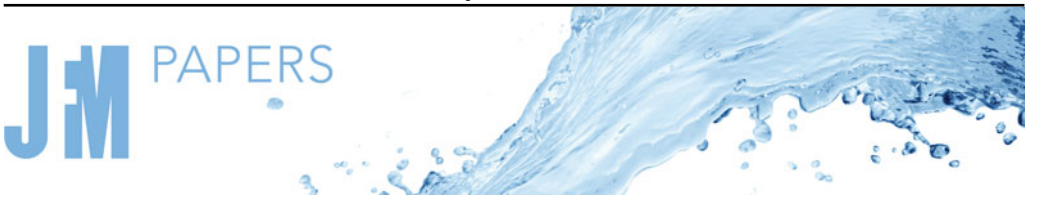
J. Fluid Mech. (2026), *vol.* 1039, A18, doi:10.1017/jfm.2026.11802



Learning to traverse convective flows at moderate to high Rayleigh numbers

Ao Xu[1,2], **Hua-Lin Wu**[1], **Ben-Rui Xu**[1] **and Heng-Dong Xi**[1,2]

[1]Institute of Extreme Mechanics, School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, PR China

[2]National Key Laboratory of Aircraft Configuration Design, Key Laboratory for Extreme Mechanics of Aircraft of Ministry of Industry and Information Technology, Xi'an 710072, PR China

Corresponding author: Heng-Dong Xi, hengdongxi@nwpu.edu.cn

(Received 18 November 2025 UTC; revised 12 June 2026 UTC; accepted 14 June 2026 UTC)

We study the navigation of a self-propelled inertial particle in two-dimensional Rayleigh–Bénard convection at Prandtl number $Pr = 0.71$ and cell aspect ratio $\Gamma = 4$ for Rayleigh numbers Ra ranging from 10^7 to 10^{11}. A reinforcement-learning (RL) controller selects the propulsive acceleration, subject to an upper bound $\mathcal{A}_{max}$, to achieve a prescribed horizontal displacement. We find that the success rate increases abruptly with $\mathcal{A}_{max}$ at moderate Ra, whereas at higher Ra the transition becomes more gradual and shifts to larger $\mathcal{A}_{max}$. Moreover, although the completion time increases with Ra, the propulsion energy required for successful traversal decreases. Proper orthogonal decomposition indicates that these performance differences are associated with reorganisation of the carrier flow. At moderate Ra, the dominant large-scale circulation partitions the domain through persistent transport barriers, requiring a finite thrust surplus to cross them; at higher Ra, energy is distributed across many modes, the barriers fragment and transient plume-assisted pathways emerge. Compared with a constant-heading baseline, the learned policy aligns with local currents and consumes significantly less energy. Lagrangian coherent structure analysis further suggests that the RL agent tends to cross repelling barriers and surf along attracting pathways. Finally, by mapping these behaviours onto the local Eulerian flow topology using Voronoi tessellation and the Q-criterion, we distil an interpretable, physics-based heuristic strategy that retains robust navigability. These results connect turbulent-flow organisation with autonomous navigation under bounded actuation.

Key words: Bénard convection, plumes/thermals, machine learning

1. Introduction

Self-propelled particles span length scales from micrometres to metres and harvest energy from their surroundings to generate motion (Bechinger *et al.* 2016). Natural examples include molecular motors, cells, bacteria, fish and birds, which are powered by chemical or biological energy (Gustavsson *et al.* 2016; Sengupta, Carrara & Stocker 2017; Qiu, Marchioli & Zhao 2022*a*). Artificial counterparts include self-propelled microparticles and macroscopic robots driven by thermal or electrical energy (Cichos *et al.* 2020). Trajectory planning is crucial in many settings, including microorganisms navigating turbulent flows (Monthiller *et al.* 2022; Qiu *et al.* 2022*c*), cells migrating along chemical gradients (Sengupta *et al.* 2017) and birds undertaking long-distance flights by thermal soaring (Weimerskirch *et al.* 2016). These navigation problems can be classified according to the environmental information used to guide motion. Short-range navigation may rely on sensing hydrodynamic disturbances (Qiu *et al.* 2022*b*; Borra *et al.* 2022; Zhu, Fang & Zhu 2022; Xu, Wu & Xi 2022), medium-range navigation on chemotaxis or infotaxis (Vergassola, Villermaux & Shraiman 2007; Loisy & Eloy 2022; Rigolli *et al.* 2022) and long-range navigation on coupled horizontal and vertical dispersion (Colabrese *et al.* 2017; Gustavsson *et al.* 2017; Biferale *et al.* 2019; Gunnarson *et al.* 2021; Xu, Wu & Xi 2023). A substantial body of work on Lagrangian navigation builds on Zermelo's problem of determining optimal routes in unsteady currents using level-set methods, graph search and stochastic control. Because models and observations are often imperfect, recent studies have incorporated estimation and adaptation into the planning process through learning-based flow prediction, belief-space planning and reinforcement-learning (RL) control (Simons 2004; Piro *et al.* 2025; Heinonen *et al.* 2025). Compared with traditional model predictive control, which requires computationally expensive forward prediction and is highly sensitive to the exponential divergence of trajectories in spatiotemporally chaotic flows (Krishna *et al.* 2022, 2023), model-free RL offers particular advantages. It can learn an instantaneous reactive policy from local sensing, thereby reducing the need for real-time flow forecasting. Furthermore, understanding how well such learned policies generalise across flow regimes is important for practical deployment (Jiao *et al.* 2025; Hang *et al.* 2026).

