciation in sulphide spring fishes. *Mol. Ecol.* **27**, 843–859 (2018).
- Greenway, R. et al. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proc. Natl Acad. Sci. USA* **117**, 16424–16430 (2020).
- Tobler, M., Riesch, R., Tobler, C. & Plath, M. Compensatory behaviour in response to sulphide-induced hypoxia affects time budgets, feeding efficiency and predation risk. *Evol. Ecol. Res.* **11**, 935–948 (2009).
- Lukas, J. et al. Diurnal changes in hypoxia shape predator-prey interaction in a bird-fish system. *Front. Ecol. Evol.* **9**, 619193 (2021).
- Riesch, R. et al. Extreme habitats are not refuges: poeciliids suffer from increased aerial predation risk in sulphidic southern Mexican habitats. *Biol. J. Linnean Soc.* **101**, 417–426 (2010).
- Lukas, J. et al. Acoustic and visual stimuli combined promote stronger responses to aerial predation in fish. *Behav. Ecol.* **32**, 1094–1102 (2021).
- Farkas, I., Helbing, D. & Vicsek, T. Mexican waves in an excitable medium. *Nature* **419**, 131–132 (2002).
- Deco, G., Jirsa, V. K. & McIntosh, A. R. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* **12**, 43–56 (2011).
- Pizoli, C. E. et al. Resting-state activity in development and maintenance of normal brain function. *Proc. Natl Acad. Sci. USA* **108**, 11638–11643 (2011).
- Poel, W. et al. Subcritical escape waves in schooling fish. *Sci. Adv.* **8**, eabm6385 (2022).
- Wang, J., Kádár, S., Jung, P. & Showalter, K. Noise driven avalanche behavior in subexcitable media. *Phys. Rev. Lett.* **82**, 855–858 (1999).
- Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
- Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S. & Couzin, I. D. Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proc. Natl Acad. Sci. USA* **112**, 4690–4695 (2015).
- Kinouchi, O. & Copelli, M. Optimal dynamical range of excitable networks at criticality. *Nat. Phys.* **2**, 348–351 (2006).
- Cavagna, A. et al. Scale-free correlations in starling flocks. *Proc. Natl Acad. Sci. USA* **107**, 11865–11870 (2010).
- Calovi, D. S. et al. Collective response to perturbations in a data-driven fish school model. *J. R. Soc. Interface* **12**, 20141362 (2015).
- Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
- Drossel, B. & Schwabl, F. Self-organized critical forest-fire model. *Phys. Rev. Lett.* **69**, 1629–1632 (1992).
- Grassberger, P. Critical behaviour of the Drossel-Schwabl forest fire model. *New J. Phys.* **4**, 17 (2002).
- Biswas, S., Chandra, A. K., Chatterjee, A. & Chakrabarti, B. K. Phase transitions and non-equilibrium relaxation in kinetic models of opinion formation. *J. Phys. Conf. Ser.* **297**, 012004 (2011).
- Hidalgo, J. et al. Information-based fitness and the emergence of criticality in living systems. *Proc. Natl Acad. Sci. USA* **111**, 10095–10100 (2014).

40. Bornholdt, S. & Rohlf, T. Topological evolution of dynamical networks: global criticality from local dynamics. *Phys. Rev. Lett.* **84**, 6114–6117 (2000).
41. Mora, T. et al. Local equilibrium in bird flocks. *Nat. Phys.* **12**, 1153–1157 (2016).
42. Daniels, B. C., Krakauer, D. C. & Flack, J. C. Control of finite critical behaviour in a small-scale social system. *Nat. Commun.* **8**, 14301 (2017).
43. Sosna, M. M. G. et al. Individual and collective encoding of risk in animal groups. *Proc. Natl Acad. Sci. USA* **116**, 20556–20561 (2019).
44. Jolles, J. W., Boogert, N. J., Sridhar, V. H., Couzin, I. D. & Manica, A. Consistent individual differences drive collective behavior and group functioning of schooling fish. *Curr. Biol.* **27**, 2862–2868 (2017).
45. Rahmani, P., Peruani, F. & Romanczuk, P. Flocking in complex environments-attention trade-offs in collective information processing. *PLoS Comput. Biol.* **16**, e1007697 (2020).
46. Cramer, B. et al. Control of criticality and computation in spiking neuromorphic networks with plasticity. *Nat. Commun.* **11**, 2853 (2020).

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

## Methods

### Collection of field observations

The field observations were performed in Teapa, a municipality in Tabasco in south-eastern Mexico (17° 33' N, 93° 00' W). In this area, there are several sulfidic springs that feed into surrounding surface-water streams and pools. The field work and data collection were performed from 6 April to 11 April 2018 in the largest pond (~600 m<sup>2</sup>) in the Baños del Azufre spring complex (Fig. 1a). A total of 89 videos were collected, each between 120 and 180 s long. This relatively short length of video recordings was chosen to ensure that, during the observation, as well as at least 180 s before the time window, the system had not been subject to a large external perturbation like a bird attack or bird flyover, to obtain the best possible approximation of spontaneous fish activity given the experimental constraints. The videos were recorded in the afternoon (17:00 to 19:00). The frame rate acquisition of all videos was 50 frames s<sup>-1</sup>. Figure 1b provides a snapshot.

### Rectification process

Each of the videos was rectified to reconstruct a top view of the pond. For this purpose, we positioned a 1.55 m × 1.55 m plastic square on the surface of the water at the end of each recording session. We rectified the videos using the four corner positions of the square and the OpenCV Python library<sup>47</sup>. The rectification process is shown in Fig. 1b.

### Background subtraction process

We implemented the background subtraction process to each rectified video using the MOG2 background subtractor from the OpenCV library in Python<sup>48</sup>. After the background subtraction process, the obtained videos consisted only of black and white pixels, as seen in Fig. 1b. For this purpose we used a threshold value of 100 and a history value of 400 in the OpenCV function createBackgroundSubtractorMOG2<sup>48</sup>. We cropped the processed videos to eliminate the black edges of the videos created during the rectification process, as well as the parts of the pond close to the shore, thus keeping only the dynamics in the centre of the pond. The size of all the processed empirical videos is 500 × 500 pixels.

### Analysis of the surface-activity signal

We quantified the surface-wave activity of the pond using the processed black-and-white videos. We computed the surface-activity signal  $\mathcal{A}(t)$ , which is defined for each time step (or frame) as the number of white pixels at a given time  $t$  divided by the total number of pixels in that frame. Thus, the surface-activity signal  $\mathcal{A}(t)$  can take values between 0 and 1. From each empirical signal we calculated two quantities that served to characterize the dynamics. These quantities are the mean value  $\langle \mathcal{A} \rangle_t$  and the threshold value  $\mathcal{A}_{th}$ . They are defined as

$$\langle \mathcal{A} \rangle_t = \frac{1}{T} \sum_{t=0}^T \mathcal{A}(t), \quad (1)$$

$$\mathcal{A}_{th} = \langle \mathcal{A} \rangle_t + 2\sqrt{\langle (\mathcal{A} - \langle \mathcal{A} \rangle_t)^2 \rangle_t}, \quad (2)$$

where  $T$  is the time duration of the signal. We computed a third quantity from each signal, which is the ratio of both previously defined values:

$$\mathcal{R} = \frac{\langle \mathcal{A} \rangle_t}{\mathcal{A}_{th}}. \quad (3)$$

We used this ratio  $\mathcal{R}$  to compare the empirical signals of  $\mathcal{A}(t)$  with the corresponding signals obtained in numerical simulations. The average value over all empirical videos was  $\mathcal{R}_{emp} = 0.33 \pm 0.15$ . The numerical values of these quantities are presented in Extended Data Fig. 4 for each video. We used the threshold value  $\mathcal{A}_{th}$  to compute the characteristic times  $\tau_1$  (inter-spike time) and  $\tau_2$  (spike-duration time), shown in

Fig. 1c. The corresponding distributions obtained from the empirical videos are presented in Fig. 1d,e.

