Learning to control active matter

Martin J. Falk, Vahid Alizadehyazdi, Heinrich Jaeger, and Arvind Murugan
Department of Physics, University of Chicago, Chicago, Illinois 60637, USA

(Received 11 May 2021; accepted 25 August 2021; published 30 September 2021)

The study of active matter has revealed novel non-equilibrium collective behaviors, illustrating their potential as a new materials platform. However, most work treat active matter as unregulated systems with uniform microscopic energy input, which we refer to as activity. In contrast, functionality in biological materials results from regulating and controlling activity locally over space and time, as has only recently become experimentally possible for engineered active matter. Designing functionality requires navigation of the high-dimensional space of spatio-temporal activity patterns, but brute force approaches are unlikely to be successful without system-specific intuition. Here, we apply reinforcement learning to the task of inducing net transport in a specific direction for a simulated system of Vicsek-like self-propelled disks using a spotlight that increases activity locally. The resulting time-varying patterns of activity learned exploit the distinct physics of the strong and weak coupling regimes. Our work shows how reinforcement learning can reveal physically interpretable protocols for controlling collective behavior in non-equilibrium systems.

DOI: 10.1103/PhysRevResearch.3.033291

I. INTRODUCTION

Active matter has revealed exciting new patterns of self-organization not found in equilibrium systems [1,2]. Pioneering theoretical and experimental work explored the complexity generated by spatio-temporally uniform systems, in particular with uniform non-equilibrium microscopic driving across space and time. The resulting phenomena are sometimes seen as a step towards achieving complex functionality shown by biological materials. However, biological functionality, e.g., cytokinesis or cell migration [3–5], can be attributed to not merely being out of equilibrium but rather, to the ability to regulate activity as a function of space and time.

Recent experimental advances have demonstrated regulation of activity in diverse engineered systems, including bacteria, colloids, and reconstituted cytoskeletal components; while details differ, these experimental platforms allow for activity to be modulated as a function of space and time, usually through optical means [6–11].

However, we do not currently have systematic computational frameworks to exploit these new experimental techniques for manipulating active matter. The high-dimensional space of spatio-temporal protocols opened up by these experimental advances cannot be explored through brute-force alone. Furthermore, activity is a scalar field, and therefore it is not immediately clear how control of this quantity can achieve complex targets like spatial structure or net momentum transfer. For example, while a colloid can be induced to self-propel with light, light only controls the scalar speed at which the colloid self-propels, and not the vector direction. Consequently, previous work relied on system-specific physical intuition [11–17] or assumed complete knowledge of the underlying dynamical equations [18].

In contrast, data-driven approaches can be model-free and have shown promise for similar control problems [19] but typically with a few coupled degrees of freedom such as single-particle navigation [20–26], or bio-inspired locomotion [27–33]. These works have established data-driven techniques, in particular reinforcement learning [34], as a powerful tool for tackling control problems in physics. However, less attention [18,35] has been given to many-body non-equilibrium systems of the kind studied here.

Here, we address the challenge of control in active matter by leveraging developments in model-free reinforcement learning (RL) (Fig. 1). We construct an RL setup that identifies time-varying patterns of a scalar activity parameter capable of inducing directed transport in a simulated system of Vicsek-like self-propelled disks. As aligning interactions between the disks are increased from zero, the nature of learned protocols changes, illustrating the flexibility of the reinforcement learning approach. We find that the learned protocols can be physically interpreted in terms of the distinct underlying physics at weak and strong coupling.

In doing so, our goal is to demonstrate that reinforcement learning is a well-suited technique for achieving functionality in a broad class of active systems. While the system under consideration here is simple and canonical, it contains two physically very distinct regimes. The success we demonstrate in each regime is therefore indicative that the performance of the approach is not due to a unique aspect of the regime-specific physics. Our approach therefore promises to be a useful, model-free tool in confronting the high-dimensional protocol search problem that is universal to optically-activated active matter.

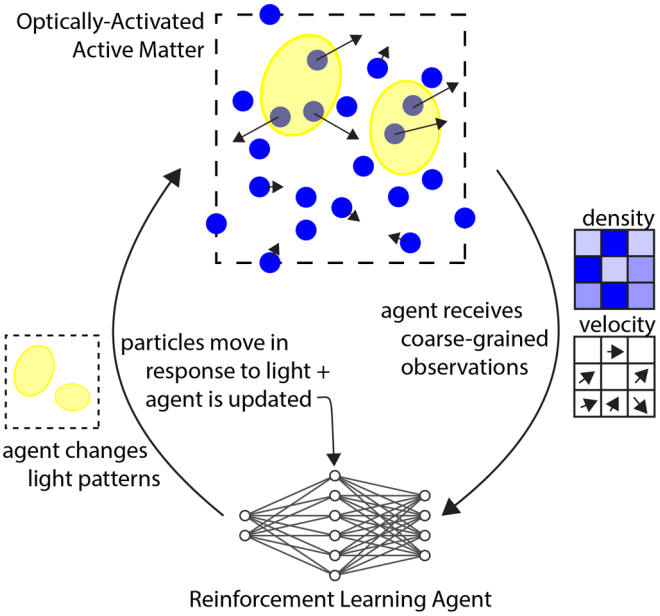

FIG. 1. Reinforcement learning provides a framework to control active matter by modulating activity over space and time. Recent experimental advances allow local modulation of activity by light in diverse active matter systems [6–12]. We consider a framework in which a reinforcement learning (RL) agent controls the illuminated region—its location, size, and shape—in order to achieve a desired non-equilibrium organization. Particle positions and velocity are coarse-grained, and passed to an RL agent, which decides where to place the light. The active matter system responds to the agent's choices of illumination; RL receives a reward based on this response and updates its control protocol accordingly.

II. METHODS

Simulation environment overview

More concretely, we set out to maximize directional transport in a 2D system of self-propelled particles by controlling activity. Particle positions are updated similarly to the canonical Vicsek model [2,36], but with the distinction that the magnitude of activity $v(\mathbf{x}, t)$ is a function of space and time:

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + v(\mathbf{x}, t)\mathbf{p}_i(t)\Delta t + \frac{1}{\gamma}F_{ex} + \eta, \quad (1)$$

$$\mathbf{p}_i(t+\Delta t) = (U_\theta \circ \mathcal{W})((1-k)\mathbf{p}_i(t) + k\bar{\mathbf{p}}_i(t)), \quad (2)$$

where $\mathbf{p}$ is the particle polarization, $\bar{\mathbf{p}}$ its local spatial average, Δt a timestep, U_θ a random rotation, $\mathcal{W}$ a normalization, and k a coupling in the interval [0,1), F_{ex} comes from a WCA excluded volume pair-potential, γ is a drag coefficient, and η is a spatial diffusion term. $v(\mathbf{x}, t)$ is now a generic spatiotemporal field controlling the speed (but not direction) of active self-propulsion.

As in recent experiments where spatio-temporal control of activity has been achieved optically [6–11], we assume that particles' polarities are unaffected by light. We stress that the only microscopic quantity changed by the light field is the active propulsion speed. All other quantities, such as rotational diffusion or coupling between polarities, are independent of optical activation. As noted, particles also experience excluded volume and a small amount of thermal noise; refer to Appendix A for more detail on how the simulations are implemented.

In what follows, we will formulate an RL setup for maximizing the x component of the system's momentum, in distinct physical regimes. We emphasize that the goal of maximizing +x momentum is relatively simple, but requires a nontrivial strategy; as light only controls the scalar speed at which a particle self-propels, not the vector direction, simply illuminating the particles will not produce transport in a specific direction (see Sup Movie 1 within the Supplemental Material [37]).

B. Formulating RL for active matter

Having described the simulation environment, we now turn to the selection of an appropriate algorithm for navigating the space of activity protocols. We also have to make choices about how this algorithm will interact with the simulation environment; we need to choose how to represent states, actions, and rewards. While most reinforcement learning algorithms are constructed with the goal of enabling successful optimization within a high-dimensional space of protocols, some algorithms will be better suited to the specific requirements of manipulating optically-activated active matter. We therefore make our choices motivated by the potential application of our approach beyond our simulated Vicsek-like environment.

1. Defining states, actions, and rewards

Reinforcement learning is commonly framed in the language of Markov decision processes, which contain four essential ingredients: states, actions, transitions, and rewards [34]. Transitions are determined by the physics of our active matter simulation, but the other three ingredients need to be defined as well.

