Learning to swim efficiently in a nonuniform flow field

Krongtum Sankaewtong,[1,*] John J. Molina,[1] Matthew S. Turner,[2,1] and Ryoichi Yamamoto[1,†]
[1]*Department of Chemical Engineering, Kyoto University, Kyoto 615-8510, Japan*
[2]*Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom*

(Received 28 December 2022; revised 1 May 2023; accepted 16 May 2023; published 7 June 2023)

Microswimmers can acquire information on the surrounding fluid by sensing mechanical queues. They can then navigate in response to these signals. We analyze this navigation by combining deep reinforcement learning with direct numerical simulations to resolve the hydrodynamics. We study how local and nonlocal information can be used to train a swimmer to achieve particular swimming tasks in a nonuniform flow field, in particular, a zigzag shear flow. The swimming tasks are (1) learning how to swim in the vorticity direction, (2) learning how to swim in the shear-gradient direction, and (3) learning how to swim in the shear-flow direction. We find that access to laboratory frame information on the swimmer's instantaneous orientation is all that is required in order to reach the optimal policy for tasks (1) and (2). However, information on both the translational and rotational velocities seems to be required to accomplish task (3). Inspired by biological microorganisms, we also consider the case where the swimmers sense local information, i.e., surface hydrodynamic forces, together with a signal direction. This might correspond to gravity or, for microorganisms with light sensors, a light source. In this case, we show that the swimmer can reach a comparable level of performance to that of a swimmer with access to laboratory frame variables. We also analyze the role of different swimming modes, i.e., pusher, puller, and neutral.

DOI: 10.1103/PhysRevE.107.065102

I. INTRODUCTION

Active matter encompasses a broad range of physical, chemical, and biological systems composed of "active" agents that consume energy from the surrounding environment in order to perform tasks (e.g., self-propel). Examples include motile cells such as spermatozoa, fish, birds, and even humans. Besides consuming energy, living agents also sense and react to environmental stimuli in order to accomplish their tasks, e.g., biological objectives such as gravitaxis [1,2], chemotaxis [3,4], or predation avoidance [5,6]. An example of predator avoidance is given by copepods (*Acartia tonsa*) [7], crustaceans that can be found in both freshwater and saltwater, which use mechanoreceptors to sense hydrodynamic signals in order to escape from predators. The recent progress in synthetic active particles has also revealed exciting possibilities for novel applications of active systems, e.g., as micromotors [8,9] or for therapeutics [10,11]. However, such applications require that the active agents be able to navigate complex environments in order to accomplish particular tasks. A natural question here is how to efficiently train the agents to achieve these objectives, when they are only able to process simple cues from their surroundings. For wet active systems, the major challenge is to fully account for the hydrodynamic interactions. Colabrese *et al.* [12] have used a reinforcement learning method (Q-learning) to develop efficient swimming strategies for a gyrotactic microswimmer, which was tasked with swimming in the vertical direction against a periodic Taylor-Green background vortex flow in two dimensions (2D). The swimmer was given information of its laboratory frame orientation and the vorticity of the background flow, which was discretized to have three possible values: positive, negative, or zero. Subsequent studies have extended the method to three dimensions, e.g., to optimize for vertical migration against gravity [13,14], avoiding predation [15], navigation near surfaces or interfaces [16], and navigation in other complex flow fields [17,18]. However, the hydrodynamic interactions were not fully taken into account in these studies, as the background flow was fixed, with the swimmer being advected or rotated by the flow. In contrast, here we investigate how to train a swimmer to navigate a complex flow (a zigzag shear flow) by performing direct numerical simulations (DNSs) [19] to account for the full hydrodynamic interactions and the particle-fluid coupling in three dimensions. First, we consider the case where the swimmer is able to perceive its current location, orientation, and translational and rotational velocities, within the laboratory frame, as well as retaining a memory of the last two actions it has performed. We then train the swimmer to achieve three separate tasks, (1) swimming in the vorticity direction, (2) swimming in the shear-gradient direction, and (3) swimming in the flow direction. We employ deep Q-learning [20] on a suitably discretized action space. Our results show the feasibility of using only the orientation and the action memory in order to learn optimal swimming strategies for tasks (1) and (2), i.e., swimming along the vorticity and the shear-gradient directions. Swimming in the flow direction proved to be a much more challenging task, as evidenced by the low performance

*aom@cheme.kyoto-u.ac.jp
†ryoichi@cheme.kyoto-u.ac.jp

compared with that of the other two. In this case, the swimmer was unable to learn to align itself with the flow streamline. While most studies on navigation [21–23] assume the agent's state to be composed of laboratory frame information, this is not appropriate for biological microswimmers, as they can only sense local information, e.g., hydrodynamic signals. Therefore we have also investigated how the same learning can be performed using only locally accessible information, i.e., the hydrodynamic force exerted on the swimmer by the surrounding fluid and the relative alignment of the swimmer with a signal direction. In what follows we will refer to this as a light source, sensed by light-sensitive receptors for which we use the shorthand of "eye," without implying the presence of a fully developed eye. In principle, microorganisms could also sense the Earth's gravitational field, or other signals coming from, e.g., persistent magnetic [24], heat [25], or chemical gradients. For the case in which the organism can sense a single laboratory frame signal direction in this way we found that a combination of these two signals (hydrodynamic forces and signal orientation), along with the memory of two recent actions, can yield the same qualitative level of performance as when using laboratory frame information. Finally, we also investigate the effect of learning for different swimming modes, i.e., pusher, puller, and neutral. When given the same set of signals, pushers show the best performance, above that of neutral swimmers, with pullers performing the worst.

II. SIMULATION METHODS

A. System of interest

We consider a swimmer navigating through a Newtonian fluid with an imposed zigzag shear flow. The coupled dynamics of the swimmer and the fluid are evaluated by solving a modified Navier-Stokes equation, which accounts for the fluid-particle interaction using the smoothed profile (SP) method [19], together with the Newton-Euler equations for the rigid-body dynamics. The squirmer is assumed to be able to perceive information from its environment and perform actions accordingly. These actions are determined by a weighted neural network, trained using a deep Q-learning algorithm, to draw actions that lead to the highest accumulated reward over a given time interval.

B. The squirmer model

Here, we consider the "squirmer" model to represent swimmers as self-propelled spherical particles with a modified stick-boundary condition [26,27]. Originally, this model was proposed to describe the dynamics of ciliated microorganisms, where the swimmers are driven by the fluid flows generated at their surfaces. The expression for this surface velocity is given as an expansion in terms of Legendre polynomials. For simplicity, the radial and azimuthal components are usually neglected, and the expansion is usually truncated to neglect modes higher than second order [28]. Thus the slip velocity of the swimmer at a given point on its surface is characterized by spherical polar variables (ϑ, φ), with $\vartheta = 0$ corresponding to the swimming direction, according to

$$\mathbf{u}^s(\vartheta, \varphi) = B_1\left(\sin\vartheta + \frac{\alpha}{2}\sin 2\vartheta\right)\hat{\boldsymbol{\vartheta}}, \tag{1}$$