### Fitting of characteristic times distributions and goodness-of-fit

We fitted the empirically obtained distributions for the characteristic times  $\tau_1$  and  $\tau_2$  using a gamma distribution:

$$P(\tau) = \frac{1}{\Gamma(\kappa) \theta^\kappa} \tau^{\kappa-1} e^{-\tau/\theta} \quad (4)$$

where  $\kappa > 0$  and  $\theta > 0$  are two free parameters of the distribution and  $\Gamma(\kappa)$  is the gamma function. We can fit the two parameters of equation (4) using the first two moments of the empirical distributions by computing

$$\kappa = \frac{\langle \tau \rangle^2}{\sigma^2}, \quad \theta = \frac{\sigma^2}{\langle \tau \rangle}, \quad (5)$$

where  $\sigma^2 = \langle (\tau - \langle \tau \rangle)^2 \rangle$ . In this way, we obtain the best candidate to fit the empirical data. To assess the goodness of this fit, we followed a procedure introduced in ref.<sup>30</sup>. We first generate a set of random data sampled using a gamma distribution with the parameters  $\kappa$  and  $\theta$  fitted from the empirical data. We then calculate the  $D$ -statistics using the KS distance:

$$D = \max_{\tau} |F_{synth}(\tau) - F_{theory}(\tau)|, \quad (6)$$

where  $F_{synth}(\tau)$  in equation (6) is the cumulative distribution function (CDF) of the synthetically generated data, and  $F_{theory}(\tau)$  is the analytic expression of the CDF of the gamma distribution in equation (4), which is

$$F_{theory}(\tau) = \frac{1}{\Gamma(\kappa)} \gamma(\kappa, \tau/\theta), \quad (7)$$

where  $\gamma(x, y)$  is the lower incomplete gamma function, defined as

$$\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt. \quad (8)$$

We repeat this process 2,500 times and generate the distribution  $P(D)$  with all the obtained values. We now calculate the value of the  $D$ -statistics using the empirical data by computing

$$D_{emp} = \max_{\tau} |F_{emp}(\tau) - F_{theory}(\tau)|, \quad (9)$$

where  $F_{emp}(\tau)$  is the CDF of the empirical data. We compare the numerical value of  $D_{emp}$  with the distribution generated with the synthetic data and compute the  $p$  value as the fraction of values of  $D$  lower than  $D_{emp}$ . We show the results in Extended Data Fig. 1 for both characteristic times  $\tau_1$  and  $\tau_2$ . The low  $p$  values obtained for both characteristic times ( $p = 0.013$  and  $p = 0.034$ , respectively) suggest that the error between the empirical measurements and the analytic expression in equation (4) can be associated only to noise<sup>30</sup>. We can thus conclude that their distributions are consistent with gamma distributions, which have exponentially decaying tails.

### Computation of activity clusters

The activity cluster analysis presented in the main text was implemented using the definition given in ref.<sup>29</sup>. We quantified the activity clusters (coherent spatiotemporal structures) by stacking the images of each black-and-white video at equal time intervals  $\Delta t$  and then counting the number of connected spatiotemporal white pixels. The cluster



volumes are then defined as the size of these spatiotemporal structures. If we collapse the cluster volumes over the temporal axis, we obtain purely spatial structures that we call cluster areas. The distributions of the cluster volumes and cluster areas are shown in Extended Data Fig. 2. We also computed the speed of each cluster to filter small-sized ones, related to moving objects on the surface of the pond. We used the same methods as used in ref.<sup>16</sup>, and we analysed only the clusters with a speed higher than 0.5 m s<sup>-1</sup> to be sure that the analysed clusters were originated by fish activity.

### Power-law analysis of the cluster distributions

We implemented an analysis on the empirical distributions of the cluster areas and volumes to quantify how far these distributions are statistically consistent with a power-law distribution of the form

$$P(v) \propto v^{-\alpha_v}, \tag{10}$$

$$P(a) \propto a^{-\alpha_a}, \tag{11}$$

where  $v$  and  $a$  stand for the cluster volumes and cluster areas, respectively. We performed the analysis following ref.<sup>30</sup> and using the Python power-law library<sup>49</sup>. There are two important values that need to be calculated from the data: the minimal value  $a_{\min}$  (and  $v_{\min}$ ) and the exponent  $\alpha_a$  (and  $\alpha_v$ ). With these minimal values  $a_{\min}$  and  $v_{\min}$ , we can calculate the exponents using maximum likelihood estimators<sup>30</sup>:

$$\alpha_a = 1 + n_a \left[ \sum_{i=1}^{n_a} \ln \frac{a_i}{a_{\min}} \right]^{-1}, \tag{12}$$

$$\alpha_v = 1 + n_v \left[ \sum_{i=1}^{n_v} \ln \frac{v_i}{v_{\min}} \right]^{-1}, \tag{13}$$

where  $a_i$  (with  $i = 1, \dots, n_a$ ) and  $v_i$  (with  $i = 1, \dots, n_v$ ) are the observed empirical values such that  $a_i > a_{\min}$  and  $v_i > v_{\min}$ . Following ref.<sup>30</sup>, we obtain the best fitting parameters  $\alpha_{\min}^*$ ,  $v_{\min}^*$ ,  $\alpha_a^*$  and  $\alpha_v^*$  and we compute the corresponding  $D$ -statistics for areas and volumes, which we call  $D_a^*$  and  $D_v^*$ . To test the goodness of the fit of these parameters, we generate random data sampled from a power-law distribution with the estimated parameters, analogous to what we explained in the previous section. We then obtain the distributions  $P(D_a)$  and  $P(D_v)$  and compute the corresponding  $p$  values as the fraction of the synthetically generated data where  $D_a \leq D_a^*$  and  $D_v \leq D_v^*$ , respectively.

### Model implementation

The system consists of a square lattice of  $L \times L$  cells. The temporal dynamics of the cells is stochastic, and it is computed with a discrete variable that we call  $q(i, j, t)$ , where  $i$  and  $j$  are integers that represent the location of the cell within the system, and  $t$  is the time. Each cell can be in one of three different states ( $s$ ,  $\mathcal{D}$ ,  $\mathcal{U}$ ), each one associated to one behaviour as explained in the main text. A snapshot of the numerical simulations is shown in Fig. 2b and in Extended Data Fig. 3a. The transition rate that controls the change of cell  $ij$  from state  $s$  to state  $\mathcal{D}$  is mathematically expressed as

$$R_{s \rightarrow \mathcal{D}}^{(ij)}(x) = R_{\text{int}}^{(ij)}(x) + R_{\text{spont}}^{(ij)}, \tag{14}$$

where  $x$  is the number of neighbours in state  $\mathcal{D}$  of the cell  $ij$ . The neighbourhood of each cell is defined as its eight nearest neighbours (Moore neighbourhood; Extended Data Fig. 3b,c). For the interaction term in equation (14) we use a sigmoid function with two parameters called  $\theta$  and  $\mu$ :

$$R_{\text{int}}^{(ij)}(x) = \frac{(x/\theta)^{\theta\mu}}{1 + (x/\theta)^{\theta\mu}}. \tag{15}$$

Examples of equation (15) are provided in Extended Data Fig. 3d,e. The spontaneous term in equation (14) is given by

$$R_{\text{spont}}^{(ij)} = \omega, \tag{16}$$

with  $\omega$  a constant. Notice that the probability of observing a spontaneous transition  $s \rightarrow \mathcal{D}$  in the complete system increases with increasing system size as  $L^2$ . We are then left with a total of three free parameters in the model:  $\omega$ ,  $\theta$  and  $\mu$ . The transitions  $\mathcal{D} \rightarrow \mathcal{U}$  and  $\mathcal{U} \rightarrow s$  were implemented in a deterministic way. This means that a cell that transitioned to state  $\mathcal{D}$  would remain in this state (and thus affect its neighbours that are in state  $s$ ) for a time  $t_{\mathcal{D}}$ . After this time, the cell would change its state to  $\mathcal{U}$ . Analogously, a cell in state  $\mathcal{U}$  remains a time  $t_{\mathcal{U}}$  in this state before changing back to state  $s$ . These transitions can also be implemented in a stochastic fashion using constant rates such that  $R_{\mathcal{D}} = 1/t_{\mathcal{D}}$  and  $R_{\mathcal{U}} = 1/t_{\mathcal{U}}$ . Notice that the stochastic implementation of the transitions  $\mathcal{D} \rightarrow \mathcal{U}$  and  $\mathcal{U} \rightarrow s$  requires the use of a total of three states ( $s$ ,  $\mathcal{D}$ ,  $\mathcal{U}$ ) to have a refractory period between two successive events where a single cell was in the susceptible state  $s$  (ref.<sup>50</sup>). For our simulations, we used the values of  $t_{\mathcal{D}} = 1$  s and  $t_{\mathcal{U}} = 3$  s. These values correspond to the observations published in ref.<sup>24</sup>, where experiments were performed in a controlled environment in the laboratory using small groups of sulphur mollies that were captured in the same sulfidic springs in Teapa. In their experiments, the fish were exposed to artificial visual and acoustic stimuli presented separately or combined. They measured—among other observables—the diving duration (called fast-start duration in Fig. 4c of ref.<sup>24</sup>), as well as the total dive duration (shown in Fig. 4e of ref.<sup>24</sup>). The numerical values of  $t_{\mathcal{D}}$  and  $t_{\mathcal{U}}$  used in our simulations were chosen to reproduce the observations done for the bimodal stimulus.