Investigations of self-propelled particles in natural systems have also inspired the design of artificial swimmers and autonomous vehicles for complex environments. For example, Muiños-Landin *et al.* (2021) studied a self-thermophoretic microswimmer in real time and found that its learning efficiency decreased under strong Brownian noise. Related studies have also examined the control of autonomous vehicles in aerial and marine environments. Reddy *et al.* (2016) simulated a glider ascending in a spiral pattern reminiscent of soaring birds (Akos, Nagy & Vicsek 2008); this concept was later demonstrated in field experiments, in which a two-metre glider navigated autonomously using atmospheric thermals (Reddy *et al.* 2018). Bellemare *et al.* (2020) deployed balloons in the stratosphere that used wind speed, altitude and direction for station keeping and travel, enabling long-duration monitoring of air quality and environmental variables. Masmitja *et al.* (2023) developed an autonomous surface vehicle capable of locating and tracking moving underwater targets, and validated the method through sea trials involving several autonomous platforms, including the Sparus II autonomous underwater vehicle.

In both atmospheric and oceanic environments, convective currents play a key role in determining navigation paths that minimise time or energy (Biferale *et al.* 2019; Laurent *et al.* 2021). These flows are characterised by buoyancy-driven turbulence spanning a wide range of scales, presenting both opportunities and challenges for navigation. Intermittent updrafts and cross-stream currents can reduce propulsion requirements, but

three main features of convective turbulence complicate control: multi-scale variability, which demands continuous sensing and adaptation across a broad range of spatial and temporal scales; rapid decorrelation, which shortens predictability horizons; and spatial intermittency, which generates localised updrafts and preferred pathways. To analyse such convective flows, we use Rayleigh–Bénard (RB) convection as an idealised canonical model for buoyancy-driven convection relevant to atmospheric and oceanic flows (Lohse & Xia 2010; Chillà & Schumacher 2012; Pandey, Scheel & Schumacher 2018; Wang, Zhou & Sun 2020*a*; Xia *et al.* 2023; Lohse & Shishkina 2023, 2024; Shishkina & Lohse 2024; Xia *et al.* 2025).

The primary control parameters of the RB system are the Rayleigh number (Ra, which measures buoyancy relative to viscous and thermal diffusion effects), the Prandtl number (Pr, which characterises the thermophysical properties of the fluid) and the cell aspect ratio (Γ, which defines the geometry of the convection domain). At moderate Ra, the flow organises into a roll-like large-scale circulation (LSC) that persists over many convective times. Boundaries between neighbouring cells, where the mean circulation reverses, are relatively well defined. Because plumes and roll edges remain coherent in space and time, transport is channelled through a few preferred pathways across the domain. As Ra increases, plume production intensifies, coherent rolls weaken and the interior develops richer small-scale structure (Zhu *et al.* 2018; Samuel, Samtaney & Verma 2022). Temporal correlations of velocity and temperature then decay more rapidly on convective time scales, and transport pathways become less uniform because plume activity is increasingly intermittent. These regime changes matter for guidance and control because they affect the navigability of the flow under a fixed actuation limit. Specifically, flows dominated by LSC exhibit distinct pathways and barriers that influence reachability and the actuation required to move between basins, whereas flows at higher Ra offer more diffuse routes whose availability varies on short time scales. This motivates a closer investigation of how a self-propelled particle adapts its propulsive acceleration as turbulence intensity varies.

In thermal soaring and energy-harvesting flight, heuristic and optimal controllers exploit updrafts to extend flight range. More recently, deep reinforcement learning has been applied to flow control and to navigation in unsteady or turbulent environments, demonstrating flow-aware behaviour under a fixed actuation limit. However, most existing studies consider a single flow intensity and assess performance primarily in terms of time to target or qualitative success. It remains unclear how turbulence intensity determines reachability thresholds, shapes the trade-off between energy and time and influences the performance of learned guidance policies. In this work, we address this question by studying the navigation of self-propelled particles in convective environments at moderate to high Rayleigh numbers. Because classical optimal-navigation methods can be sensitive to small disturbances in chaotic flows, we adopt a data-driven approach based on an RL algorithm for trajectory optimisation (Mehta *et al.* 2019; Brunton, Noack & Koumoutsakos 2020). In recent years, RL has been increasingly used in the active-matter community, from early demonstrations of RL-driven active particles in simple environments (Schneider & Stark 2019) to more complex navigation and obstacle-avoidance tasks (Nasiri & Liebchen 2022), as well as the discovery of energy-efficient swimming gaits (Zhu *et al.* 2025). The rest of this paper is organised as follows. In § 2, we present the numerical methodology and the RL framework. In § 3, we evaluate navigation performance in terms of reachability, completion time and energy expenditure across Rayleigh numbers and actuation limits. We then interpret these trends using proper orthogonal decomposition (POD) and Lagrangian coherent structures (LCS) to examine how the agent balances energy trade-offs and traverses transport barriers, and map the

learned policy onto the local Eulerian flow topology to distil an interpretable heuristic strategy. Finally, in § 4, we summarise the main findings of the present work.

2. Numerical method

2.1. *Direct numerical simulation of thermal turbulence*

We consider incompressible thermal convection under the Boussinesq approximation, in which temperature acts as an active scalar and enters the momentum equation through buoyancy. All transport coefficients are taken to be constant. The governing equations are

$$\boldsymbol{\nabla}\cdot\boldsymbol{u}_f = 0, \tag{2.1}$$

$$\frac{\partial \boldsymbol{u}_f}{\partial t} + \boldsymbol{u}_f\cdot\boldsymbol{\nabla}\boldsymbol{u}_f = -\frac{1}{\rho_0}\boldsymbol{\nabla}P + \nu\boldsymbol{\nabla}^2\boldsymbol{u}_f + g\beta(T-T_0)\,\hat{\boldsymbol{y}}, \tag{2.2}$$

$$\frac{\partial T}{\partial t} + \boldsymbol{u}_f\cdot\boldsymbol{\nabla}T = \alpha\boldsymbol{\nabla}^2 T, \tag{2.3}$$

where $\boldsymbol{u}_f$ is the fluid velocity and P and T are the fluid pressure and temperature, respectively. The parameters β, ν and α denote the thermal expansion coefficient, kinematic viscosity and thermal diffusivity, respectively. The subscript '0' indicates a reference value, and g is the magnitude of the gravitational acceleration. The unit vector $\hat{\boldsymbol{y}}$ points opposite to the direction of gravity.