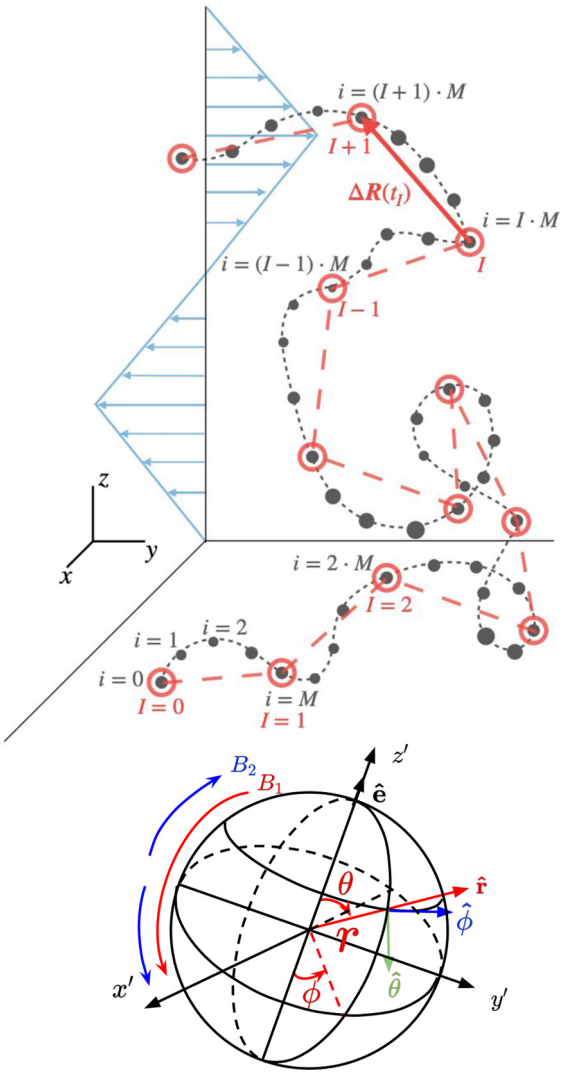

FIG. 1. Top: A schematic illustration of a particle learning to navigate a zigzag shear flow. The particle trajectory over a learning episode is illustrated as the black short-dashed line. The position of the swimmer at the discrete simulation time steps i is marked with black filled circles. This trajectory is *coarse grained* to define the action trajectory, illustrated as the red long-dashed line. This action trajectory consists of action segments, composed by taking every M simulation steps, marked with red open circles. The swimmer chooses an action a_I at the start of each action segment I, which it follows until the start of the next action segment $(I+1)$. Bottom: a schematic diagram of the squirmer "pusher" model, illustrating the surface flows generated by the B_1 ($\propto \sin\theta$) and B_2 ($\propto \sin 2\theta$) modes, which determine the swimming speed and stresslet, respectively; as well as the relevant unit vectors and angles, in spherical coordinates, i.e., $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$, where $\hat{\boldsymbol{e}}$ is the swimming direction.

where $\hat{\boldsymbol{\vartheta}}$ is the tangential unit vector in the ϑ direction, $\vartheta = \cos(\hat{\boldsymbol{r}}\cdot\hat{\boldsymbol{e}})$ is the polar angle, with $\hat{\boldsymbol{e}}$ is the swimming axis, and $\hat{\boldsymbol{r}}$ is a unit vector pointing from the center of the squirmer to the point (ϑ, φ) on the surface, as shown in the bottom panel of Fig. 1. The coefficient B_1 is the amplitude of the first squirming mode, which determines the steady-state swimming velocity of the squirmer $U = \frac{2}{3}B_1$, with $\alpha = B_2/B_1$ characterizing the type of flow field: For negative (positive)

α, the squirmer is a pusher (puller), e.g., *Escherichia coli* (*Chlamydomonas reinhardtii*). The first mode, B_1, relates to the hydrodynamic source dipole, with a decay in the velocity field proportional to $1/r^3$, while that of the second mode is related to a force dipole, which decays as $1/r^2$. For a neutral squirmer (e.g., *Paramecium*), $\alpha = 0$, the first mode dominates over B_2, and the velocity field decays as $1/r^3$.

C. The smoothed profile method

To solve for the coupled fluid and particle dynamics, we solve the equations of motion for both the viscous host fluid and the squirmer using the smoothed profile (SP) method [19]. The evolution of the particle obeys the Newton-Euler equations:

$$\dot{\boldsymbol{R}}_i = \boldsymbol{V}_i, \quad \dot{\boldsymbol{Q}}_i = \mathrm{skew}(\boldsymbol{\Omega}_i)\cdot\boldsymbol{Q}_i,$$
$$M_p\dot{\boldsymbol{V}}_i = \boldsymbol{F}_i^H + \boldsymbol{F}_i^{\mathrm{ext}}, \quad \boldsymbol{I}_p\cdot\dot{\boldsymbol{\Omega}}_i = \boldsymbol{N}_i^H + \boldsymbol{N}_i^{\mathrm{ext}}, \tag{2}$$

where i is the particle index, $\boldsymbol{R}_i$ and $\boldsymbol{V}_i$ are the particle center-of-mass position and velocity, respectively, $\boldsymbol{Q}_i$ is the orientation matrix, and $\boldsymbol{\Omega}_i$ is the angular velocity. The skew-symmetric matrix is defined such that $\mathrm{skew}(\boldsymbol{\Omega}_i)\cdot\boldsymbol{x} = \boldsymbol{\Omega}_i\times\boldsymbol{x}$ ($\forall\boldsymbol{x}\in\mathbb{R}^3$),

$$\mathrm{skew}(\boldsymbol{\Omega}_i) = \begin{pmatrix} 0 & -\Omega_i^z & \Omega_i^y \\ \Omega_i^z & 0 & -\Omega_i^x \\ -\Omega_i^y & \Omega_i^x & 0 \end{pmatrix}. \tag{3}$$

The forces exerted on the particle, appearing on the right-hand side of (2), include the hydrodynamic forces $\boldsymbol{F}^H$ and external forces $\boldsymbol{F}^{\mathrm{ext}}$, e.g., gravity. Likewise, the torques are decomposed into hydrodynamic $\boldsymbol{N}^H$ and external $\boldsymbol{N}^{\mathrm{ext}}$ contributions. Here, we have neglected interparticle forces and torques, as we only consider single-particle systems. The forces and torques are evaluated assuming momentum conservation to ensure a consistent coupling between the host fluid and the particles. The time evolution of the host fluid is determined by the Navier-Strokes equation, together with the incompressibility condition:

$$\nabla\cdot\boldsymbol{u}_f = 0, \tag{4}$$

$$\rho_f(\partial_t + \boldsymbol{u}_f\cdot\nabla)\boldsymbol{u}_f = \nabla\cdot\boldsymbol{\sigma}_f + \rho_f\boldsymbol{f}, \tag{5}$$

$$\boldsymbol{\sigma}_f = -p\boldsymbol{I} + \eta_f[\nabla\boldsymbol{u}_f + (\nabla\boldsymbol{u}_f)^T], \tag{6}$$

where ρ_f is the fluid mass density, $\boldsymbol{u}_f$ is the fluid velocity field, η_f is the shear viscosity, $\boldsymbol{\sigma}_f$ is the stress tensor, and $\boldsymbol{f}$ is an external body force.

When applying the SP method, the sharp interface between the rigid particle and the fluid domains is replaced by an interfacial region with a finite width ξ, with both regions characterized by a smooth and continuous function ϕ. This function returns a value of 0 for the fluid domain and a value of 1 for the solid domain. The total velocity field $\boldsymbol{u}$ can then be written as

$$\boldsymbol{u} = (1-\phi)\boldsymbol{u}_f + \phi\boldsymbol{u}_p. \tag{7}$$

The first term on the right-hand side of (7) represents the contribution from the host fluid, while the second term is from the rigid-body motion. Then, we can consider the system as a single-component fluid and write down a modified Navier-Stokes equation, similar to (5), but in terms of the total velocity $\boldsymbol{u}$, as

$$\rho_f(\partial_t + \boldsymbol{u}\cdot\nabla)\boldsymbol{u} = \nabla\cdot\boldsymbol{\sigma}_f + \rho_f(\phi\boldsymbol{f}_p + \phi\boldsymbol{f}_{sq} + \boldsymbol{f}_{\mathrm{shear}}), \tag{8}$$

where $\phi\boldsymbol{f}_p$ is the force density field required to maintain the rigidity of the particle, $\phi\boldsymbol{f}_{sq}$ is the force density required to maintain the squirming motion, and $\boldsymbol{f}_{\mathrm{shear}}(\boldsymbol{x},t)$ is an external force required to maintain the following zigzag velocity profile [29]:

$$v_x(y) = \begin{cases} \dot{\gamma}(-y - L_y/2), & -L_y/2 < y \leqslant -L_y/4 \\ \dot{\gamma}y, & -L_y/4 < y \leqslant L_y/4 \\ \dot{\gamma}(-y + L_y/2), & L_y/4 < y \leqslant L_y/2, \end{cases} \tag{9}$$

where $\dot{\gamma}$ is the shear rate, y is the distance in the velocity-gradient direction, and L_y is the height of the three-dimensional rectangular simulation box, of dimensions (L_x, L_y, L_z). We numerically solve the equations of motion using a fractional step procedure. First, the total velocity field is updated by solving for the advection and hydrodynamics stress contributions in the Navier-Stokes equation. Simultaneously, the particle positions and orientations are propagated forward in time. Second, we evaluate the momentum exchange over the particle domain and use it to compute the hydrodynamic contributions to the forces (torques) exerted on the particles. Third, the updated forces and torques are used to update the particle velocities, in such a way that the squirming boundary condition is maintained (through $\phi\boldsymbol{f}_{sq}$). Finally, the rigidity constraint ($\phi\boldsymbol{f}_p$) is computed, in such a way that the momentum conservation is guaranteed, and used to update the total velocity field (together with the shear-flow constraint $\boldsymbol{f}_{\mathrm{shear}}$). Detailed discussions of this procedure can be found in our earlier work [19,29,30].