There are many possibilities for defining states. We could, for instance, consider the state of the system to be a list of the positions and velocities of each particle. However, in order to respect the permutation invariance of the system, we instead construct coarse-grained density and velocity fields (Fig. 1). While it is possible that the coarse graining leads to violation of the Markov property for transitions between states, we assume that the grid is fine-grained enough to make these violations quantitatively small. Furthermore, this approach has the benefit of being easily extended to other common active matter systems, which are more readily described on larger length scales or in terms of fields.

Similarly, there are many possibilities for defining actions, corresponding to the different families of spatio-temporal activation fields. For simplicity, we constrain our optical field to be a single elliptical light source with fixed intensity, which we term the "spotlight". All particles within the spotlight experience the same active propulsion speed. All particles outside of the spotlight are inactive. The RL algorithm (Fig. 1) is allowed to take actions, which change the (a) center, (b) length, and (c) aspect ratio of the spotlight as a function of time, but do not change the intensity or tilt of the spotlight. We note that, by specifying that there is only one spotlight, we have placed constraints on the space of protocols considered by our RL setup. Any protocols identified in the RL procedure

are therefore only guaranteed to be locally optimal in this restricted space. However, our approach easily generalizes to families of protocols with multiple spotlights that may be necessary to consider in other active matter contexts.

As previously noted, our goal will be for the RL setup to maximize and maintain the x component of the system's momentum. Therefore, we choose to define the reward for any action as the subsequent instantaneous x momentum in the system. This is in contrast to other common use cases of reinforcement learning outside of physics, where the reward is frequently sparse in time.

2. Selecting an algorithm

There are a similarly wide array of possibilities for selecting a particular reinforcement learning algorithm. We wanted to choose one that would be well-suited to active matter systems in general. As such, we needed to find an RL algorithm that could take advantage of nonsparse rewards, naturally encode stochastic protocols, and accommodate continuous state and actions spaces.

These requirements suggested a class of algorithms known as online actor-critics [34]. Actor-critics have two components: An actor and a critic. The actor is a neural network, which receives coarse-grained density and velocity fields, as well as the current size and location of the spotlight. Based on this input, the actor samples a change to the pattern of light from a probability distribution, and receives its instantaneous momentum reward. Aiding the actor in its search is the critic network, which accepts the same state as the actor, but instead outputs an estimate of potential future rewards; in our case, time-integrated x momentum that can be gained in the future. Together, these two networks satisfy the requirements of selecting a reinforcement learning algorithm for an active matter context.

In order to update the two networks, the actor-critic algorithm makes use of the policy gradient theorem [34], which allows the actor loss function to be written as the product of the log probability of the actor sampling a particular choice for spotlight movement, and the temporal difference error δ_t. δ_t in turn is computed from the critic network, and can be thought of intuitively as the difference between the amount of x momentum seen in the system following spotlight motion, and how much x momentum the critic expected to see. The critic network is then updated with a loss function, which quadratically penalizes δ_t. These updates encourage the critic to become more accurate, and encourage the actor to move the spotlight in ways that will outperform the critic's expectations. Appendix B contains more detail on the actor, critic, states, actions, and rewards.

III. RESULTS

We begin by exploring control protocols for systems with different coupling k between particle polarities. Prior works have established how the physics of Vicsek-like systems qualitatively changes with this parameter [2,36], as well as the response of such systems to temporally-fixed quenched disorder of various kinds [38,39]. The no-coupling regime has been studied extensively as self-propelled hard spheres [40,41]. As the coupling is increased into a high coupling regime, the system crosses an alignment transition into a flocking phase (see Appendix E). In order to achieve +x transport, the RL policy should learn to break the symmetry of the particles' responses and exploit the distinct physics of the two regimes.

In all coupling regimes, the spotlight initially does not move in any meaningful fashion, and the net transport through the system is correspondingly low. As training proceeds, net transport through the system increases [Figs. 2(a) and 2(e)].

In the weak coupling limit, we find that the elliptical spotlight becomes fully elongated in the y direction, of finite length l_x in the x direction, and, on average, is moved at a characteristic velocity v_γ in the $+x$ direction [Figs. 2(b)–2(d)] (see Sup Movie 2 within the Supplemental Material [37]). A qualitatively distinct strategy with no well-defined spotlight speed and large fluctuations in size is learned at in the strong coupling regime [Figs. 2(f) and 2(g)] (see Sup Movie 3 within the Supplemental Material [37]).

A. Weak coupling

We next asked if we could obtain physical insight from our model-free learning approach, starting with the weak coupling limit.

Based on data in Fig. 2, we propose that the learned policy in the zero-coupling regime functions as a purification process. As the spotlight moves rightward, left-moving particles tend to exit the spotlight quickly, losing activity and reducing–x momentum. In contrast, right-moving particles tend to remain within the spotlight for longer because both move in the same direction [Fig. 3(a)], maintaining +x momentum.

We can quantify this intuition using a simplified 1D model in a region of length L with periodic boundary conditions, with an spotlight region of length $l < L$. Particles move with an active speed v_p when they are in the spotlight, and there is a conversion rate r of particles that switch their direction of motion per particle per unit time.

We limit analysis to a protocol where the spotlight moves to the right at a constant velocity v_γ and make several simplifying assumptions: (a) number density is a constant ρ_a within the spotlight and a constant ρ_i outside it. (b) Particles move with an active speed v_p when they are in the spotlight and only experience diffusive motion outside. (c) The fraction of left-moving particles is a constant f_a in the spotlight and f_i in the dark. (d) Particles instantaneously randomize their direction of motion upon exiting the spotlight. (e) The spotlight is a rectangular region of length l_x in the $+x$ direction, fully elongated in the y direction.

These assumptions are broken by real systems and by our simulated system. Furthermore, instances of the learned protocol show deviations that might reflect stochasticity inherent to learning and physical fluctuations such as nonuniform density, finite rotational decoherence time, and other violations of our assumptions. Nevertheless, we will show that this simple model of purification explains the time-averaged behavior of the learned protocol.

To make quantitative connection between the learned protocol and our purification model, we compute +x momentum as a function of purification model parameters. In the model,