For the numerical results presented in Fig. 2 in the main text, we used the Euler–Maruyama method<sup>51</sup> to compute the time evolution of the model. For all panels in the figure, the time step was  $\Delta t = 0.1$  s and the system size was  $L \times L = 250,000$  cells. For Fig. 2c,g, the number of time steps was 1,300 (equal to 130s) and only one realization was required for each plot. For Fig. 2d–f, the number of time steps was 5,000,000 and, again, only one realization was required to acquire the data for the three distributions. Finally, for Fig. 2h,i, the number of time steps was 50,000 (per realization) and 10,000 realizations were computed for each set of parameters (neighbour coupling  $\theta$  and spontaneous rate  $\omega$ ). The value of each point in Fig. 2h,i is the average correlation function over all realizations. For details on the computation of the results in Fig. 2h,i, see the Average correlation function section.

### Fitting of the parameters of the model

We fitted the parameters  $\omega$ ,  $\theta$  and  $\mu$  by comparing the empirical data with the numerical simulations via two quantities, the first being the previously defined ratio  $\mathcal{R}$  and the second one the average number of active neighbouring cells, which we call  $\mathcal{N}$ . The ratio  $\mathcal{R}$  is a quantity related to the temporal dynamics of the system, and  $\mathcal{N}$  is instead related to the spatial properties of the observed patterns in the natural system. We computed  $\mathcal{N}$  by randomly selecting time frames from each video and computing the number of neighbouring cells in state  $\mathcal{D}$  of randomly selected focal cells, as depicted in Extended Data Fig. 5. From each video we sampled 2,000 randomly selected frames, and we selected 2,500 cells from each frame as focal cells. The average value obtained over all empirical videos is  $\mathcal{N}_{\text{emp}} = 0.27 \pm 0.23$ . To find the best fitting parameters, we define two auxiliary functions,  $\epsilon_1(\omega, \theta, \mu) = |\mathcal{R}_{\text{emp}} - \mathcal{R}_{\text{sim}}(\omega, \theta, \mu)|$  and  $\epsilon_2(\omega, \theta, \mu) = |\mathcal{N}_{\text{emp}} - \mathcal{N}_{\text{sim}}(\omega, \theta, \mu)|$ , where  $\mathcal{R}_{\text{sim}}(\omega, \theta, \mu)$  and  $\mathcal{N}_{\text{sim}}(\omega, \theta, \mu)$  are the resulting values obtained in the numerical simulations for a given set of parameters  $(\omega, \theta, \mu)$ . For a fixed value of  $\mu$ , we can plot the two functions  $\epsilon_1$  and  $\epsilon_2$  in the 2D space defined by  $\omega$  and  $\theta$ . The corresponding plots for the case  $\mu = 1$  are shown in Extended Data Fig. 3f,g. When we compute the values of  $\omega$  and  $\theta$  that minimize the functions  $\epsilon_1$  and  $\epsilon_2$  (Extended Data Fig. 3f,g, orange lines),

we can find the set of parameters  $\omega^*$  and  $\theta^*$  that minimize both auxiliary functions simultaneously, namely the crossing of both orange lines, highlighted with a blue cross in Extended Data Fig. 3f,g. We can look for the corresponding parameters  $\omega^*$  and  $\theta^*$  for every given value  $\mu$ . We implemented numerical simulations for different values of  $\mu$  and compared them via the function  $\hat{\epsilon}(\omega^*, \theta^*, \mu) = \epsilon_1(\omega^*, \theta^*, \mu) + \epsilon_2(\omega^*, \theta^*, \mu)$ . The values of  $\hat{\epsilon}(\omega^*, \theta^*, \mu)$  are shown in Extended Data Fig. 3h. Notice that the minimal value is reached for a value of  $\mu = \mu^* = 1$ . Thus, the best fitting parameters are  $\omega^* = 0.6 \times 10^{-5}$ ,  $\theta^* = 2.5$  and  $\mu^* = 1$ .

### Analysis of the critical transition in the model

To provide further evidence that the transition observed in the model is indeed critical, we analysed the susceptibility of the mode as a function of the system size. For this, we highlight that the phase transition is mainly driven by the neighbour coupling  $\theta$ , as observed in Fig. 2h, Fig. 3b,e and Extended Data Fig. 3g. Thus, to simplify the analysis, we fixed the spontaneous rate  $\omega$  and parameter  $\mu$  to their optimal values ( $\omega^*, \mu^*$ ) and performed a more thorough analysis by varying only the parameter  $\theta$ . For a given value of this parameter, we computed the average mean square fluctuations of the activity signal  $\mathcal{A}(t)$ —that is, the susceptibility—defined as

$$\chi(\theta) = L^2 \times [\langle \mathcal{A}^2 \rangle - \langle \mathcal{A} \rangle^2], \quad (17)$$

where  $L$  is the system size (and  $L \times L$  is the number of cells in the system). We computed the susceptibility  $\chi$  for six different system sizes:  $L \in [25, 50, 100, 200, 400, 800]$ . We show the results in Extended Data Fig. 6a. In all cases, and for all system sizes, the susceptibility peaks in an interval between  $\theta = 2$  and  $\theta = 3$  (approximately). We can easily notice that the numerical value of the susceptibility at the peak increases with increasing system size  $L$ . We obtained the value of  $\theta$  at which the susceptibility  $\chi$  reaches its maximum value (peak) and called it  $\theta_c$ . We compared the effect of the system size  $L$  on the values of the susceptibility for three different regimes:  $\theta < \theta_c$ ,  $\theta = \theta_c$  and  $\theta > \theta_c$ . The results are shown in Extended Data Fig. 6b, where we can see that the numerical value of the susceptibility increases with increasing system size  $L$  only for the case  $\theta = \theta_c$ , and remains constant for the other two regimes. In that case, susceptibility  $\chi$  diverges with an exponent of -1.7.

To better quantify this result, we computed the width of the peak of the susceptibility  $\chi$ —called  $\mathcal{W}$ —defined as

$$\mathcal{W}^2 = \frac{1}{N} \sum_{i=1}^N \chi_i \times (\theta_i - \theta_c)^2, \quad (18)$$

where  $N$  is the sum of susceptibility values  $\sum_{i=1}^N \chi_i$ ,  $\theta_c$  is the value for which the susceptibility reaches its maximum for each system size  $L$ , and  $\chi_i$  and  $\theta_i$  are the numerical values plotted in Extended Data Fig. 6a for all system sizes. The results are plotted in Extended Data Fig. 6c for all system sizes, where we observe that the width  $\mathcal{W}$  decreases with increasing system size  $L$ . From these results we can conclude that the transition we observe in the model is indeed of a critical nature, given that the susceptibility peaks at  $\theta_c$  (a good method to identify the critical regime), and the larger the system size  $L$ , the larger and thinner the peak.

### Computation of the susceptibility in empirical data

As a complement to the analysis of the critical transition in the model, we computed the susceptibility—as defined in equation (17)—in the empirical videos. For this, we used different window sizes  $L_w$  of the original videos and computed the surface-activity signal  $\mathcal{A}(t)$  in that window. We show the results in Extended Data Fig. 7, where we notice that, in all cases, the susceptibility increases with window size, which strongly suggests that the real-life system operates at criticality. We highlight that the best fitting power laws for the different

acquisition days have an exponent between 1 and 2. As shown in the previous subsection, the exponent obtained in the model (numerical value of -1.7) lies within the same range. Furthermore, there are three days (180407, 180410 and 180411) where the best fit to the empirical data is indeed very close to the numerical result of the model. There are different potential explanations for this variability observed in the exponent: the different camera set-ups or camera position or the time of the day at which the videos were taken. Furthermore, the distribution of the fish in the region of interest may vary slightly over the course of different recordings, which, together with finite size effects, may lead to variation in the corresponding exponents. A corresponding detailed investigation is beyond the scope of this work.