We introduce the following dimensionless variables:

$$\begin{gathered}\boldsymbol{x}^* = \frac{\boldsymbol{x}}{H}, \quad t^* = \frac{t}{\sqrt{H/(g\beta\Delta_T)}}, \quad \boldsymbol{u}_f^* = \frac{\boldsymbol{u}_f}{\sqrt{g\beta\Delta_T H}}, \\ P^* = \frac{P}{\rho_0 g\beta\Delta_T H}, \quad T^* = \frac{T-T_0}{\Delta_T}.\end{gathered} \tag{2.4}$$

With these definitions, (2.1)–(2.3) become, in dimensionless form,

$$\boldsymbol{\nabla}\cdot\boldsymbol{u}_f^* = 0, \tag{2.5}$$

$$\frac{\partial \boldsymbol{u}_f^*}{\partial t^*} + \boldsymbol{u}_f^*\cdot\boldsymbol{\nabla}\boldsymbol{u}_f^* = -\boldsymbol{\nabla}P^* + \sqrt{\frac{Pr}{Ra}}\,\boldsymbol{\nabla}^2\boldsymbol{u}_f^* + T^*\hat{\boldsymbol{y}}, \tag{2.6}$$

$$\frac{\partial T^*}{\partial t^*} + \boldsymbol{u}_f^*\cdot\boldsymbol{\nabla}T^* = \sqrt{\frac{1}{Pr\,Ra}}\,\boldsymbol{\nabla}^2 T^*, \tag{2.7}$$

where H is the cell height and Δ_T is the temperature difference between the heated bottom wall and the cooled top wall. The two dimensionless control parameters appearing in the governing equations are the Rayleigh number, Ra, and the Prandtl number, Pr, defined as

$$Ra = \frac{g\beta\Delta_T H^3}{\nu\alpha}, \quad Pr = \frac{\nu}{\alpha}. \tag{2.8}$$

We perform the direct numerical simulations using the spectral element method implemented in the open-source Nek5000 solver (version v19.0). This method is suitable for the present problem because its high-order spatial accuracy and low numerical dissipation allow us to resolve a broad range of turbulent scales, capture fine intermittent structures and maintain numerical stability at high Rayleigh numbers ($Ra \geq 10^{10}$). In Nek5000, the effective resolution in each spatial direction is determined by the product of the number of spectral elements and the polynomial order. Here, we use polynomial order $N = 9$ for the spectral-element discretisation. The simulations span the range

Ra	$N_x \times N_y$	$(\Delta_g)_{max}/\eta_K$	$(\Delta_g)_{max}/\eta_B$	$(\Delta_t)_{max}/\tau_\eta$	Nu (present)	Nu (Zhu *et al.* 2018)
10^7	1152×288	0.19	0.16	0.00021	14.5	—
10^8	1152×288	0.97	0.81	0.00176	27.7	26.1
10^9	2304×576	1.23	1.03	0.00137	48.0	48.3
10^{10}	4608×1152	1.32	1.11	0.00069	95.1	95.1
10^{11}	9216×2304	1.23	1.03	0.00037	185.2	193.6

Table 1. *A posteriori* verification of spatial and temporal resolutions for the present direct numerical simulations. The columns list, from left to right: Rayleigh number Ra; grid resolution $N_x \times N_y$; maximum wall-normal grid spacing relative to the Kolmogorov length scale, $(\Delta_g)_{max}/\eta_K$, and to the Batchelor length scale, $(\Delta_g)_{max}/\eta_B$; maximum time step relative to the Kolmogorov time scale, $(\Delta_t)_{max}/\tau_\eta$; volume-averaged Nusselt number Nu from the present simulations; and reference Nu values from Zhu *et al.* (2018).

$10^7 \leq Ra \leq 10^{11}$ at fixed Prandtl number $Pr = 0.71$. To ensure adequate spatial and temporal resolution, we perform an *a posteriori* verification by examining the number of grid points within the thermal boundary layer, the maximum wall-normal grid spacing $(\Delta_g)_{max}$ relative to the Kolmogorov and Batchelor length scales and the maximum time interval $(\Delta_t)_{max}$ relative to the Kolmogorov time scale. The complete set of simulation parameters is listed in table 1. The Kolmogorov length scale is estimated as $\eta_K = (\nu^3/\langle\varepsilon_u\rangle_{V,t})^{1/4}$, while the Batchelor length scale is given by $\eta_B = \eta_K Pr^{-1/2}$. The Kolmogorov time scale is calculated as $\tau_\eta = \sqrt{\nu/\langle\varepsilon_u\rangle_{V,t}}$, where $\langle\varepsilon_u\rangle_{V,t}$ denotes the volume- and time-averaged turbulent kinetic energy dissipation rate, computed as $\langle\varepsilon_u\rangle_{V,t} = \nu\langle(\partial_j u_i')^2\rangle$, with u_i' denoting the fluctuating velocity component. As summarised in table 1, the maximum wall-normal grid spacing is of the same order as the Kolmogorov and Batchelor length scales, while the maximum time step remains substantially smaller than the Kolmogorov time scale in all cases. We also report the volume-averaged Nusselt number Nu from the present simulations alongside the reference values of Zhu *et al.* (2018), and find good agreement. Details of the spectral element method and validation of the Nek5000 solver can be found in Fischer (1997) and Kooij *et al.* (2018). As an additional verification, we also carried out simulations at $Ra = 10^7$ and 10^8 using an in-house solver based on the lattice Boltzmann method (Xu, Shi & Zhao 2017). The results from Nek5000 and the lattice Boltzmann solver show consistent flow fields and statistics.