D. Deep reinforcement learning

We employ a reinforcement learning (RL) framework [31] to obtain optimal policies for the prescribed swimming tasks. This involves training a neural network to select actions that generate a high reward. RL has proven itself to be a powerful tool for finding flow control and navigation strategies [12,32,33]. In RL, an agent (here the swimming particle), uses information received from its environment to define its current *state*, which it uses to determine its next *action*, resulting in a corresponding *reward* (assumed to be a real number) for this action. This type of agent-environment interaction allows one to control the agent decisions, in order to maximize the long-run accumulated reward without prior knowledge of the dynamics of the system. In this paper, we adopt a deep Q-learning strategy, combined with prioritized experience replay and n-step learning [20,34,35], as the search tool for finding optimal navigation strategies for a given task.

The swimmer is trained by maximizing the expected reward over a fixed time interval, called an episode (see Fig. 1). Episodes are discretized into N_s action segments of M simulation steps each, such that $T_{\mathrm{episode}} = N_s\cdot\Delta T$, with $\Delta T = M\cdot\Delta t$ being the time duration of the action segment (Δt is the simulation time step). Let s_I be the state of the

swimmer at the beginning of action segment I, which corresponds to simulation time step $i = I \cdot M$ and time $T_I = I \cdot \Delta T = (I \cdot M) \cdot \Delta t \equiv t_{i=I\cdot M}$. The swimmer uses a policy function π, which maps states to actions, in order to choose its next action a_I, which it will follow for the duration of the action segment (i.e., the next M simulation steps), after which it will be in a new state $s'_I = s_{I+1}$. The swimmer is then assigned a reward r_I, which depends on its state at the endpoints, s_I and s'_I. This gained (action-reward) *experience* is written in tuple form as (s_I, a_I, s'_I, r_I). For the purposes of the learning, the state of the system at intermediate times between the starts of subsequent action steps I and $I+1$, i.e., $T_I = t_{I\cdot M} < t < t_{(I+1)\cdot M} = T_{I+1}$, will be irrelevant. Finally, at the end of the episode, the total reward is evaluated by accumulating each of the individual action rewards acquired during the trajectory, $r = \sum_{I=0}^{N_s-1} r_I$.

To compute the (optimal) policy, define the action-value function Q_π, for a given policy π, as $Q_\pi(s_I, a_I) = r_I + \gamma r_{I+1} + \gamma^2 r_{I+2} + \cdots$, with $\gamma \in [0, 1)$ being a discount factor. This Q function gives the expected accumulated reward for adopting action a_I during step I (starting from state s_I), expressed as the reward for action step I, plus the (discounted) rewards at each subsequent step ($0 < I < N_s$). The optimal policy function π^*, whose mapping of states and actions maximizes the long-time reward, must satisfy the Bellman equation $Q_{\pi^\star}(s_I, a_I) = r_I + \gamma \max_a Q_{\pi^\star}(s_{I+1}, a)$ [31]. To learn this optimal policy, the Q value function is represented by a neural network and trained in an episode-based fashion, over N_{episode} episodes, with each episode consisting of N_s action steps, of M simulation steps each.

Throughout the training phase, at the beginning of each episode the position and orientation of the swimmer are randomly set, and the swimmer is allowed to navigate for the time duration T_{episode} (N_s action segments). At each action step I, the (current) best action is that which maximizes the (current) value function, i.e., $a_I = \text{argmax}_a Q_\pi(s_I, a)$. In order to reduce the bias in the training, a batch of stored experiences of size N_b is drawn from a "replay memory" buffer (size $N_{P_{\max}}$) at the end of each action step. The replay memory buffer will store the experiences gathered over many action steps and episodes. Each element of this batch of experiences consists of information on the environmental signals, the actions taken, and the immediate rewards, i.e., (s_I, a_I, s'_I, r_I). The drawn batch is then used to adjust the weights of Q, according to the following rule:

$$Q(s_I, a_I) \leftarrow Q(s_I, a_I) + \nu[r_I + \gamma \max_a Q(s_{I+1}, a) - Q(s_I, a_I)], \tag{10}$$

where ν is the learning rate. The Q network is trained against the following loss function (with $\boldsymbol{\theta}$ being the network weights):

$$L_I(s_I, a_I; \boldsymbol{\theta}_I) = (Y_I - Q(s_I, a_I; \boldsymbol{\theta}_I))^2, \tag{11}$$

where Y_I is the "target" at learning step I, defined as

$$Y_I = r_I + \gamma \max_a Q(s_{I+1}, a; \boldsymbol{\theta}_I^-) \tag{12}$$

with $\boldsymbol{\theta}_I^-$ being a set of target network parameters which are synchronized with the "prediction" Q-network parameters (the ones begin optimized for) $\boldsymbol{\theta}_I^- = \boldsymbol{\theta}_I$ every C steps, and otherwise held fixed between individual $\boldsymbol{\theta}$ updates. This can be understood as a loss function that depends on two identical Q networks, a prediction network and a target network, but is only trained on the former. The gradient with respect to the weights $\boldsymbol{\theta}$ can then be written as (writing only the $\boldsymbol{\theta}$ dependence)

$$\nabla_{\boldsymbol{\theta}_I} L(\boldsymbol{\theta}_I) = [Y_I - Q(\boldsymbol{\theta}_I)] \nabla_{\boldsymbol{\theta}_I} Q(\boldsymbol{\theta}_I). \tag{13}$$

We use the ADAM optimizer [36] to minimize the loss function. A detailed description of the deep Q-learning framework we have used can be found in Ref. [20].

In order to maximize exploration of the phase space, particularly at early stages of the learning, we have adopted an ϵ-greedy selection scheme. Thus the chosen policy a_I is allowed to deviate from the optimal policy. That is, the optimal policy determined from the action-value function Q is used with probability $1 - \epsilon$; otherwise the action is randomly drawn from the action space with probability ϵ. This greedy parameter is exponentially decaying in time, starting from $\epsilon = 1$, until it reaches a value of $\epsilon = 0.015$. The reason for decaying the greedy parameter is to prevent the swimmer from prematurely narrowing down the state-action space. During the early episodes of the training, the policy is very far from its optimum; therefore the swimmer needs to explore the state-action space as much as possible to gain experience interacting with the environment. Thus it is better to draw random actions, rather than sticking to a specific policy. However, after a suitable training period, the policy is expected to improve, and the swimmer should now favor the trained policy, rather than a randomly chosen action. Note that we always allow for a small probability ϵ of selecting a random action, even though the policy is expected to converge to the optimal one, since we aim to provide some room for possible improvements of the current best policy.

Finally, we will consider two forms of reward, signed rewards, in which the reward for segment I is computed from the displacement of the swimmer $\Delta\boldsymbol{R}(T_I) = \boldsymbol{R}(T_{I+1}) - \boldsymbol{R}(T_I)$, and unsigned rewards, computed from the absolute value of the displacements. Furthermore, since we consider the three distinct tasks of swimming in the shear-flow (x), shear-gradient (y), and vorticity (z) directions, the rewards are given by $r_I = \hat{\mathrm{e}}^\mu \cdot \Delta\boldsymbol{R} = \Delta R^\mu(T_I)$, in the signed case, and $r_I = |\Delta R^\mu(T_I)|$, in the unsigned case ($\hat{\mathrm{e}}^\mu$ are the unit basis vectors in the laboratory frame, $\mu = x, y, z$). The latter type of reward may be a natural choice for organisms in which there is no preferential direction of motion, but where moving to a new location may be advantageous. An example of this could be the swimming of the marine bacterium *Vibrio alginolyticus*, whose swimming pattern is a cycle of forward and backward swimming, together with turns to change its direction of motion [37]. Furthermore, we discretize the action space and define an action to be an external torque $\boldsymbol{N}^{\text{ext}} = H\hat{\boldsymbol{n}}$ that the swimmer can activate, with H being the magnitude of the torque and $\hat{\boldsymbol{n}} = \boldsymbol{m}/|\boldsymbol{m}|$ being a unit vector ($m^\mu = -1, 0, 1$). Thus the size of the action space is given by the $3^3 = 27$ possible rotation axes.