### Average correlation function

To quantify the correlation of the fluctuations of the activity of the system, we computed the average correlation function  $\langle c(\omega, \theta, \mu) \rangle$ , as done in ref. <sup>15</sup>. We first computed the time average value of the variable  $q(i, j, t)$  by calculating

$$\langle q(i, j) \rangle_t = \frac{1}{T} \sum_{t=0}^T q(i, j, t), \quad (19)$$

where  $T$  is the observation time and  $q(i, j, t) = 1$  if the cell is in the active state  $\mathcal{D}$  and  $q(i, j, t) = 0$  otherwise. We can then compute the correlation of the fluctuations of all cells in the system with its neighbours at a given time  $t$ :

$$c(t(\omega, \theta, \mu)) = \frac{\sum_{i,j=1}^L \sum_{l,p \in \Omega_{ij}} (q(i, j, t) - \langle q(i, j) \rangle_t) (q(l, p, t) - \langle q(l, p) \rangle_t)}{\sum_{i,j=1}^L (q(i, j, t) - \langle q(i, j) \rangle_t)^2}, \quad (20)$$

where  $\Omega_{ij}$  is the set of neighbours of a given cell  $ij$ . The resulting function in equation (20) depends on the values of the parameters  $\omega$ ,  $\theta$  and  $\mu$ . To compare the correlation for different parameter values, we computed the time average of equation (20) as

$$\langle c(\omega, \theta, \mu) \rangle_t = \frac{1}{T} \sum_{t=0}^T c(t(\omega, \theta, \mu)). \quad (21)$$

For the plot in Fig. 2h we computed the average value of  $\langle c(\omega, \theta, \mu^*) \rangle_t$  over 10,000 realizations, where each realization consisted of 50,000 time steps. Cuts at  $\omega = \omega^*$  and  $\theta = \theta^*$  are also shown in Fig. 2i.

### Perturbation analysis

We explored the effect of perturbations on the numerical simulations of the model. For this purpose, we considered five different perturbation intensities by perturbing a different number of cells in the PZ (Fig. 3a). The total area of the PZ—called  $\mathcal{V}_{PZ}$ —represents 1.5% of the total area of the system. The main reason is that we aim to study the effect of strong and highly localized perturbations on the system. For the results presented in the main text, we used a system of size  $L \times L = 25 \times 25$  cells (this is, a localized analysis) and thus the total area of the PZ was  $\mathcal{V}_{PZ} = 3 \times 3$  cells. The five different intensities were selected to represent a fraction of  $\mathcal{V}_{PZ}$  such that  $\mathcal{J} \in \{1, 3, 5, 7, 9\}$ .

**Sensitivity analysis.** We implemented the ROC analysis to quantify the effect of the perturbations on the system. To obtain a binary classification, we implemented numerical simulations with no perturbation (TN case) and compared them with simulations perturbed with the five different intensities (TP cases). For all cases, we sampled the dynamics of all cells in the sample zone SZ (Fig. 3a) of total area  $\mathcal{V}_{SZ} = 3 \times 3$  cells. We computed the activity signal of the cells in the SZ—called  $\mathcal{A}_{SZ}(t)$ —for an observation time  $T = 2,000$ , and we computed the time average:

$$\langle \mathcal{A}_{SZ} \rangle = \frac{1}{T} \sum_{t=0}^T \mathcal{A}_{SZ}(t). \quad (22)$$

We implemented 5,000 realizations (using the Euler–Maruyama method, a time step of  $\Delta t = 0.1$  and a system of size  $L \times L = 625$  cells) for a given set of parameters  $(\omega, \theta, \mu^*)$  and obtained the corresponding probability distribution  $P(\mathcal{A}_{SZ})$ . We then calculated the FP versus TP curve, as shown in Extended Data Fig. 8. The AUC of the FP versus TP curves for a given perturbation intensity  $\mathcal{J}$  is referred to in the following as  $\text{AUC}(\omega, \theta, \mu^*, \mathcal{J})$ . To simplify the comparison between parameters, we computed the average response to all the five external stimuli:

$$\bar{\Phi}_{\text{AUC}}(\omega, \theta, \mu^*) = \frac{1}{5} \sum_{\mathcal{J}} \text{AUC}(\omega, \theta, \mu^*, \mathcal{J}). \quad (23)$$

We plot the resulting values for the parameter space  $(\omega, \theta, \mu^*)$  in Fig. 3b. For a plot in the reduced space  $(\omega^*, \theta, \mu^*)$ , see Extended Data Fig. 8.

**Propagation of information analysis.** We studied also the propagation of the information of external stimuli in the system. For this, we used machine-learning algorithms. We generated numerical data for different values of  $(\omega, \theta, \mu^*)$  at a sampling rate of  $\Delta t = 0.1$ s, for an observation time of  $T = 40$ s and a system size of  $L = 25$ . We generated 10,000 realizations and for each one, a random perturbation time  $t_p$  was selected in the interval  $t_p \in [20\text{s}, 40\text{s}]$ . We used perturbation intensities in the interval  $\mathcal{J} \in [1, 9]$  and compared them with simulations where no perturbation was implemented. More specifically, a single realization is a sequence of 2D arrays  $\mathbf{X} \in \mathbb{R}^{(T/\Delta t) \times L \times L}$ , which records the dynamics of the system. For each realization, we computed an auxiliary perturbation temporary series, which is a single vector  $\mathbf{Y} \in \mathbb{R}^{(T/\Delta t)}$ , whose entries are 0 before the perturbation and 1 for all perturbations where the number of perturbed cells exceeds a threshold  $N_{\text{thresh}}$ . Hence, it is a binary classification task. For the perturbation analysis, we discarded the first 20 time units because they correspond to an initial stationary settling period. This results in a time-series of size  $(200, 25, 25)$  for a single sequence. From each of the realizations we sub-sample (without replacement) three  $3 \times 3$  sub-patches next to the PZ. These sub-patches aim to emulate the local-agent perspective. To reduce the memory requirements of our simulations, we furthermore downsample the time dimension and only keep every second time step. The final dataset used to train the networks consists of 30,000 sub-patch sequences of shape  $(100, 3, 3)$ .

**LSTM training.** From the numerical simulations we obtained a high-dimensional time-series of regressors and their corresponding perturbation labels across time. These arrays were reshaped into a vector format to be processed by the RNN. We first embedded  $X_t$  by using a linear layer with 128 hidden units. The embedded input patch,  $\tilde{X}_t \in \mathbb{R}^{128}$ , was then processed by the RNN. We used a common LSTM<sup>35</sup> layer, which maintained a hidden state  $h_t$  and a cell state  $c_t$ . Given an input, these states are updated using a combination of input, forget, update and output gates:

$$\begin{aligned} f_t &= \sigma(U^f \tilde{X}_t + W^f h_{t-1} + b^f) \\ i_t &= \sigma(U^i \tilde{X}_t + W^i h_{t-1} + b^i) \\ \tilde{c}_t &= \tanh(U^c \tilde{X}_t + W^c h_{t-1} + b^c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t &= \sigma(U^o \tilde{X}_t + W^o h_{t-1} + b^o) \\ h_t &= \tanh(c_t) \odot o_t. \end{aligned}$$

Afterwards, the resulting hidden state  $h_t$  is processed by another linear layer to provide the readout logits used to classify the presence of a

perturbation. The networks are trained to optimize a simple binary cross-entropy loss using backpropagation through time (BPTT) and the Adam optimizer:

$$\min_{\theta} \mathcal{L}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^2 Y_{t,c} \times \log[p_{\theta}(Y_{t,c} | X_1, \dots, X_t)], \quad (24)$$

where  $p_{\theta}(\cdot)$  denotes the class probabilities predicted by the LSTM. We rewrite  $\mathbf{Y}$  as a binary one-hot encoded matrix, where  $c$  denotes a specific perturbation class. At each time step the network receives an 18D vector input. The networks were trained on 12 CPU cores each and we used the PyTorch<sup>52</sup> automatic differentiation library. All results shown were obtained by repeating the same procedures for five independent runs with different random seeds. The technical details and parameters used for the training of the networks are listed in Extended Data Table 1.

### Reporting summary

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

### Data availability

All empirical data necessary to reproduce the corresponding figures and results presented in this Paper are available through the public repository Zenodo (<https://doi.org/10.5281/zenodo.7323527>).

### Code availability

Detailed code used to train the LSTM networks is available on Github: <https://github.com/RobertTLange/automata-perturbation-lstm>.

### References

- OpenCV: Open Source Computer Vision Library; <https://opencv.org/>
- Background Subtraction Methods in Python using OpenCV; [https://docs.opencv.org/3.4/d1/dc5/tutorial\\_background\\_subtraction.html](https://docs.opencv.org/3.4/d1/dc5/tutorial_background_subtraction.html)
- Alstott, J., Bullmore, E. & Plenz, D. powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS ONE* **9**, e85777 (2014).
- Lindner, B., García-Ojalvo, J., Neiman, A. & Schimansky-Geier, L. Effects of noise in excitable systems. *Phys. Rep.* **392**, 321–424 (2004).
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ. Press, 2007).
- Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (eds Wallach H. et al.) 8026–8037 (Curran Associates Inc., 2019).