2.2. *Kinematics of self-propelled inertial particles*

We consider particles that are sufficiently small that they do not disturb the carrier flow. Here, 'small' means that the particle diameter is smaller than the Kolmogorov length scale and larger than the molecular mean free path, so that Brownian motion is negligible. The particles are assumed to be isotropic, and in the present framework we consider only their translational motion. In a fluid environment with strong shear and vorticity, such as turbulent RB convection, particles are naturally subjected to flow-induced torques and rotational dynamics (e.g. Jeffery-type rotations). By restricting the model to translational kinematics, we effectively assume an idealised agent that either possesses an isotropic propulsion mechanism, capable of redirecting thrust without rotating its body, or has a negligible rotational relaxation time, thereby achieving instantaneous attitude control. Although this assumption allows us to isolate the higher-level trajectory-planning problem from the lower-level attitude-control problem, it also removes certain physical complexities. In particular, neglecting rotational dynamics means that the agent does not experience the continuous flow-induced torques that would otherwise perturb its orientation. Consequently, a real biological or artificial swimmer would likely incur an

additional energetic cost to generate corrective torques against the local fluid vorticity in order to maintain its desired heading.

For the sub-Kolmogorov particles considered here, this additional rotational cost can be estimated by comparing the power required for rotational correction with that required for translational propulsion. For a small spherical particle, the translational power scales as $P_{trans} \sim 6\pi\mu a U^2$, while the rotational power scales as $P_{rot} \sim 8\pi\mu a^3\Omega^2$, where a is the particle radius, U is a characteristic relative translational velocity and Ω is a characteristic relative angular velocity. Estimating the angular velocity from the local Kolmogorov-scale velocity gradient, $\Omega \sim U/\eta_K$, gives $P_{rot}/P_{trans} \sim (4/3)(a/\eta_K)^2$. In the present study, $a = 10\,\mu$m, while η_K is of the order of 1.6×10^{-3} to 1.8×10^{-3} m across the considered Rayleigh-number range. The resulting ratio is therefore $P_{rot}/P_{trans} \sim O(10^{-5})$. This scaling estimate suggests that the rotational energy correction is small compared with the translational propulsion cost for the present sub-Kolmogorov regime, although rotational dynamics may no longer be negligible for larger, anisotropic or finite-size swimmers.

The total force acting on the particle consists of the net gravitational force $\boldsymbol{F}_{gravity}$, the drag force $\boldsymbol{F}_{drag}$ and the propulsive force $\boldsymbol{F}_{propel}$. In the present heavy-particle and sub-Kolmogorov regime ($\rho_p \gg \rho_f$, $St \sim 10^{-3}$), added-mass, undisturbed-flow pressure-gradient, shear-induced lift, and Basset history forces are expected to be small compared with the drag force and are therefore neglected. The net gravitational force $\boldsymbol{F}_{gravity}$, acting along the direction of gravitational acceleration after accounting for hydrostatic buoyancy, is given by

$$\boldsymbol{F}_{gravity} = \rho_p V_p \boldsymbol{g} - \rho_f V_p \boldsymbol{g}, \tag{2.9}$$

where ρ_p and V_p denote the particle density and volume, respectively. The drag force $\boldsymbol{F}_{drag}$ arises from the relative motion between the particle and the surrounding fluid and is given by

$$\boldsymbol{F}_{drag} = \frac{m_p}{\tau_p}\left(\boldsymbol{u}_f - \boldsymbol{u}_p\right) f(Re_p), \tag{2.10}$$

where m_p and $\boldsymbol{u}_p$ denote the particle mass and velocity, respectively. The particle response time is $\tau_p = \rho_p d_p^2/(18\rho_f \nu)$, where d_p is the particle diameter. The particle Reynolds number is $Re_p = d_p\|\boldsymbol{u}_f - \boldsymbol{u}_p\|/\nu$, which determines the drag correction factor $f(Re_p)$. For $Re_p \ll 1$, Stokes drag applies and $f \approx 1$; for finite Re_p, we use the Schiller–Naumann correction $f(Re_p) = 1 + 0.15\,Re_p^{0.687}$. The propulsive force $\boldsymbol{F}_{propel} = m_p \boldsymbol{a}_{propel}$ represents the particle's self-generated actuation and yields a propulsive acceleration bounded above by a prescribed limit. In natural systems, this actuation may arise from biochemical or thermal processes, whereas in artificial systems it may be driven by thermal, electrical or mechanical energy sources.

The particle motion is governed by

$$\frac{\mathrm{d}\boldsymbol{x}_p(t)}{\mathrm{d}t} = \boldsymbol{u}_p(t), \tag{2.11}$$

$$\frac{\mathrm{d}\boldsymbol{u}_p(t)}{\mathrm{d}t} = \frac{\rho_p - \rho_f}{\rho_p}\boldsymbol{g} + \frac{\boldsymbol{u}_f(\boldsymbol{x}_p, t) - \boldsymbol{u}_p(t)}{\tau_p} f(Re_p) + \boldsymbol{a}_{propel}(t), \tag{2.12}$$

where $\boldsymbol{x}_p$ denotes the particle position. After non-dimensionalisation using the scales introduced above, the governing equations become

$$\frac{\mathrm{d}\boldsymbol{x}_p^*}{\mathrm{d}t^*} = \boldsymbol{u}_p^*, \tag{2.13}$$