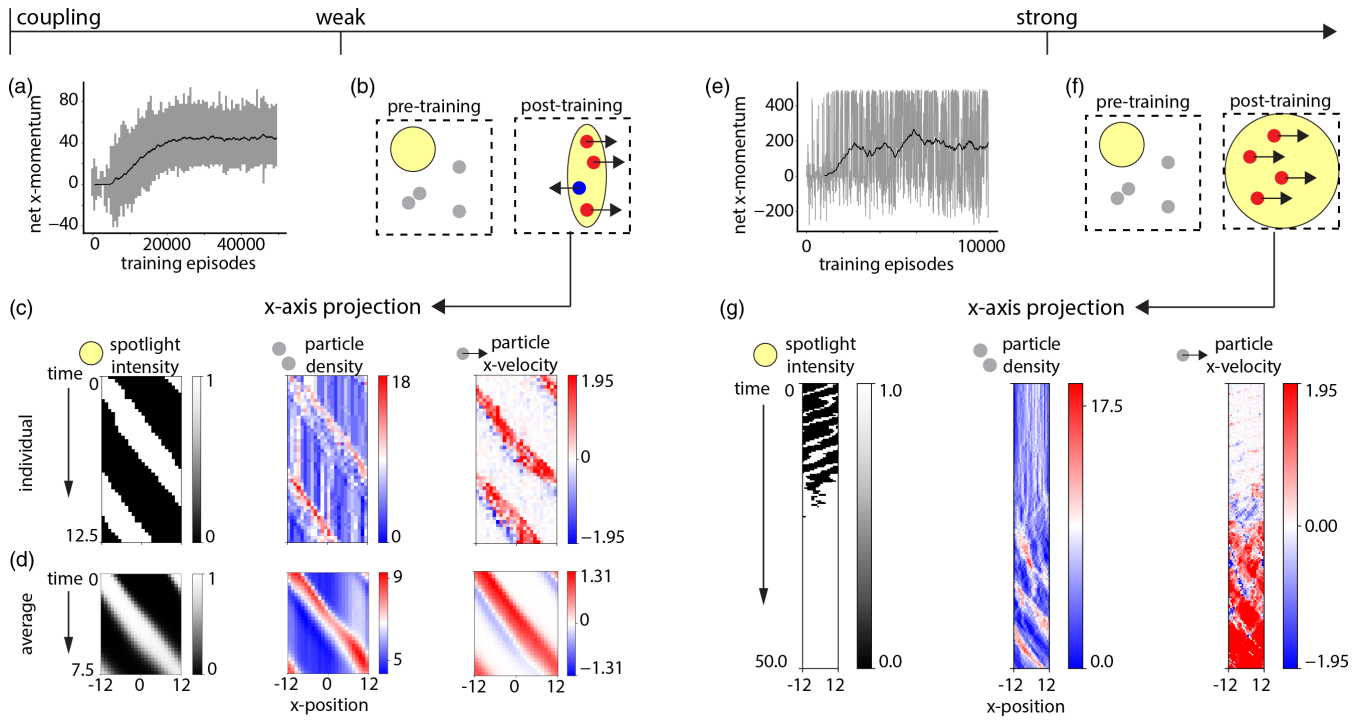

FIG. 2. Reinforcement learning generates distinct protocols to induce directional transport in self-propelled disks at weak and strong coupling. Weak coupling: (a)–(d); strong coupling: (e)–(g). [(a), (e)] As the RL setup trains, average x momentum during a training episode increases. X momentum remains highly stochastic (grey curves), but running average of (a) 500 or (e) 1000 episodes shows clear improvement (black curve). [(b), (f)] Training changes spotlight shape and movement in distinctive ways at weak and strong coupling in order to induce rightward transport. [(c), (g)] Kymographs of spotlight intensity, particle density, and particle x velocity for the learned protocols in the weak- and strong coupling regimes. Kymographs focus on the one spatial dimension important for the target behavior, but some information is lost in the x projection. Examples of the full system dynamics are available in Sup Movies 2,3 within the Supplemental Material [37]. (d) For weak coupling, averaging over multiple aligned periods of the learned policy shows that the spotlight moves from left to right at a well-defined velocity, with traveling wave of particle density with positive x velocity carried along.

we have four unknowns (ρ_a, ρ_i, f_a, f_i) with four constraints. (a) Particle number conservation:

$$\rho_a \tilde{l} + \rho_i(1-\tilde{l}) = \rho, \tag{3}$$

where ρ is the overall (linear) number density.

(b) Steady-state particle flux balance into the spotlight:

$$(1+\tilde{v})f_a\rho_a + \delta\tilde{v}(1-f_a)\rho_a = \tilde{v}\rho_i, \tag{4}$$

where $\delta\tilde{v}$ is $|1-\tilde{v}|$.

(c) Steady-state flux balance of left-moving particles into the spotlight:

$$(1+\tilde{v})f_a\rho_a + \tilde{r}\tilde{l}\rho_a f_a = \tilde{v}\rho_i f_i + \tilde{r}\tilde{l}\rho_a(1-f_a). \tag{5}$$

(d) Steady-state flux balance of left-moving particles into the dark region

$$\tilde{v}\rho_i f_i + \tilde{r}(1-\tilde{l})\rho_i f_i = (1+\tilde{v})f_a\rho_a + \tilde{r}(1-\tilde{l})\rho_i(1-f_i). \tag{6}$$

The equations then involve three nondimensional quantities, $\tilde{v} = \frac{v_\gamma}{v_p}$, $\tilde{l} = \frac{l}{L}$, and $\tilde{r} = \frac{rL}{v_p}$. To account for the excluded volume of the particles, we modify the system of equations above an adjustable density threshold ρ_{ev}. For further explanation of the various terms in the model, please see Appendix C.

We numerically solve these coupled equations, allowing us to compute a phase diagram for net x momentum as a function of $\tilde{v}$ and $\tilde{l}$ [Fig. 3(b)]. We fix $\tilde{r}$ and ρ_{ev} based on values measured in the simulation itself (Appendix D).

The phase diagram shows that maximum steady-state momentum is achieved when $v_\gamma \approx v_p$, consistent with the average velocity of our learned protocol [Fig. 3(b)] and similar to earlier physics-based single swimmer analyses of periodic activity pulses [14,15]. Static patterns of diffusivity variations have also been known to generate drift [42].

Our theory additionally provides a mechanistic explanation for the existence of an optimal length l_x for the spotlight, as we see in our RL-derived policy. Below this length, the spotlight is too small to accommodate more than a small number of particles, and excluded volume interactions prevent additional accumulation. Above this length, the spotlight is too big to purify the left-moving particles. Balancing these two competing effects yields an optimal spotlight length close to the length learned by our RL policy [Fig. 3(b)].

Finally, the structure of the equations suggests that, so long as we keep ρ_{ev} and $\tilde{r}$ fixed, the velocity of the spotlight v_γ should be approximately equal to the active speed v_p. This prediction is confirmed by simulations run at different v_p (with $\tilde{l} = .3$) [Fig. 3(c)].

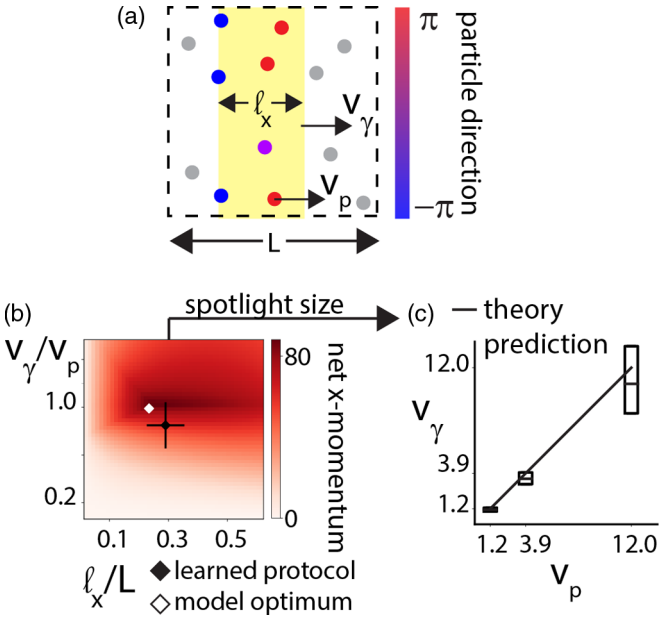

FIG. 3. Weak coupling protocol can be interpreted as a purification process. (a) As spotlight moves right, left-moving particles exit the spotlight (and become inactive) sooner than right-moving particles that cotranslate with spotlight. Consequently,–x momentum of left movers is quenched at the left edge of the spotlight while +x momentum is maintained within the spotlight. At steady state, density lost by exiting particles is replenished by particles that lie ahead of spotlight's path. (b) Phase diagram for 1D model in (a). X momentum as function of normalized spotlight velocity ($\frac{v_\gamma}{v_p}$) and normalized spotlight length ($\frac{l_x}{L}$). X momentum is maximized for $v_\gamma \approx v_p$ and for intermediate l_x (black star), close to parameters identified by reinforcement learning (RL) (white star, black-dotted line). Error bars represent the standard deviation of distributions for the trained protocol. (c) We trained RL on self-propelled systems with different light-enhanced activity v_p. Spotlight speed v_γ scales with v_p as predicted by theory (black line). Boxplots extend from lower to upper quartile of velocity distribution, with a line at the median.

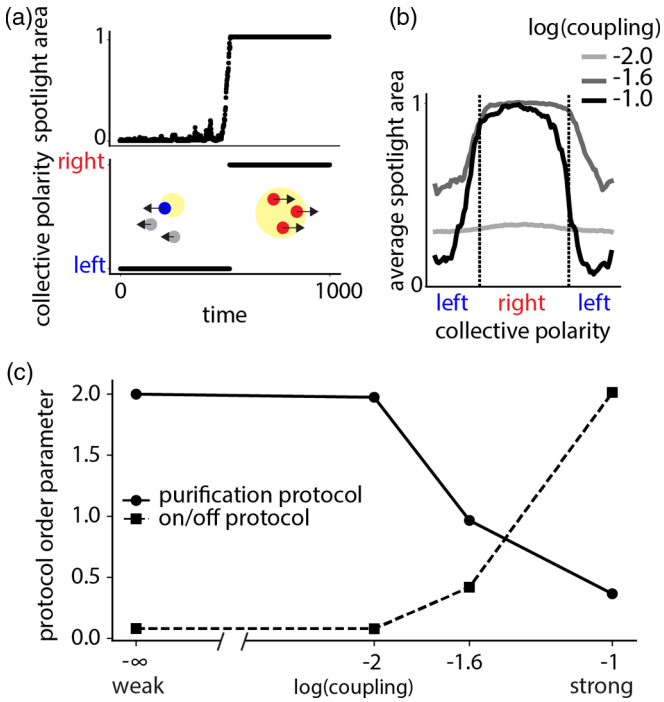

FIG. 4. Order parameters quantify how learned protocols switch from purification at weak coupling to a flocking-based strategy at strong coupling. (a) Snapshot traces of collective direction (bottom) and ratio between area of spotlight and system area (top) for the strong coupling protocol. Measurements of the two quantities are made at the same time in simulation. (b) Average spotlight area as a function of collective polarity. At strong and intermediate coupling, the spotlight provides maximum illumination when the collective polarity points rightward, and provides less illumination when the collective polarity points leftward. There is no relation between collective polarity and spotlight size at weak coupling. This transition is consistent with correlations between spotlight area and collective polarity as a function of coupling (Fig. 8). (c) Order parameters $\langle P \rangle$, $\langle O \rangle$ quantitatively track nature of protocols from weak to strong coupling.