### Acknowledgements

We are grateful to the director and staff at the Centro de Investigación e Innovación para la Enseñanza y el Aprendizaje (CIEA) field station in Teapa (Mexico) for hosting our multiple research stays. In addition, we thank B. Lindner and I. Sokolov for helpful discussions and advice, and H. Klenz for support for the acquisition of data. The acquisition of videos adhered to the ‘Guidelines for the treatment of animals in behavioral research and teaching’ (Animal Behaviour 2021) and were approved by the Mexican government (DGOPA.09004.041111.3088, PRMN/DGOPA-009/2015 and PRMN/DGOPA-012/2017 issued by SAGARPA-CONAPESCA-DGOPA). This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation): BI 1828/3-1 (J.L. and D.B.), RO 4766/2-1 (P.R.)

and under Germany's Excellence Strategy - EXC 2002/1 'Science of Intelligence' project - no. 390523135 (L.G.-N., R.T.L., D.B., J.K., H.S. and P.R.). J.L. was partially supported by the Berlin Funding for Graduates (Elsa-Neumann-Scholarship des Landes Berlin).

### Author contributions

L.G.-N., R.T.L., H.S. and P.R. designed the study. P.P.K., J.L., L.A.R., D.B., J.K. and P.R. acquired and processed the empirical data. L.G.-N. analysed the empirical data. L.G.-N. and R.T.L. performed the numerical simulations. L.G.-N., R.T.L., H.S. and P.R. analysed the results. L.G.-N., R.T.L., D.B., H.S. and P.R. wrote the paper. All authors commented on the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

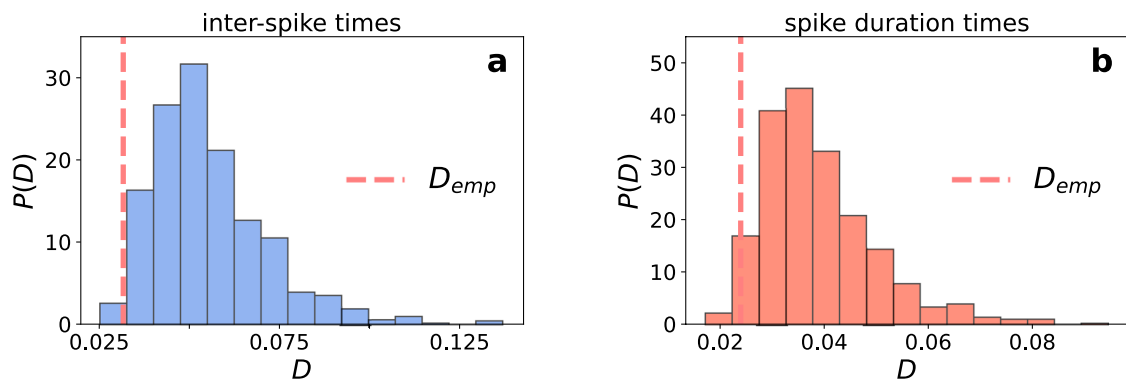
**Extended data** is available for this paper at <https://doi.org/10.1038/s41567-022-01916-1>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41567-022-01916-1>.

**Correspondence and requests for materials** should be addressed to Pawel Romanczuk.

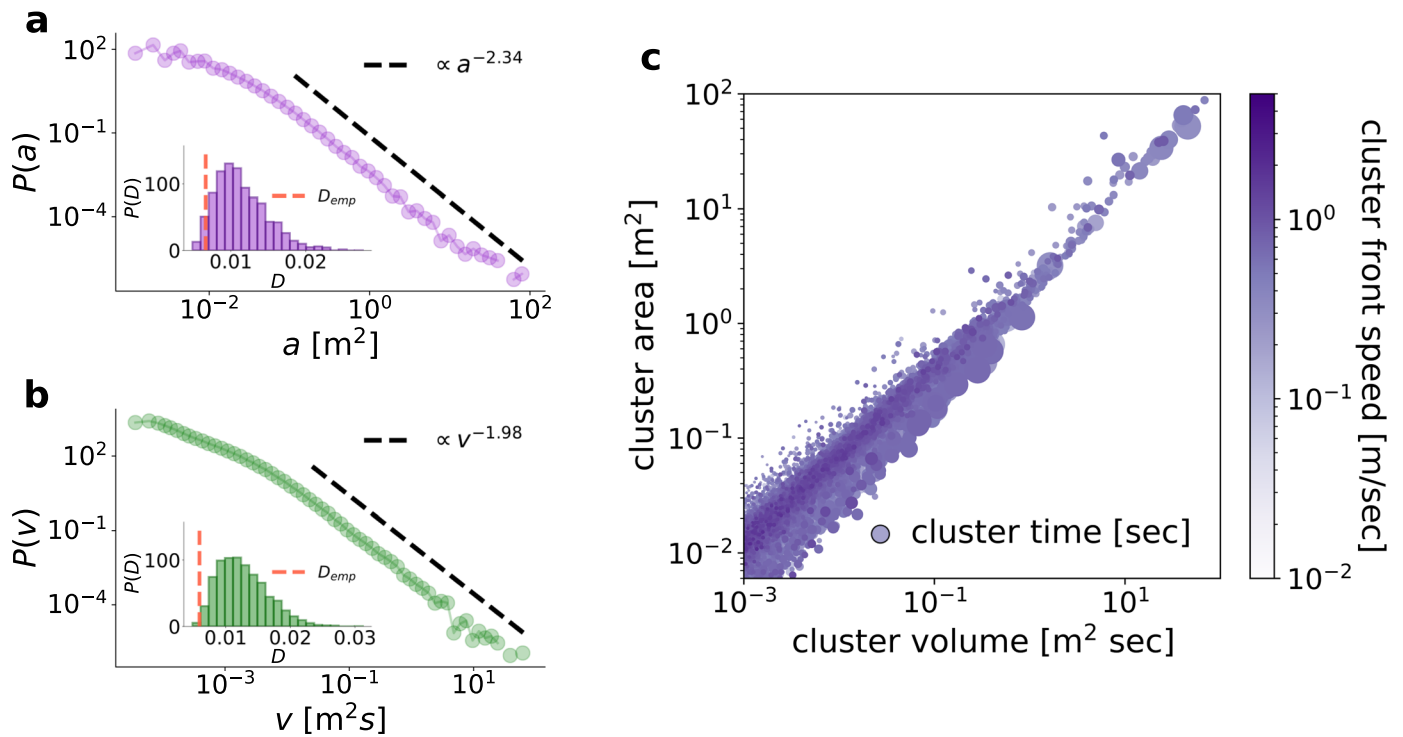
**Peer review information** *Nature Physics* thanks Clément Sire and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



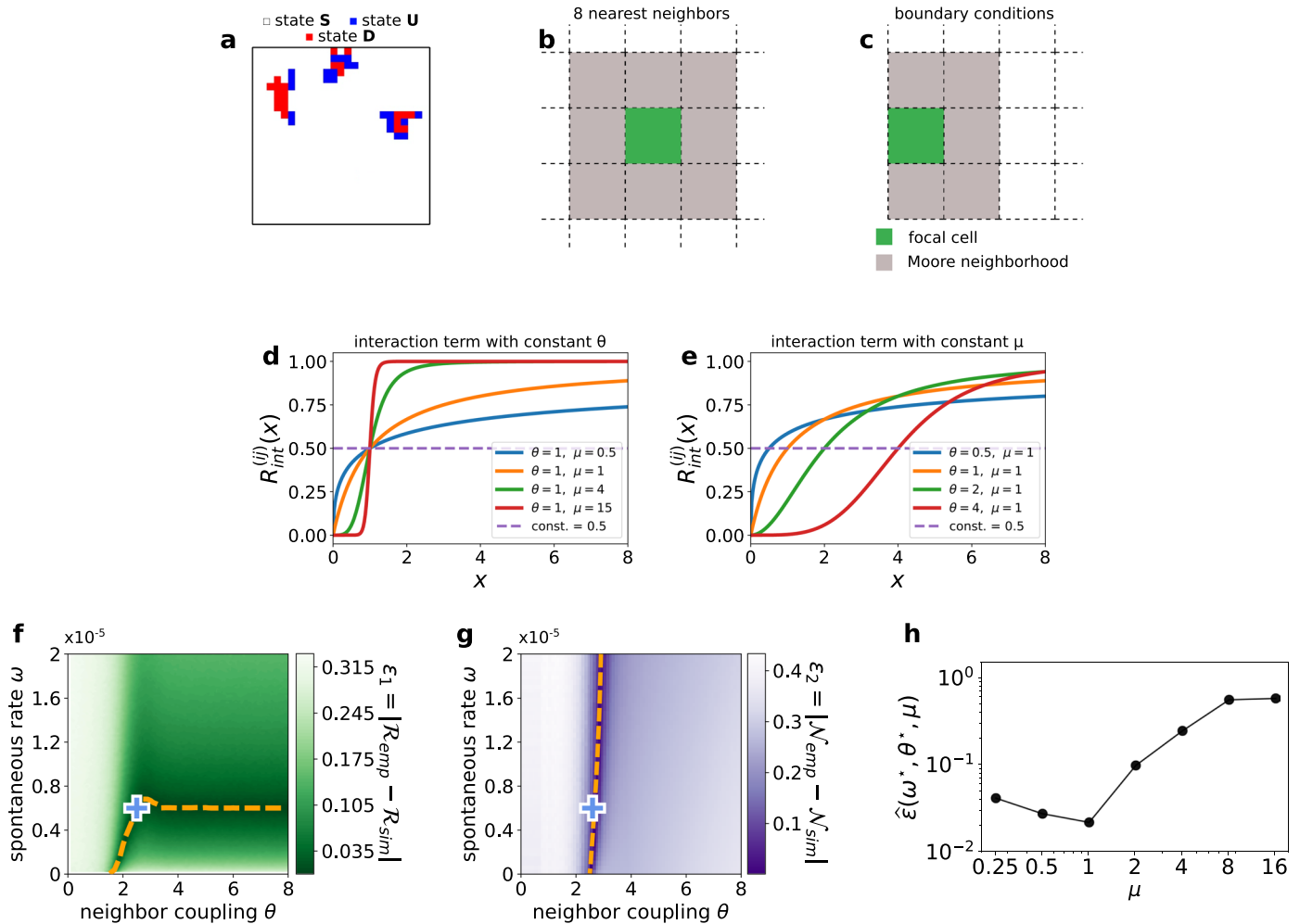
**Extended Data Fig. 1 | Goodness-of-fit of the characteristic time distributions.** **a**, Distribution  $P(D)$  generated with synthetic data from a gamma distribution (eq. (4)) with parameters  $\kappa = 0.59$  and  $\Omega = 21.72$ .  $D$  is the Kolmogorov-Smirnov distance of the empirical ( $D_{emp}$ ) or synthetically generated ( $D$ ) data to the theoretical distribution. The parameters  $\kappa$  and  $\Omega$  were fitted using the