E. The system parameters

Throughout this paper we present our results in simulation units, using as basic units of length, density, and viscosity

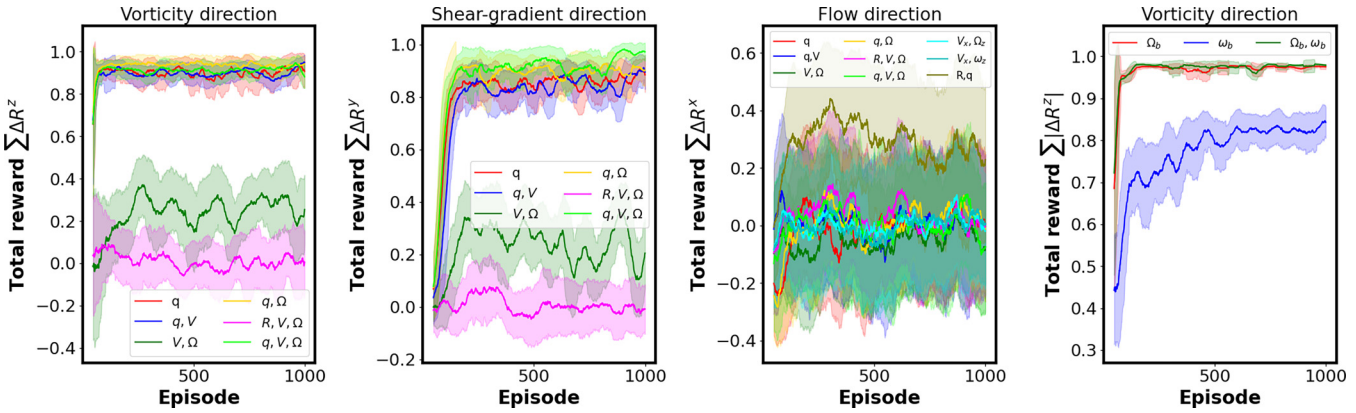

FIG. 2. Learning performance for different tasks or rewards as measured by the rolling average of the total reward, normalized by the maximum possible displacements per episode. From left to right: Learning to swim in the vorticity (z), shear-gradient (y), and shear-flow (x) directions, using laboratory frame information and signed rewards, and finally, learning to swim in the vorticity direction using body frame information and unsigned rewards. For each task, we train our swimmer using different sets of signals, which can include the orientation quaternion q, translational velocity $\boldsymbol{V}$, rotational velocity $\boldsymbol{\Omega}$, position $\boldsymbol{R}$, and background flow vorticity $\boldsymbol{\omega}$, as specified in the plot legends. The two previous actions ($\mathbf{a}$) were included in all cases (not labeled). The subscript b in the legend of the last panel denotes signal variables projected into the body frame. An averaging window size of 50 episodes was used. The shaded area represents the standard error in the mean. The signed rewards are defined as $\sum_{I=0}^{N_s-1} \Delta R^{\mu}(T_I) = \sum_{I=0}^{N_s-1} (R^{\mu}(T_{I+1}) - R^{\mu}(T_I))$, where I is the action or learning step and N_s is the total number of action segments per episode. Likewise, the unsigned rewards are given as $\sum_{I=0}^{N_s-1} |\Delta R^z(T_I)| = \sum_{I=0}^{N_s-1} |R^z(T_{I+1}) - R^z(T_I)|$.

the grid spacing $\Delta = 1$, fluid density $\rho_f = 1$, and viscosity $\eta = 1$. The units of time and mass are $\rho_f(\Delta)^2/\eta = 1$ and $\rho_f \Delta^3 = 1$, respectively. The radius of the spherical swimmer is $\sigma = 5\Delta$, and the size of our rectangular simulation box is $32\Delta \times 64\Delta \times 32\Delta$, with full periodic boundary conditions along all dimensions. Other parameters used in the SP simulator are the particle-fluid interface thickness $\xi = 2\Delta$, the particle density $\rho_p = \rho_f$, and the magnitude of the external torque $H = 400\eta^2\Delta/\rho_f$. The applied shear rate is $\dot{\gamma} = 0.04\eta/(\rho_f \Delta^2)$, which corresponds to a Reynolds number $\mathrm{Re} \approx 1$. For most of the cases presented here, and unless stated otherwise, the swimmer is set to be a puller with $\alpha = 2$ and $B_1 = 0.1\eta/(\rho_f \Delta)$, corresponding to a particle Reynolds number $\mathrm{Re} \approx 6 \times 10^{-2}$, comparable to that of *E. coli* in water [38].

To further characterize our system, we introduce the following three dimensionless ψ parameters:

$$\psi_1 = \frac{\frac{2}{3}B_1}{\frac{\dot{\gamma}}{2}L_y}, \tag{14}$$

$$\psi_2 = \frac{H/(\pi\sigma^3\eta)}{\dot{\gamma}/2}, \tag{15}$$

$$\psi_3 = \frac{\frac{2}{3}B_1 T_{\mathrm{episode}}}{L_y}. \tag{16}$$

These measure the strength of the swimming in three natural ways: ψ_1 is the ratio of the baseline swimmer speed to the maximum shear-flow speed (twice the typical shear-flow speed); ψ_2 is the ratio of the active rotation rate of the swimmer to that induced by the shear flow; and ψ_3 is the ratio of the maximum total active displacement of a swimmer over one episode, duration T_{episode}, to the largest system size L_y. Unless otherwise specified, our system parameters correspond to the following: $\psi_1 \simeq 5 \times 10^{-2}$, meaning that the swimmer moves slowly compared with the fluid speed; $\psi_2 \simeq 6$, meaning that the swimmer can actively rotate faster than the rotation induced by the shear flow, necessary for it to have meaningful control of its orientation; and $\psi_3 \simeq 20$, meaning the swimmer can explore regions with different shear gradients. The characteristic angular rotation (per epoch) caused by the external torque corresponds to $\frac{H}{\pi\sigma^3\eta}T_{\mathrm{episode}} \simeq 180$ rad, such that the agent can fully rotate $\gtrsim$10 times during one episode. For the learning parameters, we use a discount rate $\gamma = 0.93$, learning rate $\nu = 2.5 \times 10^{-4}$, and batch size $N_b = 128$, with a replay memory size of $N_{P_{\max}} = 10^5$ and a greedy parameter ϵ with decay rate $k = 0.992$ (see Supplemental Material Sec. IV [39] for a discussion of the choice of decay rate). The neural network consists of one input layer with the number of neurons equal to the number of state-defining variables, with three hidden layers of 100 neurons each, and one output layer with 27 neurons, corresponding to the size of the action space. Finally, a learning episode consists of $N_s = 2 \times 10^3$ action steps of $M = 10$ simulation steps each. The precise numerical values for all our parameters can be found in the Supplemental Material, Sec. I [39].

III. RESULTS AND DISCUSSION

One of the main challenges of RL is defining an appropriate state. Gunnarson *et al.* [40] have shown that in unsteady two-dimensional flow fields, different sets of environmental cues lead to significantly different levels of performance for a given task. Here, to define the state, we use combinations of the swimmer's laboratory frame configuration, i.e., position $\boldsymbol{R}$ ($n_d^R = 3$), translational velocity $\boldsymbol{V}$ ($n_d^V = 3$), rotational velocity $\boldsymbol{\Omega}$ ($n_d^{\Omega} = 3$), and rotation quaternion q ($n_d^q = 4$), together with the background flow information, in particular, the flow vorticity $\boldsymbol{\omega} = \boldsymbol{\nabla} \times \boldsymbol{u}$ ($n_d^{\omega} = 3$). For simplicity, we encode the orientation using quaternions $q = (\cos\theta/2, \sin\theta/2\boldsymbol{n})$, defined by rotating the laboratory frame around a unit vector $\boldsymbol{n}$, by an angle θ. The number of degrees of freedom (inputs)