B. Strong coupling

The learned protocol at strong coupling does not have a well-defined spotlight velocity. Instead, spotlight size is correlated with the polarity of particles.

We find that the strong coupling protocol exploits flocking physics inherent to this regime of the Vicsek model [2]. Due to the coupling, there is limited heterogeneity in the polarity of individual particles, creating a well-defined collective polarity (Appendix E). To understand the protocol identified in the strong coupling regime, we compared the collective polarity of the system to the area of the spotlight [Fig. 4(a)]. We found that RL maximized the area of the spotlight when collective polarity pointed in the desired direction, and fluctuated the spotlight area when collective polarity pointed in the undesired direction [Fig. 4(b)]. This protocol has appealing parallels to other on/off strategies studied, but in the context of single-colloid navigation [20–22].

In fact, we can systematically distinguish the learned strategies in the two coupling limits by defining two order parameters that characterize the control protocol. We define the purification parameter $\langle P \rangle$ to be the inverse ratio of the standard deviation and the mean of the v_γ distribution; intuitively, $\langle P \rangle$ is high if the spotlight has persistent motion in one direction. We define the on/off parameter $\langle O \rangle$ to be the ratio of spotlight area when the collective polarity points right versus left; intuitively, $\langle O \rangle$ is high if the spotlight's intensity is correlated with the collective polarity of particles within the spotlight.

We repeated the learning algorithm for coupling values between the strong and weak regimes [Fig. 4(c)]. We find that $\langle P \rangle$ is high for low coupling and falls with increasing coupling while $\langle O \rangle$ is high for strong coupling and falls with decreasing coupling. Thus, the model-free RL setup learns distinct protocols to exploit different physics at different coupling values.

In the crossover regime [$\log(k) = -1.6$] where spontaneous collective motion begins to emerge, the learned protocol adopts aspects of both the purification and on/off strategies [Fig. 4(c)]. Like the strong coupling protocol, the agent has maximal spotlight area when the collective polarity points to

the right [Fig. 4(b)]. However, when the collective polarity points to the left, the spotlight area is still half its maximal value on average, potentially reflecting the larger spread of individual particle polarities in the crossover regime.

Finally, we repeated the learning procedure at strong coupling but lower densities; in this regime, the particles break up into clusters, each with a tightly coupled alignment (Sup Movie 4 within the Supplemental Material [37]). The learned protocol is harder to directly interpret but achieves a peak momentum transfer closer to the weak coupling regime, despite exhibiting characteristics more similar to the policy learned in the strong coupling regime (Appendix F, Fig. 7). We additionally repeated the learning procedure at the original density but in larger systems with 4 times the number of particles. In this case, the RL setup re-identified the same strategies as it did in the smaller systems (Fig. 9).

IV. CONCLUSIONS

Over the past two decades, the physics of active matter systems with homogeneous activity has been illuminated with great success. Our work proposes that reinforcement learning can be used to explore the collective physics of active particles in spatio-temporally complex environments. For the simple models investigated here, we were able to extract physical insight from our initially physics-blind approach, in the strong and weak coupling regimes. Such insight is valuable, particularly given known concerns about the ability of RL to learn reproducible protocols [43].

Control of active matter by modulating where and when energy is dissipated is broadly applicable since microscopic energy dissipation (i.e., activity) is a universal aspect of active matter systems ranging from bacteria to colloids. As such, we envision that the approach we applied here to particulate active matter can be readily extended to systems where the features of interest are emergent, e.g., topological defects [6,7]. In those systems, nonlocal effects, e.g., the change in nematic texture due to the motion of topological defects, also suggest the possibility that more complex functional goals will require counter-intuitive solutions. Similar considerations might also apply to geometrical constraints introduced by confining active matter within deformable containers [44–47], or attempting to manipulate an active container filled with passive matter [48–50]. Broadening the nature of the problem, it might also be fruitful to consider the application of reinforcement learning to design the interaction protocols between objects, which carry their own sources of illumination [51]. We therefore propose that reinforcement learning provides an appealing, model-free method for generating intuition and functionalizing the effects of localized activity in systems hosting topological excitations or otherwise complex dynamics.

ACKNOWLEDGMENTS

We are indebted to Jonathan Colen, John Devaney, Ryo Hanai, Kabir Husain, Shruti Mishra, Riccardo Ravasio, Matthew Spenko, and Bryan VanSaders for insightful feedback on the manuscript. We would also like to thank Weerapat Pittayakanchit and Steven Redford for helpful discussions at the beginning of the project. This work was primarily supported by NSF EFRI Grant No. 1830939. We acknowledge the University of Chicago's Research Computing Center for their support of this work.

APPENDIX A: SIMULATION DETAILS

Simulations are run in two dimensional periodic boundary conditions using HOOMD-blue [52] (2.9.0). Translational dynamics of the system are fairly simple. A WCA potential between particles is used to enforce excluded volume beginning at a radius of .5. Positions are updated under Langevin dynamics, with a drag coefficient of 5, and a small kT of .03. Self-propulsion is incorporated into the dynamics via the addition of a constant force whose magnitude is the product of the drag coefficient and the active speed referenced in the text. Unless otherwise noted, this active speed is set to 3.9. The direction of the force is updated every five timesteps, and points along the particle's instantaneous polarity at the time of the update.

The angular dynamics of the polarity are also fairly simple. Each particle carries with it a unit vector polarity, which points in the plane. Every five timesteps, this polarity is rotated by a random angle drawn from the distribution $\frac{\pi}{\sqrt{1000}}\mathcal{N}(0,1)$.

For simulations that incorporate a Vicsek-like coupling between physically proximal particles, every five timesteps the polarity update begins by computing a list of each particle's neighbors, which are within a radius of 1.33, including the central particle. In the following the central particle will be referenced to with the index i. From this list of particles, a mean polarity $\overline{p_i}$ is computed. The polarity p_i of particle i is then updated to be $(1-k)p_i + k\overline{p_i}$, where the coupling k can take on values in [0,1). The updated polarity is then normalized to be of unit length. As before in the noninteracting case, each polarity is subsequently rotated by a random angle drawn from the distribution $\frac{\pi}{\sqrt{1000}}\mathcal{N}(0,1)$.

In all simulations performed here, 144 particles are initialized on a square lattice with an overall number density of 0.25 particles per unit area. The mass of each particle is set to 1. Polarities are initialized by drawing from a uniform distribution between 0 and 2π. Timesteps were set to be 5×10^{-3}.

APPENDIX B: LEARNING ALGORITHM DETAILS

We implemented a simple TD actor-critic algorithm [53] based on the implementation found in Ref. [54], using Tensorflow (2.0.0).

Reinforcement learning is based on the framework of Markov decision processes (MDPs), which involve a time-series of states, actions, and rewards. The formulation of the three essential components of the states, the actions, and the rewards are independent of the specific reinforcement learning algorithm. We outline those three key components before briefly describing our implementation of the actor-critic algorithm.

1. States, actions, rewards

States are represented by concatenating the coarse-grained number density field, the coarse-grained x-velocity field, the

coarse-grained y-velocity field, and the current position and shape of the spotlight(s). The fields are all flattened into vectors before concatenation. In our current study, all fields are 3×3, so the size of the state space is $3\times(3\times3)+4\times(\#spotlights)$, which comes to a total of 31. The coarse grained velocity fields are constructed by averaging the velocities of all the particles in a particular grid square; if no particles are present, then the velocity is assigned to be zero. This procedure is accomplished using the SciPy function binned_statistic_2d, and the corresponding density field construction is done using the NumPy function histogram2d. While it is possible that the coarse-graining leads to violation of the Markov property for transitions between states, we assume that the grid is fine-grained enough to make these violations quantitatively small.