empirical values of the inter-spike times  $\tau_1$ . The p-value results to be  $p = 0.015$ . **b**, A similar procedure was applied to the spike-duration times  $\tau_2$ , where the obtained numerical values of the parameters are  $\kappa = 0.69$  and  $\Omega = 0.99$ . In this case, the p-value results to be  $p = 0.027$ .



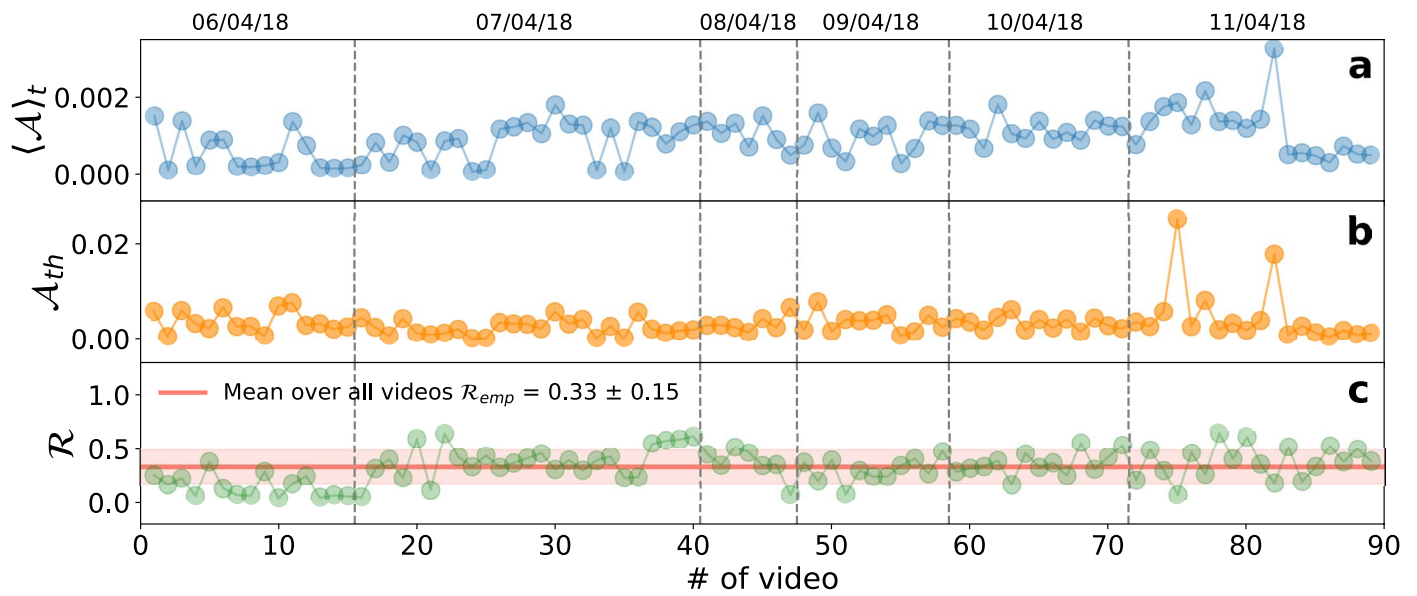
**Extended Data Fig. 2 | Power-law analysis of the empirical cluster distributions. a,b,** Plots of the cluster areas distribution  $P(a)$  and the cluster volumes distribution  $P(v)$ . In each plot, we present the distribution of KS distances  $P(D)$  obtained from the synthetic data that served us to calculate the

corresponding p-values. The p-value for the cluster areas results to be  $p=0.044$ , and for the cluster volumes we obtain  $p=0.0056$ . **c,** Plot of the information of each cluster: area, speed, time-duration and volume.



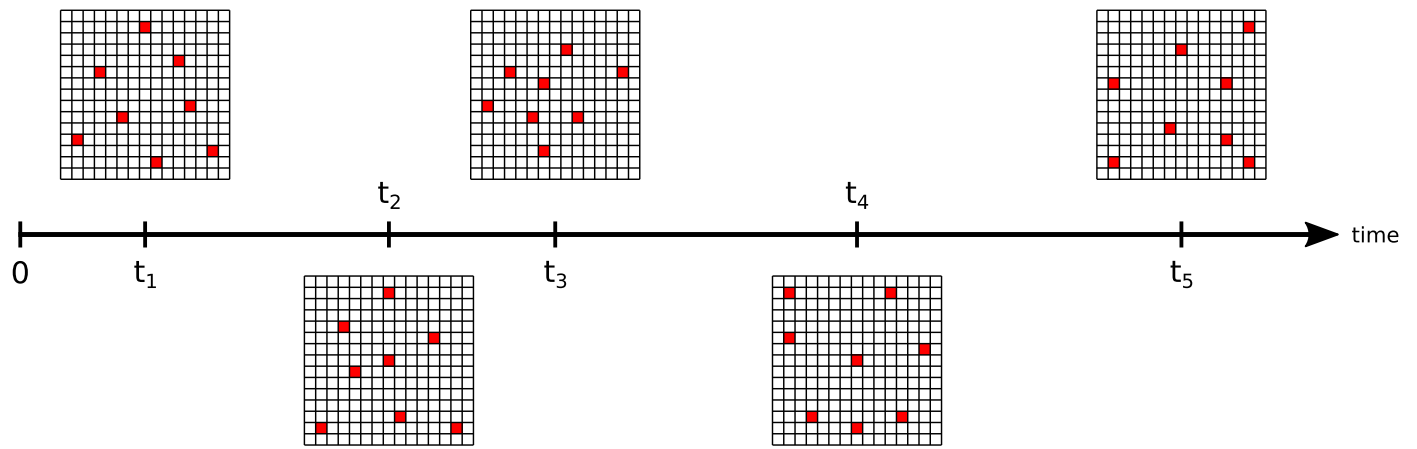
**Extended Data Fig. 3 | Numerical implementation of the model and parameter selection.** **a**, Snapshot of the numerical simulations of a system of size  $L = 25$ . The cells can adopt three different states, depicted with three colors: white (state  $S$ ), red (state  $D$ ) and blue (state  $U$ ). **b,c**, Scheme of the interaction neighborhood of a given cell  $ij$  and the absorbing boundary conditions used for the numerical simulations. Each cell is affected by its 8 nearest neighbors (Moore neighborhood). **d,e**, Plot of the interaction term of the transition rate given in

equation (15) for constant neighbor coupling  $\theta$  and constant parameter  $\mu$  respectively. **f,g**, Plot of the auxiliary functions  $\epsilon_1(\omega, \theta, \mu)$  and  $\epsilon_2(\omega, \theta, \mu)$  for  $\mu = \mu^* = 1$ . In each plot, the orange line denotes the region where the auxiliary functions are minimized. The crossing of both lines is denoted by blue arrows. **h**, Plot of  $\epsilon(\omega^*, \theta^*, \mu)$  for different values of  $\mu$ . The minimal value is reached at  $\mu^* = 1$ .