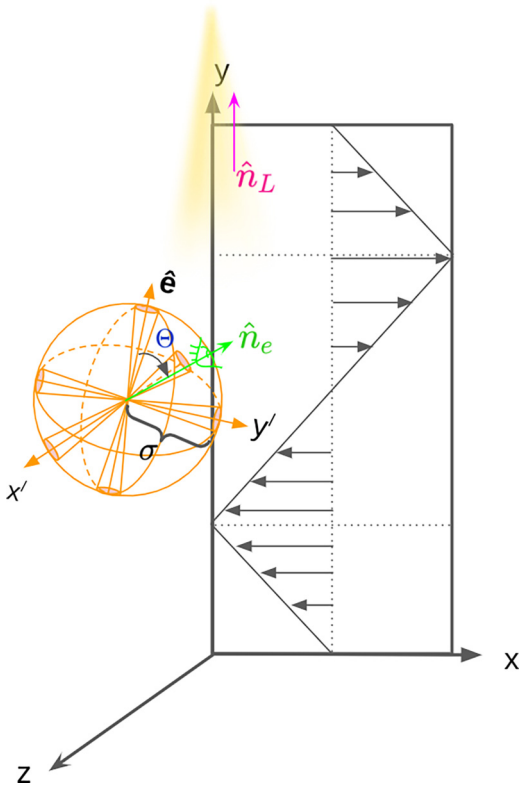

FIG. 3. Schematic representation of a spherical microswimmer that measures (local) body frame information while navigating a zigzag shear flow. Here, x', y', and z' are the principal axes in the body reference frame, with $\hat{\boldsymbol{e}} = z'$ being the swimming direction. We consider a swimmer with six sensors distributed across its surface, each one aligned with one of the principal axes. The stress sensors occupy spherical caps (highlighted) with an area of $\pi/2\sigma^2(1-\cos 30) \approx 0.21\sigma^2$ each. The swimmer is assumed to have a single sensor (eye) that can sense a signal (light), with a location specified by the unit vector $\hat{\boldsymbol{n}}_e$ (green arrow). The signal source can lie in any direction $\hat{\boldsymbol{n}}_L$ (magenta arrow), but here we illustrate the case in which it is aligned with the shear-gradient direction (magenta arrow).

required to specify each of the state variables Γ is given by n_d^Γ ($\Gamma \in \{R, V, \Omega, \ldots\}$). Taken together, these sets of signals provide a complete specification of the swimmer's current configuration. We also include the swimmer's previous two actions, denoted **a** in all examples, with the goal of improving the convergence of the policy. Such limited memory may be accessible, even in microorganisms [41]. However, it turns out that this memory term is not essential for the learning (Supplemental Material Sec. V [39]). Figure 2 shows the rolling average of the normalized total rewards for the different swimming tasks. The normalization constant is defined as the maximum possible displacements per episode, calculated as $(2/3B_1)T_{\text{episode}}$.

For the tasks of swimming in the vorticity and shear-gradient directions, it is sufficient for the swimmer to receive orientation information, via the rotation quaternion q, along with the memory of the last two actions taken, in order to develop an efficient policy. We see that the signal variable combinations that include the orientation q can approach an optimal policy, while those that do not show inferior performance. In contrast, for the task of swimming in the flow direction, the swimmer was unable to reliably perform the task. In these cases, however, swimmers given position and orientation information were able to outperform all other swimmers (as might be expected), even if the overall score was still poor (relative to the other tasks). As shown in Fig. 2 (flow direction), swimmers with this privileged information achieve a reward that is significantly higher ($\simeq 0.4$) than that achieved by those without the information ($\simeq 0$) and can actually learn to swim in the flow direction even though the performance is still inferior compared with the other two tasks. A detailed quantitative analysis of the performance obtained for this task can be found in the Supplemental Material, Sec. II [39]. This (relatively) poor performance is associated with the fact that the swimmer is unable to learn how to align itself with the x-y (shear) plane; otherwise it would be able to locate the position (height) of maximum flow and remain there in order to maximize its reward. The last panel of Fig. 2 shows the results obtained when using unsigned rewards, defined as $r_I = |\Delta R^z(T_I)| = |R^z(T_{I+1}) - R^z(T_I)|$, for the task of swimming in the vorticity (z) direction. The signals used for this set of simulations are different (compared with learning with signed rewards), as the swimmer only needs to perceive how it is orientated relative to the flow vorticity $\boldsymbol{\omega}$, as measured by the angular velocity in the body frame $\boldsymbol{\Omega}_b$, in order to develop an efficient policy. While the signal from the background vorticity also provides information about the orientation of the swimmer, the performance in this case is not as good as that obtained when using the swimmer's rotational velocity.

We have demonstrated the ability of idealized microswimmers to efficiently perform swimming tasks using laboratory frame information. We have also shown that information on the body frame rotational velocity can be used to efficiently swim in the (unsigned) vorticity direction, since it can give a hint as to the axis of rotation, which itself can be related to the alignment in the vorticity direction. However, active microorganisms in nature may not have such privileged information. Typically, they can only obtain certain body frame signals. A good example is that of copepods, which are able to sense the proximity of predators through the induced bending patterns of their setae, hairlike structures on the surface [7]. Recent work on RL for swimming in nonuniform flows has included local signals, e.g., local fluid strain rate and slip velocity, but the full hydrodynamic effects have not been taken into account [14,42,43]. Furthermore, these studies have only considered the task of swimming against gravity. Given that swimmers can be expected to sense local surface stress signals, we train a swimmer to accomplish a similar suite of swimming tasks to those discussed before (i.e., swimming in the shear-flow, shear-gradient, and vorticity directions), but using a set of physiologically reasonable body frame signals. First, the swimmer is assumed to be able to detect surface stresses, through the hydrodynamic forces exerted by the surrounding fluid on the swimmer, denoted as $\boldsymbol{\tau}_i$ ($n_{st} = 3$). Here, we assume that the spherical swimmer has six surface sensors ($0 \leqslant i < 6$), located on the antipodal points at the intersection of the three principal body axes (see Fig. 3). Each sensor is assigned a surface area of $0.21\sigma^2$, where σ is the particle diameter. These sensors are given a finite size in

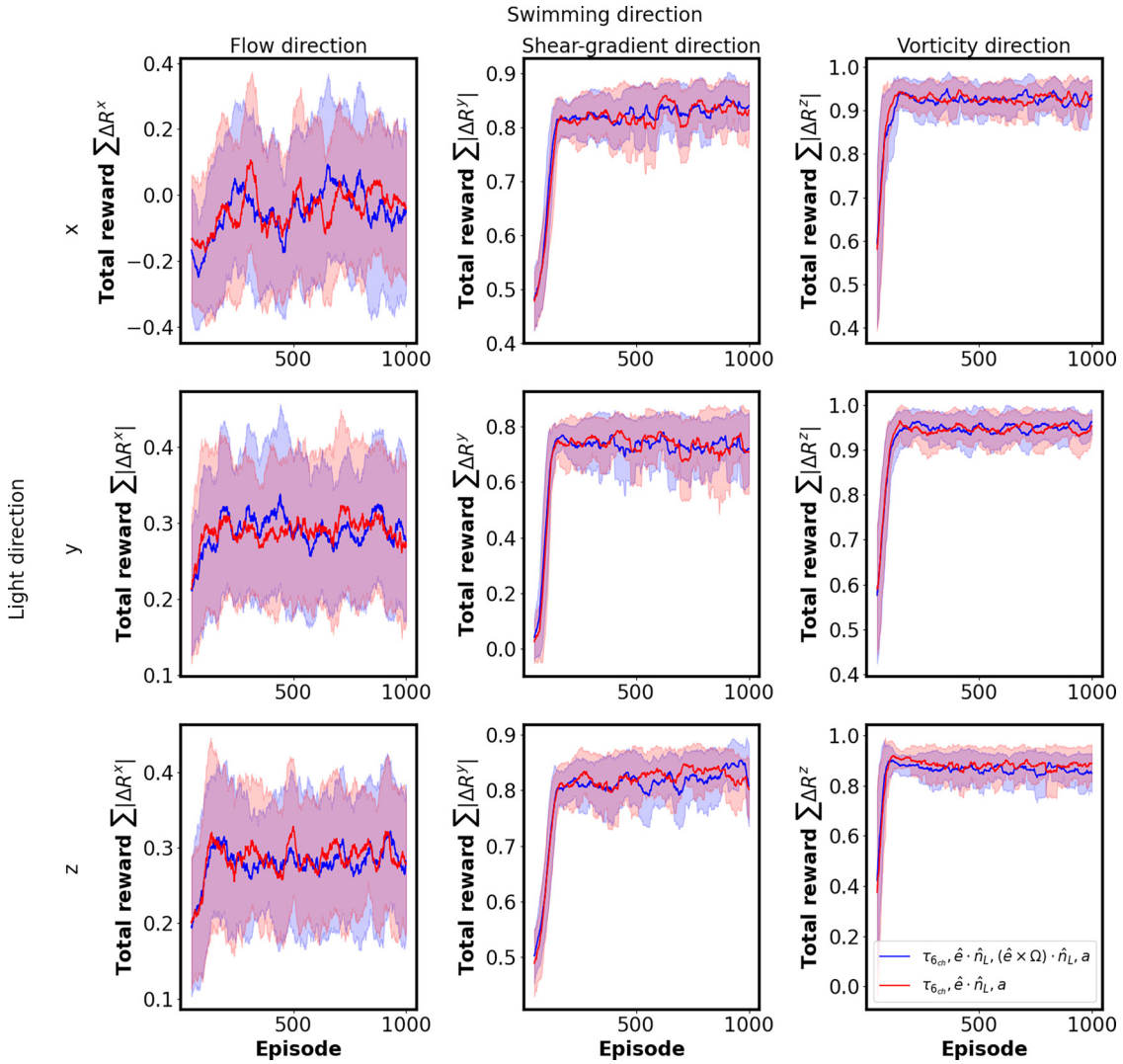

FIG. 4. The rolling average of the normalized total rewards, with an averaging window of 50 episodes, for swimmers tasked with migrating along different directions μ, under light sources shining along a direction β ($\mu, \beta = x, y, z$). Columns represents the reward direction μ, and rows represent the signal direction β. The total rewards are defined in terms of either signed displacements $\sum_I R^\mu(T_I)$ ($\mu = \beta$) or unsigned displacements $\sum_I |\Delta R^\mu(T_I)|$ ($\mu \neq \beta$). Here we consider two sets of input signals, $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L, (\hat{\boldsymbol{n}}_e \times \boldsymbol{\Omega}) \cdot \hat{\boldsymbol{n}}_L\}$ and $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$, with the eye's location at $\Theta = 30^\circ$.