The action space is of dimension $4\times(\#spotlights)$. Each arm is specified as a 4-tuple: $(x, y, ratio, length)$ where x is the x position of the center of the arm, y is the y-position of the center of the arm, $ratio$ is the ratio between the extent of the spotlight in the x direction compared to the y direction, and $length$ is the extent of the spotlight in the x direction. In other words, the region activated by a given arm is an ellipsoid centered at (x, y), with an x dimension of $length$ and a y dimension of $\frac{length}{ratio}$. However, the output of the actor network is not directly these variables, but is instead a vector $(\delta_1, \delta_2, \delta_3, \delta_4)$. This effectively regularizes and constrains the policy to be relatively continuous in time. This vector then updates the current $(x, y, ratio, length)$ to the following values:

$$x \to (x + d\delta_1 + l_x)\%(2l_x) - l_x, \tag{B1}$$

$$y \to (y + d\delta_2 + l_y)\%(2l_y) - l_y, \tag{B2}$$

$$ratio \to \text{clip}(ratio + .1\delta_3, .1, 3), \tag{B3}$$

$$length \to \text{clip}(length + .1\delta_4, .2l_x, 1.8l_x), \tag{B4}$$

where clip is a function that clips the first entry to be within the bounds in the second and third entries, % is the mod function, and l_x, l_y are half the widths of the simulation box in the x and y dimensions respectively. For the intermediate and strong coupling regimes discussed in Fig. 4, the constraint that l_x must be smaller than the full simulation box length is relaxed, and the spotlight is allowed to occupy the full volume of the simulation box:

$$length \to \text{clip}(length + .1\delta_4, .2l_x, 2.2l_x). \tag{B5}$$

Note that d sets the maximum distance the agent can move the spotlight center between updates. On physical grounds, this should not be faster than the maximum distance a particle can move between updates. Hence, for a given level of activity, we set d so that the spotlight can move 4x the distance that particles can move in the time between spotlight updates.

In order to improve training, we preprocess states before they are fed into the actor and the critic. Specifically, we generate 10^5 random samples of states the agent is likely to encounter, and then using the sklearn StandardScaler function to define a function, which scales states relative to the random samples. For density fields, this means we sample a 3-by-3 matrix where each entry is drawn from a uniform distribution between 0 and 1, and then normalize and scale this matrix so that the sum is equal to the number of particles in the system. For velocity fields, we sample two 3-by-3 matrices where each entry is drawn from a uniform distribution between 0 and 1, and then scale them so that entries run between $+/-$ the active speed of the particles. For the entries corresponding to the spotlight descriptors, we sample four scalars from the [0,1] uniform distribution and scale the first so that it runs between $+/-l_x$, the second between $+/-l_y$, the third between .1 and 3, and the fourth between $0.2l_x$ and $1.8l_x$. For the intermediate and strong coupling regimes, the fourth was sampled between $0.2l_x$ and $2.2l_x$

The reward is simply the sum of the x velocities in the system.

2. TD actor-critic

Given this setup for states, actions, and rewards, we now describe implementation of a simple TD actor-critic agent; more detail can be found in standard RL references, e.g., Ref. [34].

Our actor-critic agent has two neural networks, the actor and the critic. We represent the critic as a simple feed-forward neural network with two hidden layers, each with 400 neurons, and a single output node. The actor is represented as a feed-forward neural network with two hidden layers, each with 40 neurons, and two output layers, each with $4\times(\#spotlights)$ nodes. Both networks use ELU activation functions for the hidden layers. For the actor, one output layer has a linear activation, while the other has a softplus activation. The one output node for the critic has a linear activation.

During training, the actor receives a state, and its outputs are used to parameterize the means and standard deviations of $4\times(\#spotlights)$ Gaussian distributions. These distributions are then sampled and the samples are clipped to be between $+/-1$. These sampled numbers are then turned into actions as described in Appendix B1, the system transitions to its next state according to the physics described in Appendix A, and then a reward is generated based on the next state.

In order to update the two networks, the actor-critic algorithm makes use of the policy gradient theorem [34], which allows the actor loss function to be written as the product of the log probability of having sampled the action, and the temporal difference error δ_t. δ_t in turn is computed from the critic network, which is a bootstrapped approximation for the difference between the value the critic network assigns to the next state and what value it should actually be. Since this is a bootstrapped estimate, we approximate what the value should actually be the sum of the reward received in the transition and the value of the previous state. In order to insure that the critic network outputs will not diverge, the value of the previous state is discounted by a scalar $0 < \gamma < 1$. The critic network is then updated with a loss function, which quadratically penalizes δ_t. All updates are performed just for one learning step, and with a batch size of one, using the Adam optimizer.

In regards to hyperparameters, we attempted to choose standard values, which did not need much changing in between the various parameter regimes we considered in this work. The size of the networks remained the same, and for all networks trained, γ is set to $1-10^{-3}$. The learning rate is 10^{-4}

for the critic and 2×10^{-6} for the actor. We initially start these values to be 10 times higher, and then quadratically decay the learning rate to their stated values using the tensorflow polynomial_decay function.

Training is done in episodes of 100 training steps, over which we record the total reward. Each training step consists of 50 timesteps. Every 900 episodes, the system restarts to the initial square lattice with randomized polarities. Figure 2 agent training was run for 4.95×10^4 episodes, Fig. 3 agents were trained for 6.3×10^4 episodes, and in Fig. 4, the low coupling agent was trained for 2.7×10^4 episodes, while the high and intermediate agents were trained for 9×10^3 episodes.

APPENDIX C: THEORY DESCRIPTION

Here we provide a more detailed description of Eqs. (3)–(6), as well as a description of the modifications to account for excluded volume effects.

Equation (3) is a statement of particle number conservation in our system, which can be written as

$$\rho_a l + \rho_i(1-l) = N_{\rm tot}. \tag{C1}$$

The first term on the right -hand side (rhs) represents the number of particles in the active region, and the second term represents the number of particles in the inactive region. Dividing through by L yields Eq. (3).

Equation (4) is a statement that at steady-state, the number of particles exiting and entering the active region must be equal:

$$(v_p+v_\gamma)f_a\rho_a + |v_p - v_\gamma|(1-f_a)\rho_a = v_\gamma\rho_i. \tag{C2}$$

The rhs first term represents left-moving particles exiting from the back of the active region either because the region has moved past them, or they have propelled themselves into the inactive region. The second term represents right-moving particles exiting either from the back or the front of the active region due to mismatch between the active velocity and the velocity of the active region. The left-hand side (lhs) accounts for inactive particles moving into the active region as the active region advances. Dividing through by v_p yields Eq. (4).

Equation (5) is a statement that at steady-state, the number of left-moving particles exiting and entering the active region must be equal, and this conservation is independent of the balance of total number of particles entering and exiting:

$$(v_p+v_\gamma)f_a\rho_a + rl\rho_a f_a = v_\gamma\rho_i f_i + rl\rho_a(1-f_a). \tag{C3}$$

The rhs first term represents left-moving particles exiting from the back of the active region. The second term represents the loss of the bulk left-moving active population, as formerly left-moving particles flip direction to become right-moving particles. The lhs terms analogously represent incoming left-moving particles from the inactive region, and addition from particles in the active bulk that switch from right to left. Dividing through by v_p yields Eq. (5).

Finally, Eq. (6) is a statement that at steady-state, the number of left-moving particles exiting and entering the inactive region must be equal:

$$v_\gamma\rho_i f_i + r(1-l)\rho_i f_i = (v_p+v_\gamma)f_a\rho_a + r(1-l)\rho_i(1-f_i). \tag{C4}$$

The physical meaning of the terms is analogous to Eq. (C3), and dividing through by v_p yields Eq. (6).