$$\frac{\mathrm{d}\boldsymbol{u}_p^*}{\mathrm{d}t^*} = -\Lambda\,\hat{\boldsymbol{y}} + \frac{\boldsymbol{u}_f^* - \boldsymbol{u}_p^*}{St}\,f(Re_p) + \boldsymbol{a}_{propel}^*, \tag{2.14}$$

where $\Lambda = (\rho_p - \rho_f)/(\rho_p \beta \Delta_T)$ is the buoyancy ratio, which quantifies the particle buoyancy relative to the thermal buoyancy scale of the carrier flow. The particle Stokes number is defined as $St = \tau_p / t_f$, where the free-fall time scale is $t_f = \sqrt{H/(g\beta\Delta_T)}$. The propulsion strength is characterised by the dimensionless activity number

$$\mathcal{A} = \frac{\|\boldsymbol{a}_{propel}\|}{g}, \tag{2.15}$$

which measures the ratio of the particle's propulsive acceleration to gravity. The dimensionless propulsion term can therefore be written as

$$\boldsymbol{a}_{propel}^* = \mathcal{A}\frac{\rho_p}{\rho_p - \rho_f}\Lambda\,\hat{\boldsymbol{a}}_{propel}, \tag{2.16}$$

where $\hat{\boldsymbol{a}}_{propel} = \boldsymbol{a}_{propel}/\|\boldsymbol{a}_{propel}\|$ is the unit vector in the actuation direction. In the heavy-particle limit ($\rho_p \gg \rho_f$), $\rho_p/(\rho_p - \rho_f) \approx 1$, and the expression simplifies to

$$\boldsymbol{a}_{propel}^* = \mathcal{A}\Lambda\,\hat{\boldsymbol{a}}_{propel}. \tag{2.17}$$

Accordingly, the dimensionless momentum equation reduces to

$$\frac{\mathrm{d}\boldsymbol{u}_p^*}{\mathrm{d}t^*} = -\Lambda\,\hat{\boldsymbol{y}} + \frac{\boldsymbol{u}_f^* - \boldsymbol{u}_p^*}{St}\,f(Re_p) + \mathcal{A}\Lambda\,\hat{\boldsymbol{a}}_{propel}. \tag{2.18}$$

2.3. *Optimal control of self-propelling particles*

We control the particle through its propulsive acceleration. The control input is given by

$$\boldsymbol{a}_{propel} = \frac{\boldsymbol{F}_{propel}}{m_p}, \tag{2.19}$$

with its magnitude bounded by $\|\boldsymbol{a}_{propel}\|/g = \mathcal{A} \leqslant \mathcal{A}_{max}$, where $\mathcal{A}_{max}$ is the prescribed actuation limit. At each decision step, the self-propelled particle serves as the agent in an RL framework. It receives an observation of the local flow state, selects a propulsive acceleration and then receives a reward together with the next observation. Through repeated interactions, the agent seeks a policy that maximises the expected cumulative return (Mehta *et al.* 2019; Brunton *et al.* 2020).

Reinforcement-learning methods are commonly classified as either policy-based or value-based. Policy-based methods, such as policy-gradient methods, optimise the parameters $\boldsymbol{\theta}$ of a stochastic policy to maximise a performance objective $J(\pi_{\boldsymbol{\theta}})$. They naturally accommodate continuous actions, but can require many samples and may converge slowly, making them less suitable for complex flow problems. By contrast, value-based methods, such as Q-learning, learn an action–value function $Q_{\boldsymbol{\theta}}(s, a)$ and derive a deterministic policy by selecting the action that maximises it, i.e. $a(s) = \arg\max_a Q_{\boldsymbol{\theta}}(s, a)$. They are efficient for discrete actions but are generally not well suited to continuous control.

For navigation problems with continuous state and action spaces, the soft actor–critic (SAC) method is well suited because it combines policy-based and value-based components. The actor network selects the propulsive acceleration, while the critic

network evaluates the corresponding value function to guide learning. The SAC algorithm maximises not only the expected cumulative reward but also the policy entropy, thereby encouraging exploration. The optimal policy π^* is obtained by solving

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left\{r_t(s_t, a_t, s_{t+1}) + \alpha\mathcal{H}[\pi(\cdot|s_t)]\right\}\right], \tag{2.20}$$

where γ is the discount factor, and the coefficient α controls the trade-off between the entropy term and the reward. The reward r_t depends on the state s_t, the chosen action a_t and the resulting transition to s_{t+1}. The entropy $\mathcal{H}$ of the policy at state s_t measures the randomness of the selected actions and is computed from the action distribution as $\mathcal{H}[\pi(\cdot|s_t)] = \mathbb{E}_{a_t\sim\pi(\cdot|s_t)}[-\log\pi(a_t|s_t)]$. The SAC algorithm therefore seeks a policy that maximises a weighted sum of expected reward and entropy, enabling efficient exploration while maintaining training stability (Haarnoja *et al.* 2018). Detailed descriptions of the SAC formulation are provided in Appendix A.

Key components of the RL framework include the environmental cues available to the agent (i.e. the current state s_t), the agent's actions (i.e. a_t) and the feedback it receives for its behaviour (i.e. the reward r_t). The observation at time t^* is given by

$$s_t = \left\{\boldsymbol{x}_p^*, \boldsymbol{u}_p^*, \boldsymbol{a}_p^*, \boldsymbol{u}_f^*, \boldsymbol{\nabla}\boldsymbol{u}_f^*, T_f^*, (\boldsymbol{\nabla}T_f)_x^*\right\}, \tag{2.21}$$