order to average the surface stress derived from the numerical simulations. Furthermore, we also assume that this swimmer has a sensor (eye) located on its surface, at an angle Θ from the swimming direction $\hat{\mathbf{e}}$. This sensor can be considered to detect visual cues (i.e., light), via the signal $\hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L$ ($n_{st} = 1$). This might, therefore, serve as a model for a microorganism capable of migrating towards or away from light sources. We consider cases in which the light can come from one of three directions: parallel to the flow direction (x), parallel to the shear-gradient direction (y), or parallel to the vorticity direction (z). In addition, we also consider the case in which the model microorganism has the ability to sense the "flashing" of the light due to its relative reorientation. This flashing signal is encoded as $(\hat{\boldsymbol{n}}_e \times \boldsymbol{\Omega}) \cdot \hat{\boldsymbol{n}}_L$. We utilize these as state-defining parameters and repeat the same Q-learning procedure used previously, in order to obtain the optimal policy for each of the three swimming tasks, i.e., swimming in the shear-flow, shear-gradient, and vorticity directions.

Figure 4 shows the learning results for such a model swimmer. Each column of panels in this figure represents a reward direction, i.e., the desired swimming direction, and each row represents the direction of the signal source. In cases where the desired or target swimming direction is aligned with the signal direction, we use a signed reward $\sum_I \Delta R^\mu(T_I)$; otherwise we use the corresponding unsigned reward $\sum_I |R^\mu(T_I)|$. The rationale behind this is that the swimmer is able to distinguish whether it swims towards or away from a light source, hence the use of signed rewards in that case. However, while there may be an evolutionary advantage for the swimmer to move from its present location (e.g., to locate additional food sources), in the absence of any external signals, there is no reason to choose any particular (signed) direction. Inspired by the fact that the eye location of particular biological microorganisms, e.g., *Chlamydomonas*, is at 30° away from the front [44], we consider a microorganism-inspired swimmer with $\Theta = 30^\circ$. This swimmer can perceive three types

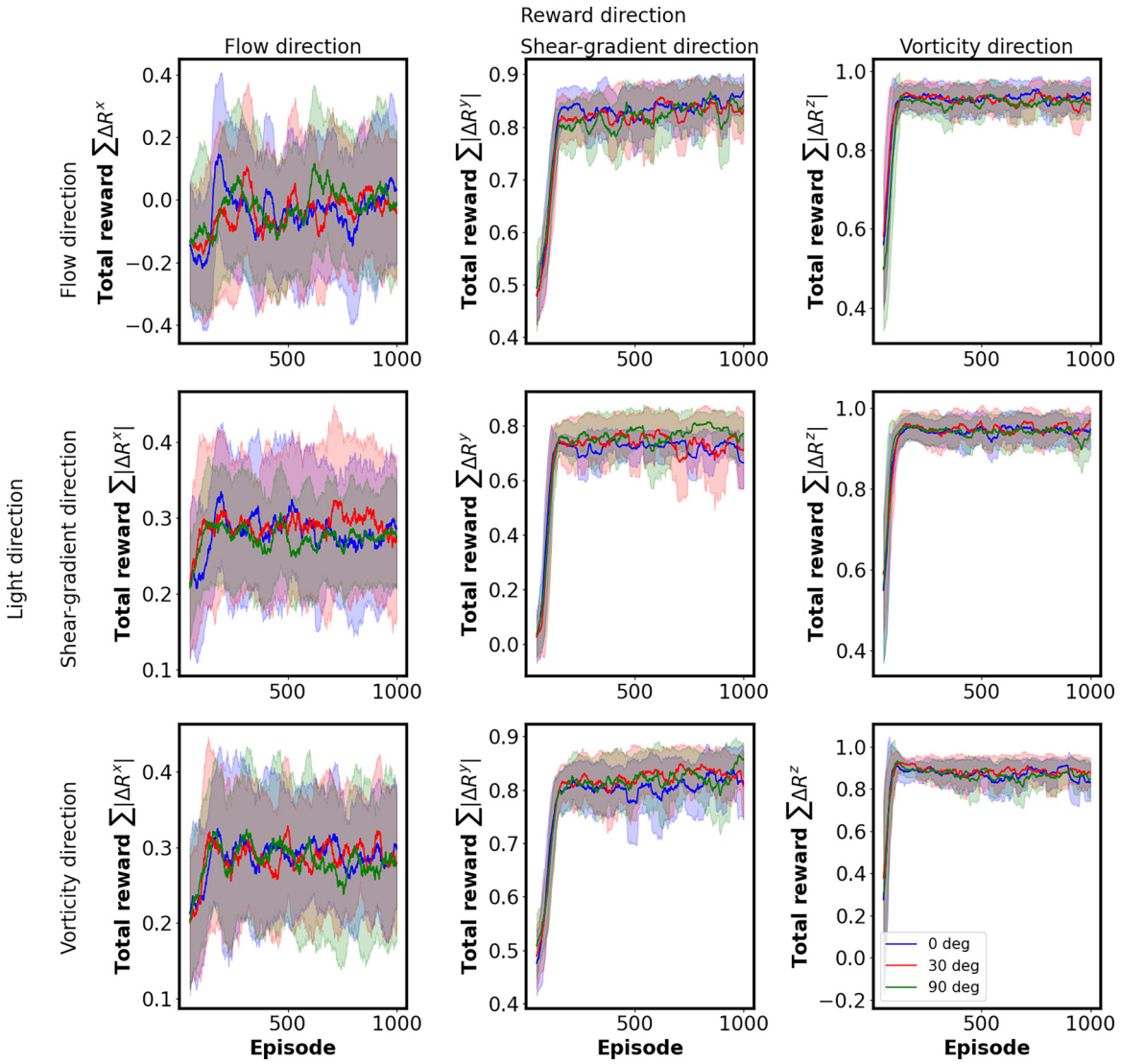

FIG. 5. The rolling average of the normalized total rewards, with an averaging window of 50 episodes, for visually aware swimmers tasked with migrating along different directions μ, under light sources shining along direction β. We consider swimmers with eyes located at $\Theta = 0^\circ$, 30°, and 90°, using $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$ as the state-defining variables.

of signals: surface stresses, light alignment, and light rotation (flashing). The results using these signals are found to be broadly similar to the results for swimmers able to use laboratory frame information; see Fig. 2. In Fig. 4, two different combinations of input signals were studied, i.e., $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L, (\hat{\boldsymbol{n}}_e \times \boldsymbol{\Omega}) \cdot \hat{\boldsymbol{n}}_L\}$ and $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$ alone. One can see that there is little qualitative difference between these two sets of parameters. Thus the surface stresses and the alignment with the light provide sufficient information to allow for efficient swimming. It is somewhat surprising that the swimmer learns to swim so well in the vorticity and shear-gradient directions. However, as for the swimmers with access to laboratory frame information, these swimmers are still unable to efficiently swim with the shear flow. We consider that this is due to the difficulty, for an unconstrained agent in 3D, of identifying the appropriate rotation that would lead it to the high-flow regions. Simply put, there is not enough sensory information to narrow down the rotational degrees of freedom to allow the swimmer to perform this task efficiently. To test this hypothesis, we constrained an agent to only be able to orient within the shear-flow–shear-gradient (x-y) plane, resulting in effective 2D motion with one rotational degree of freedom. We found that this restriction allowed the agent to efficiently target the high-flow regions, strengthening our argument that it is the excessive degrees of freedom, and lack of information, associated with rotating in 3D that make this such a difficult task in general. We further conducted an auxiliary simulation, where the agent was rewarded for staying oriented in the x-y plane, to test whether the agent could bypass this issue by first learning to move in the x-y plane and only then learning how to reorient within it. We found that the agent still struggled to target the x-y plane, even for this simpler task. The details are available in the Supplemental Material, Sec. VI [39].