In order to account for excluded volume effects, we assume that density within the active region is capped at a value ρ_{ev}. If Eqs. (3)–(6) initially provide a solution where $\rho_a > \rho_{ev}$, then we instead fix $\rho_a = \rho_{ev}$. Note that Eq. (3) still holds, and therefore this implies that both ρ_a and ρ_i are fixed. What prevents Eqs. (4)–(6) from being over-determined is that the physics requires an additional variable W to be introduced in order to account for the particles that are pushed out of the active region as a result of the excluded volume interactions. In this regime, we solve the following set of equations:

$$(v_p+v_\gamma)f_a\rho_a + |v_p - v_\gamma|(1-f_a)\rho_a + W = v_\gamma\rho_i, \tag{C5}$$

$$(v_p+v_\gamma)f_a\rho_a + rl\rho_a f_a + W f_a = v_\gamma\rho_i f_i + rl\rho_a(1-f_a), \tag{C6}$$

$$v_\gamma\rho_i f_i + r(1-l)\rho_i f_i = (v_p+v_\gamma)f_a\rho_a + r(1-l)\rho_i(1-f_i) + W f_a. \tag{C7}$$

These equations are identical in meaning to Eqs. ((C2)–(C4), the only difference being the mean-field accounting for the extra population of left-moving particles carried away from the active region by W.

APPENDIX D: PARAMETER ESTIMATION

In order to compute our numerical phase diagram in Fig. 3 B, we need to set two parameters r and ρ_{ev}.

In order to estimate r, we followed the same process that generates polarity diffusion in our simulation. We generated 144 random walks of length 100 000, where every five timesteps, the value of the walk was changed by a sample drawn from the distribution $\frac{\pi}{\sqrt{1000}}\mathcal{N}(0,1)$. Particle polarities were initialized uniformly around the unit circle. Operationally, r is defined as the fraction of particle polarities, which switch x direction in a unit of time. As the value of the timestep in our simulation is 0.005 time units, we then downsample our random walk by taking every 200th entry. We then count the number of times the cosine of the random walks switches sign, and divide that total sum by the number of random walks, and the length of the downsampled walks. In doing so, we find that for our system, $r = 0.32$.

In order to estimate ρ_{ev}, we make the crude assumption that the ρ_{ev} is the active density ρ_a for any optimal protocol, and specifically the optimal protocol found in Fig. 2. This assumption is based on the intuition that the optimal protocol seeks to maximize ρ_a until excluded volume effects saturate the active region, which can be seen in the resultant phase diagram in Fig. 3(b). Measuring this from simulations of the fully converged protocol found in Fig. 2 yields a linear number density of $\rho_{ev} \approx 9$. Since the simulation box is a square of length 24, this corresponds to an area number density of 0.375.

APPENDIX E: ALIGNING INTERACTIONS IN THE VICSEK MODEL

Central to our analysis of the strong coupling regime is the existence of a well-defined "collective polarity", in the sense

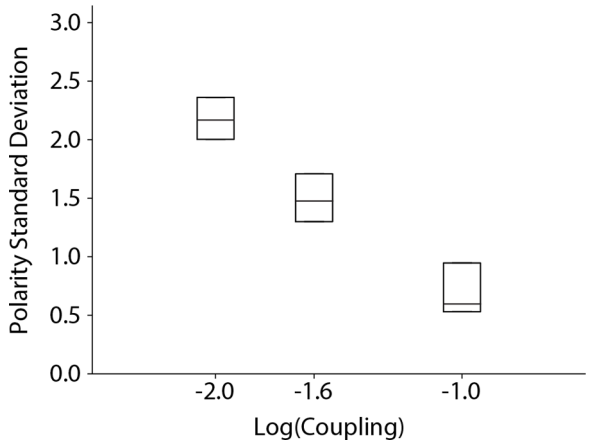

FIG. 5. Increasing coupling sharpens the collective polarity distributions generated in fully-trained trained protocols. For a range of alignment couplings, we train reinforcement learning agents and evaluate the polarity distributions their converged policies generate. We compute the circular standard deviation at each instance in time and create boxplots to assess the resulting distributions. Each distribution consists of $N = 2.0\times10^5$ circular standard deviations, drawn from the frames of the same simulations analyzed in Fig. 4. Each circular standard deviation in turn is computed from the polarities of the $n = 144$ particles in the simulation. Boxplots indicate the lower and upper quartiles, with an interior line indicating the median. As coupling increases, the polarity standard deviation decreases, indicating the development of a well-defined collective polarity.

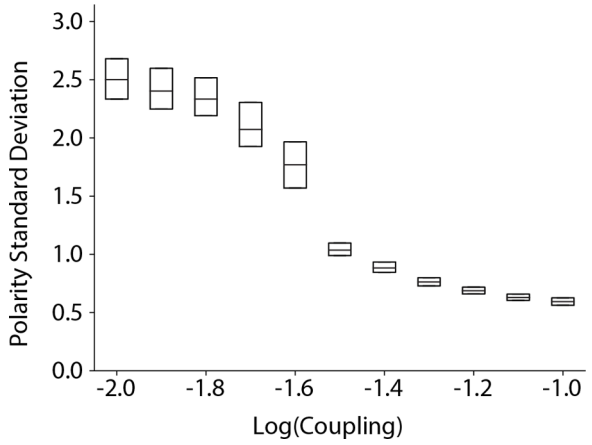

FIG. 6. In a fully-activated Vicsek system, increasing coupling sharpens the collective polarity distribution. For simulations with permanently activated particles run at a range of polarities, we compute the circular standard deviation at each instance in time and create boxplots to assess the resulting distributions. Each distribution consists of $N = 5.0\times10^3$ circular standard deviations, drawn from the frames of a simulation where all particles are activated and coupled at the corresponding coupling values. Each circular standard deviation in turn is computed from the polarities of the $n = 144$ particles in the simulation. Boxplots indicate the lower and upper quartiles, with an interior line indicating the median. The system exhibits a crossover regime for polarity standard deviation at a coupling of approximately $k = 10^{-1.5}$, where the standard deviation decreases and the distribution of standard deviations becomes more peaked as well. This indicates the development of a well-defined collective polarity.

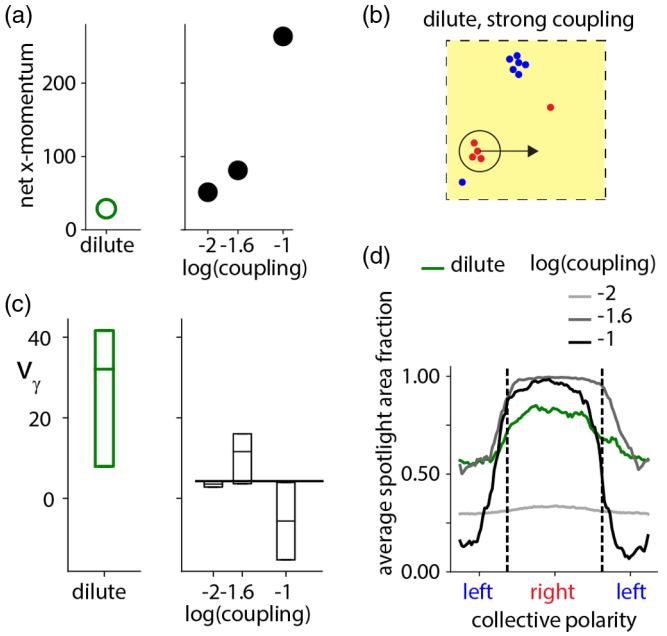

FIG. 7. Learning to induce transport in a strongly coupled dilute system with flocking domains. All nondilute data are taken from simulations discussed in Fig. 4. (a) In the dilute regime, average x momentum transferred by a fully-trained reinforcement learning agent is approximately similar to the amount in the weak coupling regime. (b) Schematic of flocking domains in the dilute, strongly-coupled regime. (c) We evaluate the spotlight x velocity v_γ in the dilute regime, as well as over the range of couplings discussed in Fig. 4. At weak coupling, the spotlight moves at the speed predicted by theory (black line). At intermediate and strong coupling v_γ diverges from the prediction and additionally is more stochastic. In the dilute regime, v_γ has an even larger spread. Boxplots extend from lower to upper quartile of velocity distribution, with a line at the median. (d) Average spotlight area as a function of collective polarity. In the dilute regime, the spotlight is larger when the average polarity points to the right.

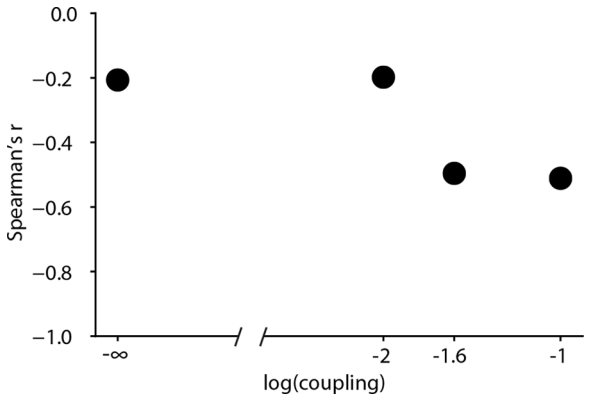

FIG. 8. Spotlight size is anticorrelated with collective motion away from the preferred direction at strong couplings. We calculate Spearman's r between spotlight area and the absolute value of the angle between collective polarity and the +x direction for increasing values of coupling. As coupling increases, spotlight area and angle magnitude become more strongly anticorrelated.