where $\boldsymbol{\nabla}\boldsymbol{u}_f^*$ denotes the local velocity gradient and $(\boldsymbol{\nabla}T_f)_x^*$ the local horizontal temperature gradient. We emphasise that the agent relies on locally measurable environmental cues evaluated at its instantaneous position, rather than on any globally available description of the flow field. Although sensing the absolute location $\boldsymbol{x}_p$ is typically unfeasible for microscopic swimmers, it is readily available for macroscopic autonomous vehicles through standard positioning systems. Recent studies have cautioned that including absolute coordinates may allow agents to artificially 'memorise' static flow signatures in steady or periodic flows (Gunnarson *et al.* 2021; Jiao *et al.* 2025). However, at the moderate to high Rayleigh numbers considered here, thermal convection is spatiotemporally chaotic, and the locations of plumes and coherent structures continuously drift and evolve. This inherent unsteadiness reduces the likelihood that the agent simply overfits to absolute coordinates. We therefore retain the spatial information $\boldsymbol{x}_p$ because the task involves a fixed-displacement target and the agent requires global context to assess its progress. Furthermore, unlike in our previous study at lower Ra (Xu *et al.* 2023), the inclusion of local gradients $(\boldsymbol{\nabla}\boldsymbol{u}_f, (\boldsymbol{\nabla}T_f)_x)$ becomes important at high Rayleigh numbers owing to the emergence of finer characteristic scales. The information carried by $(\boldsymbol{\nabla}T_f)_x$ cannot be inferred solely from the instantaneous velocity gradient $\boldsymbol{\nabla}\boldsymbol{u}_f$. Whereas $\boldsymbol{\nabla}\boldsymbol{u}_f$ encodes local shear and strain, $(\boldsymbol{\nabla}T_f)_x$ is related to vorticity production in two-dimensional thermal convection. It therefore provides the agent with additional information about local flow acceleration, a physical mechanism specific to active-scalar turbulence.

A systematic ablation study examining the role and non-redundancy of these positional and gradient-based observations is provided in Appendix B. Briefly, the ablation results show that the relative importance of different sensory inputs depends on the flow regime. At moderate Ra, for example $Ra = 10^8$, the flow remains relatively organised by LSC, and local gradient cues are correlated with the large-scale transport structure, leading to a clear improvement in navigation performance. At higher Ra, for example $Ra = 10^{10}$, the flow becomes more intermittent and less spatially coherent. In this regime, explicit positional information $\boldsymbol{x}_p$ becomes important for maintaining global progress towards the fixed-displacement target. The ablation study also shows that, at high Ra, providing the agent with local velocity-gradient information $\boldsymbol{\nabla}\boldsymbol{u}_f$ yields better navigation performance

than providing the local temperature field T alone. A possible physical explanation is that thermal plumes become increasingly fragmented and intermittent at high Ra, so the scalar temperature field may become less directly predictive of subsequent transport. By contrast, the velocity-gradient tensor encodes local kinematic topology, including shear layers and vortex boundaries through quantities such as the Q-criterion. It therefore provides the agent with local cues that help it avoid trapping in vortical regions and access strain-dominated transport pathways.

The action corresponds to the particle's propulsive acceleration. We parametrise it as a bounded continuous action in polar form,

$$\boldsymbol{a}_{propel} = \mathcal{A}_{max} g[c_a \cos(\theta),\ c_a \sin(\theta)], \tag{2.22}$$

with $c_a = (a_0 + 1)/2 \in [0, 1]$ and $\theta = \pi a_1/2 \in [-\pi/2, \pi/2]$. This half-plane restriction reflects the physical constraint that the particle cannot generate reverse thrust. To assess the robustness of this choice, we conducted sensitivity checks for representative cases, $Ra = 10^8$ and 10^{10}, by evaluating agents with a full action space, $\theta \in [-\pi, \pi]$. These unconstrained agents learned to avoid backward thrust in order to maximise the reward, yielding success-rate curves nearly identical to those of the constrained cases. The variables a_0 and a_1 are the normalised outputs of the policy and lie in $[-1, 1]$. The realised dimensionless actuation magnitude therefore satisfies $\|\boldsymbol{a}_{propel}\|/g = \mathcal{A} = \mathcal{A}_{max} c_a \leq \mathcal{A}_{max}$.

The reward function balances progress towards the target against the cost of propulsion,

$$r_t = \frac{1}{R + Q}\left[R\ V^*_{eff}(t) - Q\ \|\boldsymbol{a}^*_{propel}(t)\|\right], \tag{2.23}$$

where $V^*_{eff}(t) = \boldsymbol{u}^*_p(t) \cdot \boldsymbol{n}$ is the dimensionless particle velocity along the target direction $\boldsymbol{n}$. The first term promotes rapid motion towards the goal, whereas the second penalises excessive propulsive effort. The coefficients R and Q determine the trade-off between completion time and energy expenditure. The normalisation factor $1/(R + Q)$ renders the reward a normalised weighted sum of the two physical objectives. This helps ensure that the magnitude of the single-step reward remains bounded by the characteristic physical scales, namely the dimensionless velocity V^*_{eff} and the actuation limit $\mathcal{A}_{max}$, thereby reducing the risk of arbitrary inflation of the reward signal and unstable gradients during training for different choices of R and Q. To guide the reader through the subsequent analyses, we note that in the primary evaluations presented in § 3 we fix the reward-weighting ratio at a large value of $R/Q = 5000$. As shown in § 3.3, a sufficiently large R/Q ratio is useful in highly chaotic RB convection. It helps the agent prioritise reachability and time efficiency, which are required to cross transport barriers between convective rolls, while the energy-penalty term simultaneously discourages redundant or purely random actuation. A detailed parametric investigation of this trade-off, including the limiting behaviours of time-optimal and energy-optimal policies, is deferred to § 3.3.