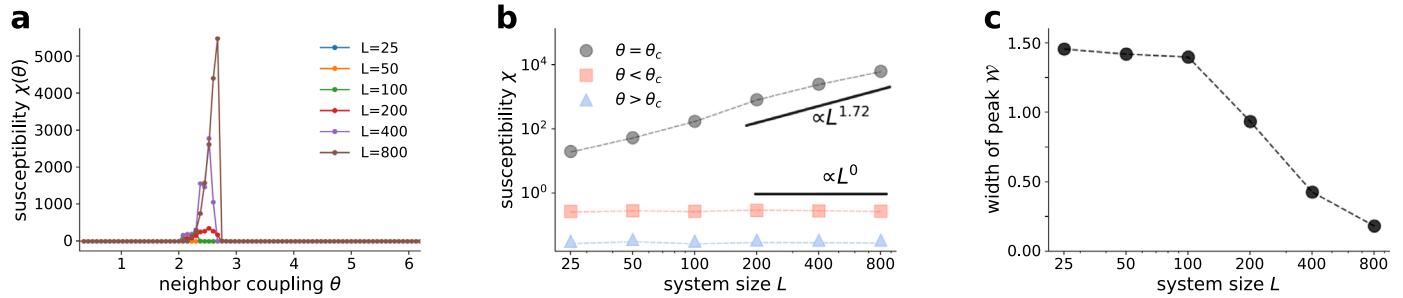
**Extended Data Fig. 4 | Results from the analysis of the empirical surface-activity signals. a-c,** Numerical values of the three characteristic quantities  $\langle \mathcal{A} \rangle_t$ ,  $\mathcal{A}^{th}$  and  $\mathcal{R}$  that we used to analyse the surface-activity signals. The grey-dashed vertical lines divide the data by days of acquisition on the field station. In **c**, we also show the mean value of the ratio  $\mathcal{R}_{emp}$  calculated over all videos.





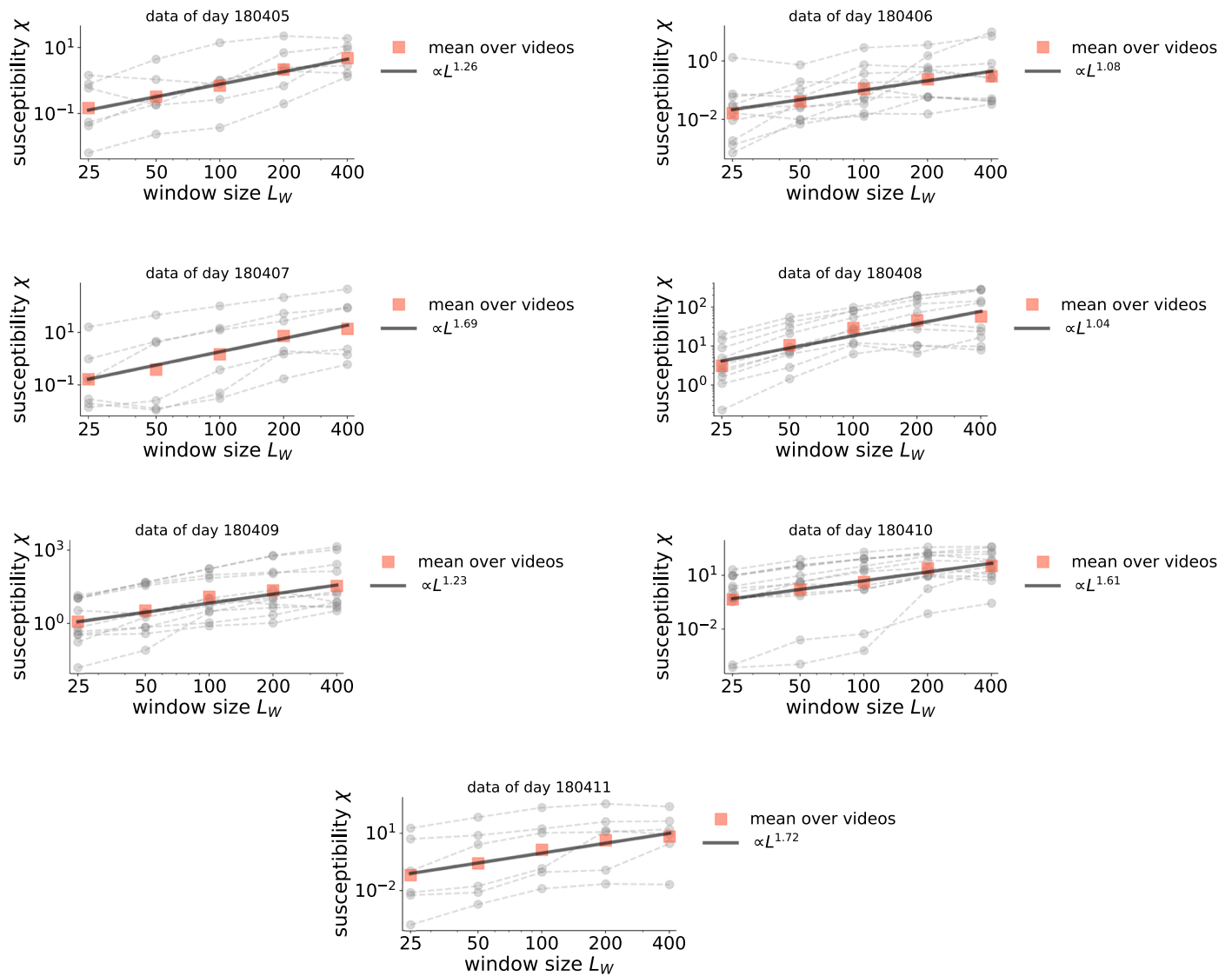
**Extended Data Fig. 5 | Average number of active neighbors.** Scheme of the computation of the average number of active neighbor cells for a given set of focal cells. The times  $t_i$  represent the randomly selected times at which a set of 8 focal cells are randomly selected. The focal cells are depicted in red. We

computed the number of active neighboring cells for each focal cell and averaged over all focal cells and over all randomly selected times  $t_i$ . The value obtained from the empirical videos is  $\mathcal{N}_{\text{emp}} = 0.27 \pm 0.23$ .



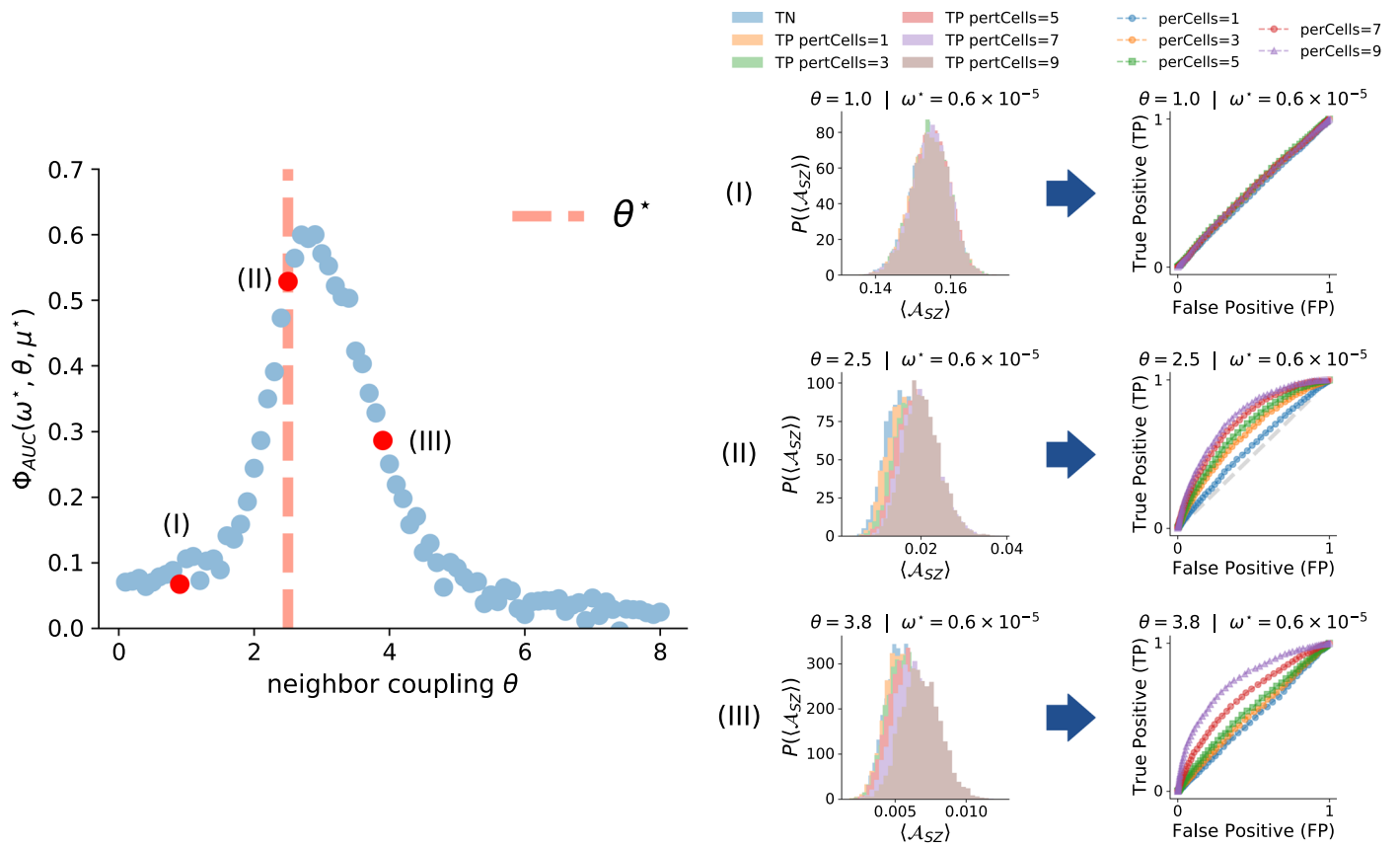
**Extended Data Fig. 6 | Study of the critical transition. a**, Susceptibility computed for six different system sizes  $L_w \in [25, 50, 100, 200, 400, 800]$  and for different values of the neighbor coupling parameter  $\theta$ . **b**, Value of the susceptibility in three different regimes: i)  $\theta = \theta_c$ , ii)  $\theta < \theta_c$  and iii)  $\theta > \theta_c$ , where  $\theta_c$  is