To clarify what role, if any, is played by the positioning of the eye, we have compared the learning efficiency for different locations, $\Theta = 0^\circ$, 30°, and 90°, with $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$

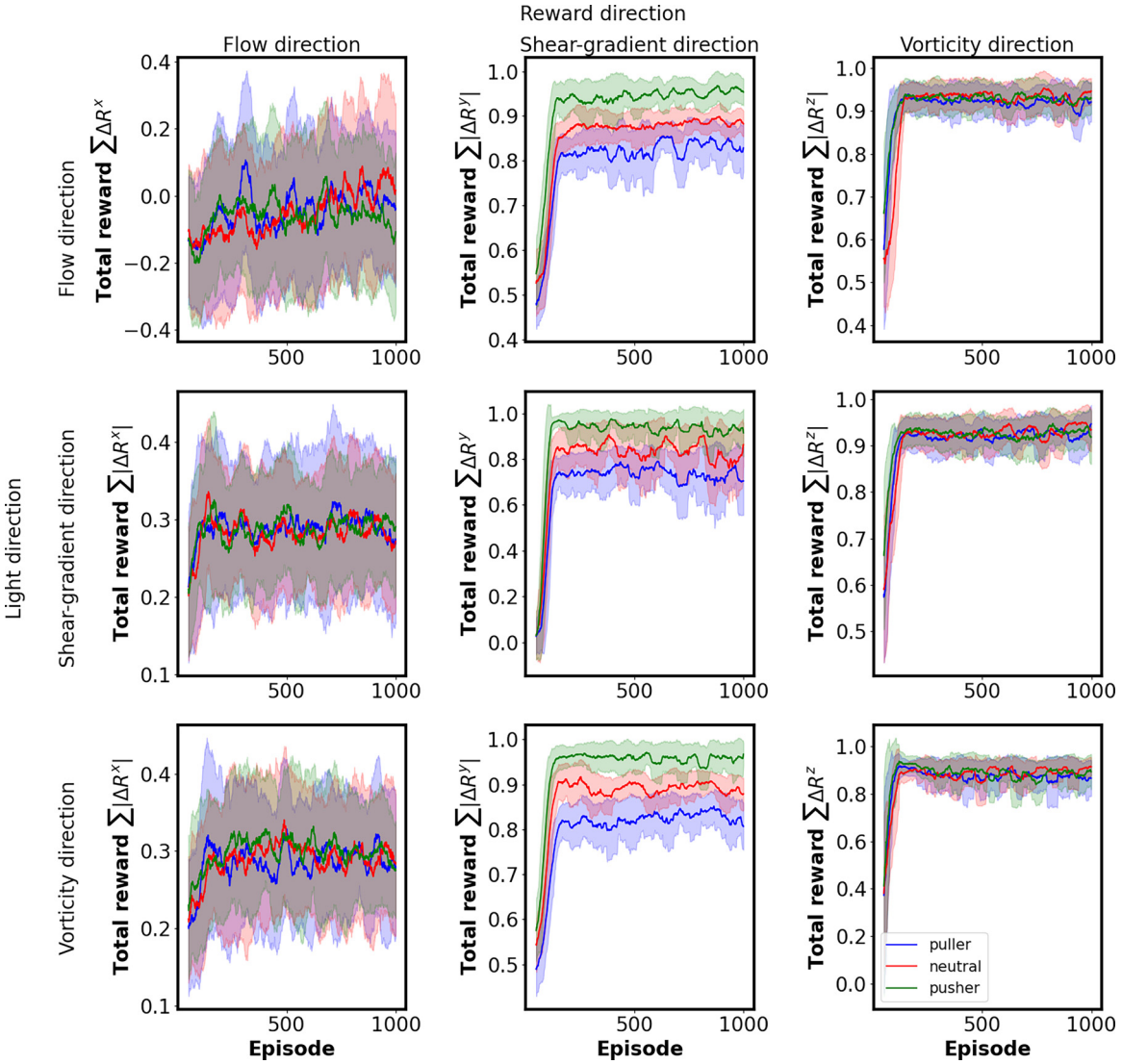

FIG. 6. The rolling average of the normalized total rewards, with an averaging window of 50 episodes, for different types of swimmers, with eyes located at $\Theta = 30^\circ$, using $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$ as the state-defining variables. Here we consider three distinct swimmer types: puller ($\alpha = 2$), neutral ($\alpha = 0$), and pusher ($\alpha = -2$).

as the input signals. The results are shown in Fig. 5, where a similar performance is obtained in all cases. Thus, for the tasks considered here, the choice of the eye's location seems to be relatively unimportant.

Finally, we also compare the performance for different swimmer types by considering the squirming parameters $\alpha = -2$ (pusher) and $\alpha = 0$ (neutral), in addition to the puller studied in the rest of this paper. The plots in Fig. 6 show the results obtained when using the surface stresses and light-alignment signals $\{\boldsymbol{\tau}_i, \hat{\boldsymbol{n}}_e \cdot \hat{\boldsymbol{n}}_L\}$. Here we found a surprising result for the task of swimming in the shear-gradient direction, in which there are clear differences between the policy efficiencies developed under different swimming modes. The pushers achieve the best performance, followed by neutral swimmers, with pullers being the least efficient. This distinction between swimming modes has also been observed in a confined system [45,46]. This results from the fact that each type of swimmer perceives the surrounding flow field differently (see Supplemental Material Sec. III [39]).

IV. CONCLUSIONS

We have performed direct numerical simulations, using the smoothed profile method, coupled with a deep reinforcement learning algorithm to investigate the learning performance of a swimmer under an applied zigzag flow. We considered three different swimming assignments, in which the swimmer is tasked with moving in the shear-flow (x), shear-gradient (y), or vorticity (z) directions. We demonstrated how different state information provided to the swimmer during the learning could result in vastly different performance. We studied the learning in cases where the swimmer receives either laboratory frame or body frame (local) variables. For the former, an efficient policy for migrating in the vorticity and shear-gradient directions emerged for swimmers given only their instantaneous orientation and the memory of the last two actions. However, for the task of swimming in the flow direction, the swimmer was unable to develop an efficient policy, meaning it could not target regions in which the flow was maximal.

For swimmers more closely inspired by microorganisms, e.g., copepods, we assumed that the swimmer has six force- or stress-sensing channels distributed on its surface, allowing it to sense relative differences with the local fluid velocity. This swimmer was also assumed to possess a sensor, e.g., a crude eye or photoreceptor that can sense light. Thus the model microorganism can also measure its alignment relative to the direction of a light source, and perhaps also the flashing of this light due to its relative rotation. We found that providing hydrodynamic forces, along with the light-alignment signal, was sufficient to develop efficient strategies to perform the swimming tasks. In particular, we observed similar efficiency to the case of swimmers trained on (global) laboratory frame information. Additionally, we found that the location of the eye was inconsequential to achieve these swimming tasks, with this particular flow field. We also investigated the differences in learning as a function of type of swimmer, by comparing pullers, pushers, and neutral swimmers. The pushers outperform the other two modes, with neutral swimmers performing better than pullers. This distinction in the performance among different swimming types arises from the different ways the swimmers sense the surrounding flow field. We hope that our work may help motivate future studies on efficient swimming strategies for active particles and also further our understanding of model biological swimmers. One of the remaining challenges to address relates to our use of an external torque to define the action of the swimmer. Thus, while our swimmers are force-free, they are not torque-free. This might be reasonable for certain artificial swimmers, but most biological microswimmers are both force-free and torque-free. In future work we will consider learning under torque-free conditions.

ACKNOWLEDGMENTS

This work was supported by the Grants-in-Aid for Scientific Research (JSPS KAKENHI) under Grants No. JP 20H00129, No. 20H05619, and No. 20K03786 and SPIRITS 2020 of Kyoto University. R.Y. acknowledges helpful discussions with Prof. Hajime Tanaka and Prof. Akira Furukawa. M.S.T. acknowledges funding from Warwick University's Turing AI fund. We acknowledge the Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures and the High Performance Computing Infrastructure in Japan (Projects No. jh21001-MDH and No. jh220054) for providing the computational resources of the Wisteria/BDEC-01 at the Information Technology Center, The University of Tokyo.