2.4. *Simulation settings*

We study the motion of self-propelled particles in a two-dimensional convection cell of size $L \times H$, with aspect ratio $\Gamma = L/H = 4$. The top and bottom boundaries are maintained at constant cold and hot temperatures, respectively, and satisfy no-slip velocity boundary conditions. The left and right boundaries are periodic. Simulations are performed over the Rayleigh-number range $10^7 \leq Ra \leq 10^{11}$ at fixed Prandtl number $Pr = 0.71$. Because modelling self-propelled particles introduces more parameters than passive tracers and each simulation is computationally expensive, an exhaustive exploration of parameter space is not practical. Unless stated otherwise, we therefore

focus on the effect of the maximum propulsive acceleration, $\mathcal{A}_{max} = \max(\mathcal{A}) = \max(\|\boldsymbol{a}_{propel}\|/g) \in [0, 5]$, which characterises the particle's actuation limit. The particle diameter and density are fixed at $d_p = 20\,\mu\text{m}$ and $\rho_p = 3000\,\text{kg m}^{-3}$, respectively; over the Rayleigh-number range considered, the corresponding Stokes number St varies from 5.52×10^{-3} at $Ra = 10^7$ to 1.19×10^{-3} at $Ra = 10^{11}$.

We consider a fixed-displacement task. Throughout training, the self-propelled particle serves as the agent and is trained for a fixed budget of 1×10^6 training steps without early stopping. Its task is to achieve a displacement $\ell = [\boldsymbol{x}(t) - \boldsymbol{x}(0)] \cdot \boldsymbol{n}$ along the direction $\boldsymbol{n}$. For simplicity, we set $\boldsymbol{n} = (1, 0)$ and $\ell = 4$. Because length is non-dimensionalised by the cell height H, this target distance corresponds to traversing one full domain width of $L/H = 4$. An episode is counted as successful as soon as the particle achieves a net horizontal displacement of 4, regardless of its final vertical position. This choice is motivated by the lateral traversal of autonomous vehicles in atmospheric and oceanic environments. Our computational domain is intentionally anisotropic to reflect this setting. Specifically, it is periodic in the horizontal direction but bounded by solid walls in the vertical direction. Consequently, horizontal and vertical motions are inequivalent. The horizontal task allows the particle to traverse the domain continuously, thereby focusing on the challenge of crossing transport barriers (e.g. roll edges) and navigating intermittent cross-flows. An episode terminates unsuccessfully if the particle settles on the bottom wall or if the maximum allowed time, defined as $t^*_{max} = 30$, is reached. As shown in § 3.1, successful traversals complete within $15 \leqslant t^* \leqslant 20$. Thus, $t^*_{max} = 30$ provides a sufficient margin, corresponding to 1.5–2.0 times the typical completion time. Sensitivity tests conducted for representative Rayleigh numbers, namely $Ra = 10^8$ and 10^{10}, in which the time limit was extended to $t^*_{max} = 60$, yielded unchanged success rates. This suggests that failures due to the time limit primarily reflect particles becoming trapped in recirculating vortex cores because of insufficient propulsive thrust, rather than slower trajectories that would otherwise succeed being misclassified as failures.

2.5. *Comparative constant-heading baseline*

To provide a benchmark for propulsion-energy consumption, we construct a constant-heading baseline. This baseline acts as an inverse-dynamics trajectory-tracking controller. Specifically, it uses the completion time t_{RL} achieved by the RL policy to prescribe uniform horizontal progress across the domain. Although this gives the baseline privileged kinematic information, its role is purely comparative. By requiring the baseline to complete the traversal in the same total time as the RL agent, we obtain a controlled comparison of propulsion-energy consumption under an identical kinematic time constraint. The baseline therefore uses the same total duration and prescribes equal horizontal progress during each control interval of length Δt_c. For the target direction $\boldsymbol{n} = (1, 0)$, the dimensional desired displacement over one control interval is

$$\Delta \boldsymbol{x}_{des} = \frac{\ell}{t_{RL}} \Delta t_c \boldsymbol{n} = \left(\frac{\ell \Delta t_c}{t_{RL}}, 0 \right), \tag{2.24}$$

where the desired vertical displacement is zero.

Assuming constant particle acceleration over each control interval Δt_c, the displacement satisfies

$$\Delta \boldsymbol{x} = \boldsymbol{u}_p \Delta t_c + \frac{1}{2} \boldsymbol{a}_p \Delta t_c^2, \tag{2.25}$$

where $\boldsymbol{u}_p$ is the particle velocity at the start of the interval. The corresponding total force is therefore

$$\boldsymbol{F}_{total} = m_p \boldsymbol{a}_p = \frac{2m_p}{(\Delta t_c)^2}\left(\Delta \boldsymbol{x}_{des} - \boldsymbol{u}_p \Delta t_c\right). \tag{2.26}$$

Using the gravitational and drag forces defined above, the required propulsive force is

$$\boldsymbol{F}_{propel}(t) = \boldsymbol{F}_{total}(t) - \boldsymbol{F}_{gravity}(t) - \boldsymbol{F}_{drag}(t). \tag{2.27}$$

To ensure a fair comparison, the baseline agent operates with the same control interval Δt_c as the learned RL policy. The particle applies $\boldsymbol{F}_{propel}$ over one control interval Δt_c, which contains N_{sub} particle-integration substeps of size Δt_p, and we set $\Delta t_c = 30\,\Delta t_p$. For the RL agent, this ratio represents a practical compromise between learning efficiency and physical control responsiveness. For numerical stability, the dimensionless particle-integration time step must remain small (e.g. $\Delta t_p \sim 10^{-3} t^*$). If the agent were to make decisions at every integration step ($\Delta t_c = \Delta t_p$), a single episode would contain $\mathcal{O}(10^4)$ training steps. Such long horizons make the RL credit-assignment process difficult and can impede convergence. Conversely, choosing Δt_c too large would prevent the agent from responding to rapidly evolving transient structures in the convective flow. The choice $\Delta t_c = 30\,\Delta t_p$ gives a complete trajectory involving $\mathcal{O}(10^2)$ decision steps, keeping training computationally tractable while maintaining sufficient control frequency to navigate intermittent turbulence. The baseline agent adopts the same interval in order to match the decision frequency of the RL agent.