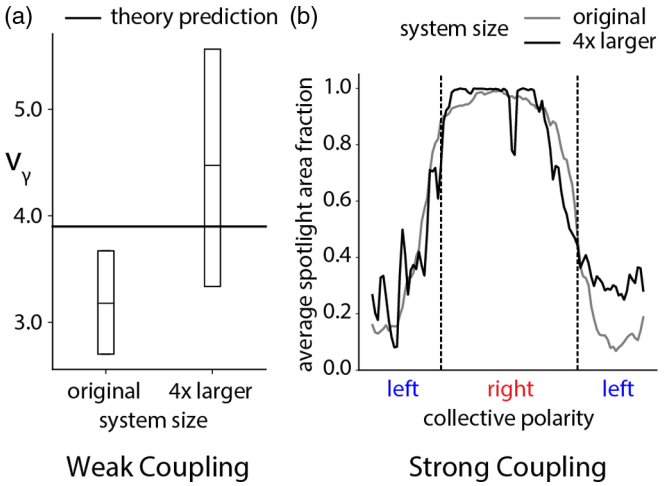

FIG. 9. Learning on 4x larger systems recapitulates the same strategies found in smaller systems for both weak and strong coupling regimes. Simulations are run with 576 particles, keeping all other physical parameters constant unless specifically noted. The RL approach and hyperparameters were also identical to those used in the smaller systems. (a) In the weak-coupling regime with $k = 0$, we find a qualitatively similar strategy to the one identified in Fig. 2, where the spotlight extends laterally and moves rightward at a well-defined velocity, close to the prediction of the simple model proposed in the main text; see also Sup Movie 5 within the Supplemental Material [37]. Boxplots extend from lower to upper quartile of velocity distribution, with a line at the median. Training was performed for 1.4×10^4 episodes. (b) In the strong coupling regime with $k = .1$, we find a quantitatively similar strategy to the one identified in Fig. 2, where the spotlight rectifies rightward motion by increasing activity when particles collectively point to the right; see also Sup Movie 6 within the Supplemental Material [37]. Training was performed for 4.2×10^3 episodes.

that the majority of particles have their polarities aligned in the same direction at any given point in time. That this quantity exists is not surprising considering the well-developed literature on the flocking transition in Vicsek and Vicsek-like models, though the order of the transition is sensitive to computational details [55].

Here, we can identify on which side of the transition the activated system is using qualitative visual indications (Sup Movie 3 within the Supplemental Material [37]), and also on a quantitative analysis of the standard deviation of the angular distribution of polarities across time (Fig. 5). These distributions are drawn from the same simulations used to generate Fig. 4. We find that as the alignment coupling increases from weak to strong, the distribution of particle polarities has lower average circular standard deviations. Therefore, even during the periods of inactivity that characterize the agent's policy in the strong coupling regime, the system retains a relatively well-defined collective polarity.

The physics of the Vicsek model do not apply to the inactive periods, and therefore we should not necessarily expect to see a collective flocking polarity during the inactive periods. However, it is entirely possible that the strength of the coupling and the spatial density of the particles allow for long decorrelation times of polarities of the individual particles, which were aligned during the active periods. If the decorrelation timescale is long compared to the inactive periods of the policy, then we should still expect to observe a well-defined collective polarity, as we do in the strong coupling regime (Fig. 5).

To further evidence that the physics of the Vicsek model [2] underlie the policy learned by the agent in the strong coupling regime, we run simulations identical to those discussed in the main text, except that all particles are activated (Fig. 6). We do this across the range of couplings explored in Fig. 4. As before in Fig. 5, we see that the collective polarity becomes more well defined at stronger couplings. Additionally, the spread of the distribution of circular standard deviations collected at different time points decreases with coupling. This decrease indicates that in the strong coupling regime, the majority of particles are aligned the majority of the time, as long as they are constantly activated.

The wider spread at the equivalent coupling values in Fig. 5 are therefore likely to be the result of decorrelation during periods of inactivity.

Measurements of the circular standard deviation were performed using SciPy's circstd function.

APPENDIX F: DILUTE REGIME POLICY

While Fig. 4 in the main text reports on policies learned for generating transport in systems, which exhibited a system-spanning flocking transition, we were also interested in how an RL agent might respond to a system with strong coupling but no system-spanning collective variable. This is precisely the sort of system, which is realized by the Vicsek model when simulated in a dilute regime [2]. In the dilute regime, the system-wide flocks break up into flocking domains, with different domain-scale collective polarities (Sup Movie 4 within the Supplemental Material [37]).

Following training, we find that an RL agent can learn to induce positive x-momentum in a dilute regime [Fig. 7(b)], to a degree similar to the weak coupling regime [Fig. 7(a)]. Unlike in the weak coupling regime, the spotlight does not move at a well-defined velocity [Fig. 7(c)]. The dynamics of the spotlight seems superficially more similar to those trained in the strong coupling regime, with the spotlight larger when the average polarity points in the positive x direction [Fig. 7(d)].

The dilute system has a number density of 0.033 and an alignment coupling $k = 0.9$. All other physical constants are the same as those given in Appendix A. Agents were trained for 1.4×10^4 episodes before training was stopped and the policies were evaluated.

[1] G. Gompper, R. G. Winkler, T. Speck, A. Solon, C. Nardini, F. Peruani, H. Löwen, R. Golestanian, U. B. Kaupp, L. Alvarez *et al.*, The 2020 motile active matter roadmap, J. Phys.: Condens. Matter **32**, 193001 (2020).