[1] D.-P. Häder and R. Hemmersbach, Gravitaxis in euglena, in *Euglena: Biochemistry, Cell and Molecular Biology*, Advances in Experimental Medicine and Biology (Springer, New York, 2017), Vol. 979, pp. 237–266.

[2] B. ten Hagen, F. Kümmel, R. Wittkowski, D. Takagi, H. Löwen, and C. Bechinger, Gravitaxis of asymmetric self-propelled colloidal particles, Nat. Commun. **5**, 4829 (2014).

[3] H. C. Berg and D. A. Brown, Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking, Nature (London) **239**, 500 (1972).

[4] S. de Oliveira, E. E. Rosowski, and A. Huttenlocher, Neutrophil migration in infection and wound repair: Going forward in reverse, Nat. Rev. Immunol. **16**, 378 (2016).

[5] B. J. Gemmell, D. Adhikari, and E. K. Longmire, Volumetric quantification of fluid flow reveals fish's use of hydrodynamic stealth to capture evasive prey, J. R. Soc., Interface **11**, 20130880 (2014).

[6] F.-G. Michalec, S. Souissi, and M. Holzner, Turbulence triggers vigorous swimming but hinders motion strategy in planktonic copepods, J. R. Soc., Interface **12**, 20150158 (2015).

[7] T. Kiørboe, E. Saiz, and A. Visser, Hydrodynamic signal perception in the copepod *Acartia tonsa*, Mar. Ecol. Prog. Ser. **179**, 97 (1999).

[8] L. K. E. A. Abdelmohsen, F. Peng, Y. Tu, and D. A. Wilson, Micro- and nano-motors for biomedical applications, J. Mater. Chem. B **2**, 2395 (2014).

[9] E. Karshalev, B. Esteban-Fernández de Ávila, and J. Wang, Micromotors for "Chemistry-on-the-Fly", J. Am. Chem. Soc. **140**, 3810 (2018).

[10] R. Fernandes and D. H. Gracias, Self-folding polymeric containers for encapsulation and delivery of drugs, Adv. Drug Delivery Rev. **64**, 1579 (2012).

[11] B. E.-F. de Ávila, P. Angsantikul, J. Li, M. Angel Lopez-Ramirez, D. E. Ramírez-Herrera, S. Thamphiwatana, C. Chen, J. Delezuk, R. Samakapiruk, V. Ramez, M. Obonyo, L. Zhang, and J. Wang, Micromotor-enabled active drug delivery for in vivo treatment of stomach infection, Nat. Commun. **8**, 272 (2017).

[12] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Flow Navigation by Smart Microswimmers via Reinforcement Learning, Phys. Rev. Lett. **118**, 158004 (2017).

[13] K. Gustavsson, L. Biferale, A. Celani, and S. Colabrese, Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning, Eur. Phys. J. E **40**, 110 (2017).

[14] J. Qiu, N. Mousavi, L. Zhao, and K. Gustavsson, Active gyrotactic stability of microswimmers using hydromechanical signals, Phys. Rev. Fluids **7**, 014311 (2022).

[15] G. Zhu, W.-Z. Fang, and L. Zhu, Optimizing low-Reynolds-number predation via optimal control and reinforcement learning, J. Fluid Mech. **944**, A3 (2022).

[16] A. Daddi-Moussa-Ider, H. Löwen, and B. Liebchen, Hydrodynamics can determine the optimal route for microswimmer navigation, Commun. Phys. **4**, 15 (2021).

[17] M. Nasiri and B. Liebchen, Reinforcement learning of optimal active particle navigation, New J. Phys. **24**, 073042 (2022).

[18] L. Biferale, F. Bonaccorso, M. Buzzicotti, P. Clark Di Leoni, and K. Gustavsson, Zermelo's problem: Optimal point-to-point navigation in 2D turbulent flows using reinforcement learning, Chaos **29**, 103138 (2019).

[19] R. Yamamoto, J. J. Molina, and Y. Nakayama, Smoothed profile method for direct numerical simulations of hydrodynamically interacting particles, Soft Matter **17**, 4226 (2021).

[20] H. van Hasselt, A. Guez, and D. Silver, Deep reinforcement learning with double Q-learning, arXiv:1509.06461.

[21] Y. Zhu, F.-B. Tian, J. Young, J. C. Liao, and J. C. S. Lai, A numerical study of fish adaption behaviors in complex environments with a deep reinforcement learning and immersed boundary–lattice Boltzmann method, Sci. Rep. **11**, 1691 (2021).
[22] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Smart inertial particles, Phys. Rev. Fluids **3**, 084301 (2018).
[23] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, Reinforcement learning with artificial microswimmers, Sci. Rob. **6**, eabd9285 (2021).
[24] C. L. Monteil and C. T. Lefevre, Magnetoreception in microorganisms, Trends Microbiol. **28**, 266 (2020).
[25] H. Almblad, T. E. Randall, F. Liu, K. Leblanc, R. A. Groves, W. Kittichotirat, G. L. Winsor, N. Fournier, E. Au, J. Groizeleau, J. D. Rich, Y. Lou, E. Granton, L. K. Jennings, L. A. Singletary, T. M. L. Winstone, N. M. Good, R. E. Bumgarner, M. F. Hynes, M. Singh *et al.*, Bacterial cyclic diguanylate signaling networks sense temperature, Nat. Commun. **12**, 1986 (2021).
[26] M. J. Lighthill, On the squirming motion of nearly spherical deformable bodies through liquids at very small Reynolds numbers, Commun. Pure Appl. Math. **5**, 109 (1952).
[27] J. R. Blake, A spherical envelope approach to ciliary propulsion, J. Fluid Mech. **46**, 199 (1971).
[28] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics* (Prentice-Hall, Englewood Cliffs, NJ, 1983).
[29] T. Iwashita and R. Yamamoto, Short-time motion of Brownian particles in a shear flow, Phys. Rev. E **79**, 031401 (2009).
[30] J. J. Molina, Y. Nakayama, and R. Yamamoto, Hydrodynamic interactions of self-propelled swimmers, Soft Matter **9**, 4923 (2013).
[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).
[32] M. Gazzola, 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).
[33] 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).
[34] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, Prioritized experience replay, arXiv:1511.05952.
[35] R. S. Sutton, Learning to predict by the methods of temporal differences, Mach. Learn. **3**, 9 (1988).
[36] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[37] K. Son, J. S. Guasto, and R. Stocker, Bacteria can exploit a flagellar buckling instability to change direction, Nat. Phys. **9**, 494 (2013).
[38] E. M. Purcell, Life at low Reynolds number, Am. J. Phys. **45**, 3 (1977).
[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.107.065102 for a detailed description of system parameters; a detailed discussion of the performance of a swimmer learning to perform the task of navigation in the flow direction in Fig. 2; principal component analysis (PCA) of the perceived hydrodynamic signals from the sensors on the surface of the swimmer to account for the difference in performance across types of swimmer; a detailed description of the choice of decay rate of the greedy parameter; a discussion about including the last two taken actions as the input signals; and a discussion about learning to swim in the flow direction in the x-y plane explaining the inferior performance of the task shown in Figs. 4–6.
[40] P. Gunnarson, I. Mandralis, G. Novati, P. Koumoutsakos, and J. O. Dabiri, Learning efficient navigation in vortical flow fields, Nat. Commun. **12**, 7143 (2021).
[41] C.-Y. Yang, M. Bialecka-Fornal, C. Weatherwax, J. W. Larkin, A. Prindle, J. Liu, J. Garcia-Ojalvo, and G. M. Süel, Encoding membrane-potential-based memory within a microbial community, Cell Syst. **10**, 417 (2020).
[42] J. Qiu, N. Mousavi, K. Gustavsson, C. Xu, B. Mehlig, and L. Zhao, Navigation of micro-swimmers in steady flow: The importance of symmetries, J. Fluid Mech. **932**, A10 (2022).
[43] F. Borra, L. Biferale, M. Cencini, and A. Celani, Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number, Phys. Rev. Fluids **7**, 023103 (2022).
[44] G. Kreimer, The green algal eyespot apparatus: A primordial visual system and more? Curr. Genet. **55**, 19 (2009).
[45] N. Oyama, J. J. Molina, and R. Yamamoto, Hydrodynamic alignment of microswimmers in pipes, arXiv:1612.00135.
[46] L. Zhu, E. Lauga, and L. Brandt, Low-Reynolds-number swimming in a capillary tube, J. Fluid Mech. **726**, 285 (2013).