the value of the neighbor coupling parameter  $\theta$  where the susceptibility reaches its maximum value. **c**, Width of the susceptibility peak  $\mathcal{W}$  - as defined in equation (18) - for different system sizes  $L$ .



**Extended Data Fig. 7 | Susceptibility in empirical data.** Susceptibility computed as defined in equation (17) using the empirical videos and five different window sizes  $L_W \in [25, 50, 100, 200, 400]$  pixels. The results of the analysis for the different window sizes of each of the videos are connected by a gray dashed line. The average susceptibility over videos is plotted as a single red

square for each window size  $L_W$ . The best power-law fit to the average values is shown as a black line and the label shows the exponent that fits best the data. The data is shown separately for all acquisition days. In all cases, the best power-law fit has an exponent that is between 1 and 2.



**Extended Data Fig. 8 | Susceptibility analysis.** Plots of the true positive (TP) vs true negative (TN) curves for three different values of the coupling parameter  $\theta$ , as well as the plot of the average response to stimuli  $\Phi_{AUC}(\omega^*, \theta, \mu^*)$  for different values of  $\theta$ .

**Extended Data Table 1 | LSTM training parameters. Hyperparameters (architecture and training procedure) of the perturbation LSTM network**

Parameter	Value	Parameter	Value	Parameter	Value
Training episodes	30k	Learning rate	0.001	$L_2$ Weight decay $\lambda$	0.01
Batch-size	64	Number of epochs	20	Train-Test Split	0.8, 0.2
Optimizer	Adam	Activation function	ReLU	Learning rate schedule	Multiplicative decay
Embedding size	128	LSTM cells	64		

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Each of the videos was rectified to reconstruct a top-view of the pond. For this purpose, we positioned a 1.55m x 1.55m plastic square on the surface of the water at the end of each recording session. We rectified the videos using the four corner positions of the square and the OpenCV Python library. We then implemented the background subtraction process to each rectified video using the MOG2 background subtractor from the OpenCV library in Python to obtain black and white videos. The size of all the processed videos is 500 x 500 pixels.

Data analysis

Data and statistical analyses were carried out using dedicated Python scripts (Python >= 3.6). All statistical analysis are explained in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data necessary to reproduce the figures and results presented in this paper will be available through the institutional repository of the Humboldt University of Berlin.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<i>Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data exclusions	<i>Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Replication	<i>Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.</i>
Blinding	<i>Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).</i>
Research sample	<i>State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.</i>
Sampling strategy	<i>Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.</i>
Data collection	<i>Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.</i>
Timing	<i>Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Non-participation	<i>State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.</i>
Randomization	<i>If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.</i>

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Investigation of collective surface waves of large fish schools under natural conditions. Quantitative data extracted from videos acquired on the field.</i>
-------------------	---

Research sample	Video recordings of large shoals of sulphur mollies in their natural environment (freshwater, sulphuric streams/ponds in Tabasco, Mexico).
Sampling strategy	Quantification of individual waves across large sample of independent time window.
Data collection	All data were collected in the field at the Baños del Azufre sulphur spring system in Tabasco, Mexico.
Timing and spatial scale	Recordings over multiple hours over multiple days, single time window of 2min, duration of waves in the order of seconds
Data exclusions	Data exclusion was made according to pre-established criterion: direct perturbation by bird attacks (N=4 out of total=93 videos). Thus, we report the analysis of the remaining 89 videos.
Reproducibility	Observed collective behaviors were reproducible across different days, further observations across different years confirm reproducibility.
Randomization	N/A
Blinding	No blinding necessary because of automatic processing of the full dataset.
Did the study involve field work?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

## Field work, collection and transport

Field conditions	Subtropical climate; All field recordings were made during the dry season (April-May), temperatures in the range 30-40°C
Location	ur study location, the Banos del Azufre site (17°33'10.1"N 92°59'49.7"W), is part of an active sulphide spring complex in southern Mexico. It is located near the city of Teapa.
Access & import/export	The study location can be accessed easily by car. We obtained sampling permits from the Mexican government (H. Ayuntamiento Constitucional Tacotalpa, Direction Fomento Economico y Turismo) and numbers are included in the main text.
Disturbance	The disturbances to the system were kept at a minimum and were similar to natural occurring perturbances e.g. movement on the shore of the pond, naturally occurring bird attacks / bird overflights .

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	<i>Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Validation	<i>Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.</i>

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	<i>State the source of each cell line used.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>



Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines  
(See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

## Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

No laboratory animals involved.

Wild animals

Poecilia sulphuraria, all ages classes as present in the natural system, no capture of fish

Field-collected samples

No samples collected from the field.

Ethics oversight

The acquisition of the videos adhered to the "Guidelines for the treatment of animals in behavioral research and teaching" (Animal Behaviour 2021) and were approved by the Mexican government (DGOPA.09004.041111.3088, PRMN/DGOPA-009/2015, and PRMN/DGOPA-012/2017 issued by SAGARPA-CONAPESCA-DGOPA).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from [ClinicalTrials.gov](#) or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

## Dual use research of concern

Policy information about [dual use research of concern](#)

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | No                                  | Yes   |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Public health              |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> National security          |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Crops and/or livestock     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Ecosystems                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Any other significant area |

### Experiments of concern

Does the work involve any of these experiments of concern:

- | No                                  | Yes  |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Demonstrate how to render a vaccine ineffective                             |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Confer resistance to therapeutically useful antibiotics or antiviral agents |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enhance the virulence of a pathogen or render a nonpathogen virulent        |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Increase transmissibility of a pathogen                                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Alter the host range of a pathogen  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable evasion of diagnostic/detection modalities                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable the weaponization of a biological agent or toxin                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Any other potentially harmful combination of experiments and agents         |

## ChIP-seq

### Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

#### Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

#### Files in database submission

Provide a list of all files available in the database submission.

#### Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

### Methodology

#### Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

#### Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

#### Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

#### Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

#### Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

#### Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

## Flow Cytometry

### Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

- Sample preparation *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
- Instrument *Identify the instrument used for data collection, specifying make and model number.*
- Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
- Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
- Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*
- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

## Magnetic resonance imaging

### Experimental design

- Design type *Indicate task or resting state; event-related or block design.*
- Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
- Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

### Acquisition

- Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*
- Field strength *Specify in Tesla*
- Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
- Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*
- Diffusion MRI  Used  Not used

### Preprocessing

- Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
- Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
- Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
- Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

## Statistical modeling &amp; inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis:  Whole brain  ROI-based  BothStatistic type for inference  
(See [Eklund et al. 2016](#))

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

## Models &amp; analysis

n/a | Involved in the study

  Functional and/or effective connectivity  Graph analysis  Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.