[2] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Novel Type of Phase Transition in a System of Self-Driven Particles, Phys. Rev. Lett. **75**, 1226 (1995).
[3] M. F. Staddon, D. Bi, A. P. Tabatabai, V. Ajeti, M. P. Murrell, and S. Banerjee, Cooperation of dual modes of cell motility promotes epithelial stress relaxation to accelerate wound healing, PLoS Comput. Biol. **14**, e1006502 (2018).
[4] F. Cheng and J. E. Eriksson, Intermediate filaments and the regulation of cell motility during regeneration and wound healing, Cold Spring Harbor Perspect. Biol. **9**, a022046 (2017).
[5] S. J. Streichan, M. F. Lefebvre, N. Noll, E. F. Wieschaus, and B. I. Shraiman, Global morphogenetic flow is accurately predicted by the spatial distribution of myosin motors, eLife **7**, e27454 (2018).
[6] T. D. Ross, H. J. Lee, Z. Qu, R. A. Banks, R. Phillips, and M. Thomson, Controlling organization and forces in active matter through optically defined boundaries, Nature (London) **572**, 224 (2019).
[7] R. Zhang, S. A. Redford, P. V. Ruijgrok, N. Kumar, A. Mozaffari, S. Zemsky, A. R. Dinner, V. Vitelli, Z. Bryant, M. L. Gardel *et al.*, Spatiotemporal control of liquid crystal structure and dynamics through activity patterning, Nat. Mater. **20**, 875 (2021).
[8] G. Volpe, I. Buttinoni, D. Vogt, H.-J. Kümmerer, and C. Bechinger, Microswimmers in patterned environments, Soft Matter **7**, 8810 (2011).
[9] I. Buttinoni, G. Volpe, F. Kümmel, G. Volpe, and C. Bechinger, Active Brownian motion tunable by light, J. Phys.: Condens. Matter **24**, 284129 (2012).
[10] J. Palacci, S. Sacanna, A. P. Steinberg, D. J. Pine, and P. M. Chaikin, Living crystals of light-activated colloidal surfers, Science **339**, 936 (2013).
[11] G. Frangipane, D. Dell'Arciprete, S. Petracchini, C. Maggi, F. Saglimbeni, S. Bianchi, G. Vizsnyiczai, M. L. Bernardini, and R. Di Leonardo, Dynamic density shaping of photokinetic E. coli, eLife **7**, e36608 (2018).
[12] J. Stenhammar, R. Wittkowski, D. Marenduzzo, and M. E. Cates, Light-induced self-assembly of active rectification devices, Sci. Adv. **2**, e1501850 (2016).
[13] S. Das, M. Lee Bowers, C. Bakker, and A. Cacciuto, Active sculpting of colloidal crystals, J. Chem. Phys. **150**, 134505 (2019).
[14] A. Geiseler, P. Hänggi, F. Marchesoni, C. Mulhern, and S. Savel'ev, Chemotaxis of artificial microswimmers in active density waves, Phys. Rev. E **94**, 012613 (2016).
[15] A. Geiseler, P. Hänggi, and F. Marchesoni, Taxis of artificial swimmers in a spatio-temporally modulated activation medium, Entropy **19**, 97 (2017).
[16] J. Colen, M. Han, R. Zhang, S. A. Redford, L. M. Lemma, L. Morgan, P. V. Ruijgrok, R. Adkins, Z. Bryant, Z. Dogic *et al.*, Machine learning active-nematic hydrodynamics, Proc. Natl. Acad. Sci. USA **118**, (2021).
[17] S. Shankar and M. C. Marchetti, Hydrodynamics of Active Defects: From Order to Chaos to Defect Ordering, Phys. Rev. X **9**, 041047 (2019).
[18] M. M. Norton, P. Grover, M. F. Hagan, and S. Fraden, Optimal Control of Active Nematics, Phys. Rev. Lett. **125**, 178005 (2020).
[19] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, Machine learning for active matter, Nat. Mach. Intell. **2**, 94 (2020).
[20] D. F. Haeufle, T. Bäuerle, J. Steiner, L. Bremicker, S. Schmitt, and C. Bechinger, External control strategies for self-propelled particles: Optimizing navigational efficiency in the presence of limited resources, Phys. Rev. E **94**, 012617 (2016).
[21] T. Mano, J.-B. Delfau, J. Iwasawa, and M. Sano, Optimal run-and-tumble–based transportation of a Janus particle with active steering, Proc. Natl. Acad. Sci. USA **114**, E2580 (2017).
[22] Y. Yang, M. A. Bevan, and B. Li, Micro/nano motor navigation and localization via deep reinforcement learning, Adv. Theory Simul. **3**, 2000034 (2020).
[23] Y. Yang, M. A. Bevan, and B. Li, Efficient navigation of colloidal robots in an unknown environment via deep reinforcement learning, Adv. Intell. Syst. **2**, 1900106 (2020).
[24] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Smart inertial particles, Phys. Rev. Fluids **3**, 084301 (2018).
[25] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Flow Navigation by Smart Microswimmers Via Reinforcement Learning, Phys. Rev. Lett. **118**, 158004 (2017).
[26] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, Reinforcement learning with artificial microswimmers, Sci. Robot. **6**, eabd9285 (2021).
[27] S. Mishra, W. M. van Rees, and L. Mahadevan, Coordinated crawling via reinforcement learning, J. R. Soc. Interface **17**, 20200198 (2020).
[28] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, Learning to soar in turbulent environments, Proc. Natl. Acad. Sci. USA **113**, E4877 (2016).
[29] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, Glider soaring via reinforcement learning in the field, Nature (London) **562**, 236 (2018).
[30] G. Novati, L. Mahadevan, and P. Koumoutsakos, Controlled gliding and perching through deep-reinforcement-learning, Phys. Rev. Fluids **4**, 093902 (2019).
[31] M. Gazzola, A. A. Tchieu, D. Alexeev, A. de Brauer, and P. Koumoutsakos, Learning to school in the presence of hydrodynamic interactions, J. Fluid Mech. **789**, 726 (2016).
[32] M. Gazzola, B. Hejazialhosseini, and P. Koumoutsakos, Reinforcement learning and wavelet adapted vortex methods for simulations of self-propelled swimmers, SIAM J. Sci. Comput. **36**, B622 (2014).
[33] G. Novati, S. Verma, D. Alexeev, D. Rossinelli, W. M. Van Rees, and P. Koumoutsakos, Synchronisation through learning for two self-propelled swimmers, Bioinspir. Biomim. **12**, 036001 (2017).
[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, 2018).
[35] M. Durve, F. Peruani, and A. Celani, Learning to flock through reinforcement, Phys. Rev. E **102**, 012601 (2020).
[36] H. Chaté, F. Ginelli, G. Grégoire, F. Peruani, and F. Raynaud, Modeling collective motion: Variations on the Vicsek model, Eur. Phys. J. B **64**, 451 (2008).
[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.3.033291 for Movies 1–6.
[38] J. Toner, N. Guttenberg, and Y. Tu, Swarming in the Dirt: Ordered Flocks with Quenched Disorder, Phys. Rev. Lett. **121**, 248002 (2018).
[39] Y. Duan, B. Mahault, Y.-Q. Ma, X.-Q. Shi, and H. Chaté, Breakdown of Ergodicity and Self-Averaging in Polar Flocks with Quenched Disorder, Phys. Rev. Lett. **126**, 178001 (2021).

[40] M. E. Cates and J. Tailleur, Motility-induced phase separation, Annu. Rev. Condens. Matter Phys. **6**, 219 (2015).

[41] C. O. Reichhardt and C. Reichhardt, Ratchet effects in active matter systems, Annu. Rev. Condens. Matter Phys. **8**, 51 (2017).

[42] P. Lançon, G. Batrouni, L. Lobry, and N. Ostrowsky, Drift without flux: Brownian walker with a space-dependent diffusion coefficient, Europhys. Lett. **54**, 28 (2001).

[43] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, Deep reinforcement learning that matters, Proceedings of the Thirthy-Second AAAI Conference on Artificial Intelligence **32**, 1 (2018).

[44] A. Quillen, J. Smucker, and A. Peshkov, Boids in a loop: Self-propelled particles within a flexible boundary, Phys. Rev. E **101**, 052618 (2020).

[45] M. Paoluzzi, R. Di Leonardo, M. C. Marchetti, and L. Angelani, Shape and displacement fluctuations in soft vesicles filled by active particles, Sci. Rep. **6**, 34146 (2016).

[46] C. Abaurrea-Velasco, T. Auth, and G. Gompper, Vesicles with internal active filaments: Self-organized propulsion controls shape, motility, and dynamical response, New J. Phys. **21**, 123024 (2019).

[47] M. S. Peterson, A. Baskaran, and M. F. Hagan, Vesicle shape transformations driven by confined active filaments, arXiv:2102.02733.

[48] M. A. Karimi, V. Alizadehyazdi, B.-P. Busque, H. M. Jaeger, and M. Spenko, A Boundary-Constrained Swarm Robot with Granular Jamming, in *Proceedings of the 2020 3rd IEEE International Conference on Soft Robotics (RoboSoft)* (IEEE, Piscataway, NJ, 2020), pp. 291–296.

[49] K. Tanaka, M. A. Karimi, B.-P. Busque, D. Mulroy, Q. Zhou, R. Batra, A. Srivastava, H. M. Jaeger, and M. Spenko, Cable-driven jamming of a boundary constrained soft robot, in *Proceedings of the 3rd IEEE International Conference on Soft Robotics (RoboSoft)* (IEEE, Piscataway, NJ, 2020), pp. 852–857.

[50] J. W. Booth, D. Shah, J. C. Case, E. L. White, M. C. Yuen, O. Cyr-Choiniere, and R. Kramer-Bottiglio, OmniSkins: Robotic skins that turn inanimate objects into multifunctional robots, Sci. Robot. **3**, eaat1853 (2018).

[51] G. Wang, T. V. Phan, S. Li, M. Wombacher, J. Qu, Y. Peng, G. Chen, D. I. Goldman, S. A. Levin, R. H. Austin, and L. Liu, Emergent Field-Driven Robot Swarm States, Phys. Rev. Lett. **126**, 108002 (2021).

[52] J. A. Anderson, J. Glaser, and S. C. Glotzer, HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations, Comput. Mater. Sci. **173**, 109363 (2020).

[53] V. R. Konda, and J. N. Tsitsiklis, Actor-critic algorithms, in *Advances in Neural Information Processing Systems* (Citeseer, Princeton, NJ, 2000), p. 1008.

[54] A. Psai, MountainCar_ActorCritic, https://github.com/andy-psai/MountainCar_ActorCritic, 2019.

[55] M. Aldana, V. Dossetti, C. Huepe, V. Kenkre, and H. Larralde, Phase Transitions in Systems of Self-Propelled Agents and Related Network Models, Phys. Rev. Lett. **98**, 095702 (2007).