Because the turbulent background flow perturbs the actual trajectory, after each interval we update the remaining time and the remaining horizontal displacement, ℓ_{rem}, projected along $\boldsymbol{n}$. The target increment for the next interval is then given by

$$\Delta \boldsymbol{x}_{des} = \frac{\ell_{rem}}{t_{rem}} \Delta t_c \boldsymbol{n}, \tag{2.28}$$

and (2.26)–(2.28) are reapplied until the particle reaches the target or the maximum allowed time is reached.

3. Results and discussion

3.1. *Flow organisation and macroscopic navigation performance*

Figure 1 shows instantaneous temperature contours at several values of Ra. At $Ra = 10^7$ (see figure 1*a*), a four-roll LSC spans the domain. Thin thermal boundary layers develop along the horizontal walls, while the roll interiors remain relatively quiescent. At $Ra = 10^8$ (see figure 1*b*), the roll structure persists, but the number of rolls increases from four to six; see the discussion in Wang *et al.* (2020*b*) and Xu *et al.* (2023). Small-scale instabilities develop, producing wavy shear layers and secondary eddies near the corners and roll edges, and plume detachment becomes more frequent. At $Ra = 10^9$ (see figure 1*c*), the LSC weakens, and the interior contains smaller, less coherent vortices. Plume ejection and impingement become more frequent and exhibit greater spatial variability. At $Ra = 10^{10}$ (see figure 1*d*), additional small-scale vortices emerge, and the bulk temperature field becomes well mixed, although regions of hot-plume ejection and cold-plume impingement remain identifiable. At $Ra = 10^{11}$ (see figure 1*e*), the flow is dominated by small scales and is mixed throughout the bulk. A sustained multi-roll LSC is no longer evident, but plume activity near the horizontal boundaries remains pronounced. Overall, the flow patterns in the $\Gamma = 4$ cell are consistent with those reported for the $\Gamma = 2$ cell by Zhu *et al.* (2018) and for the $\Gamma = 1$ cell by Zhang, Zhou & Sun (2017), He, Bao & Chen (2024) and

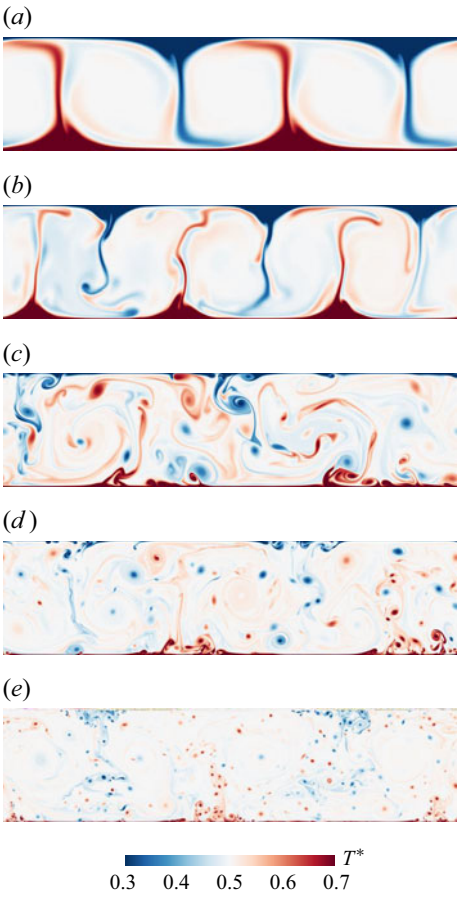


Figure 1. Instantaneous dimensionless temperature field T^* at $Pr = 0.71$ in a cell of aspect ratio $\Gamma = 4$ for (*a*) $Ra = 10^7$, (*b*) $Ra = 10^8$, (*c*) $Ra = 10^9$, (*d*) $Ra = 10^{10}$ and (*e*) $Ra = 10^{11}$. Here, $T^* = (T - T_0)/\Delta_T$, where Δ_T denotes the temperature difference between the heated bottom and cooled top walls.

Gao *et al.* (2024). A detailed comparison of the flow statistics in terms of the Nusselt number is provided in table 1. For reviews of LSC and plume dynamics, see Lohse & Xia (2010), Chillà & Schumacher (2012), Xia *et al.* (2023), Lohse & Shishkina (2023, 2024) and Xia *et al.* (2025).

Figure 2 shows representative trajectories for the fixed-displacement task. The particle starts from the blue square and is required to reach a target displaced by $\ell \boldsymbol{n}$. Here we set $\boldsymbol{n} = (1, 0)$ and $\ell = 4$, so the target, marked by a red star, lies one domain width to the right in a cell of aspect ratio $\Gamma = 4$. Because the lateral boundaries are periodic, a trajectory that exits one lateral boundary re-enters from the opposite side and continues rightward. At $Ra = 10^8$ and $\mathcal{A}_{max} = 1.5$ (see figure 2*a*), the trajectory largely follows the roll edges (Emran & Schumacher 2010). At $Ra = 10^{10}$ and $\mathcal{A}_{max} = 5.0$ (see figure 2*b*), the trajectory becomes more irregular, reflecting enhanced small-scale activity. The particle advances in short, plume-assisted segments and occasionally crosses between rolls. Figures 2(*c*) and 2(*d*) further show ensembles of particle trajectories obtained with the bounded-acceleration policy. Successful attempts are shown in cyan, and unsuccessful attempts in magenta. At $Ra = 10^8$, unsuccessful trajectories are trapped within recirculating cores before reaching the time limit t^*_{max} or hitting the bottom wall, whereas successful

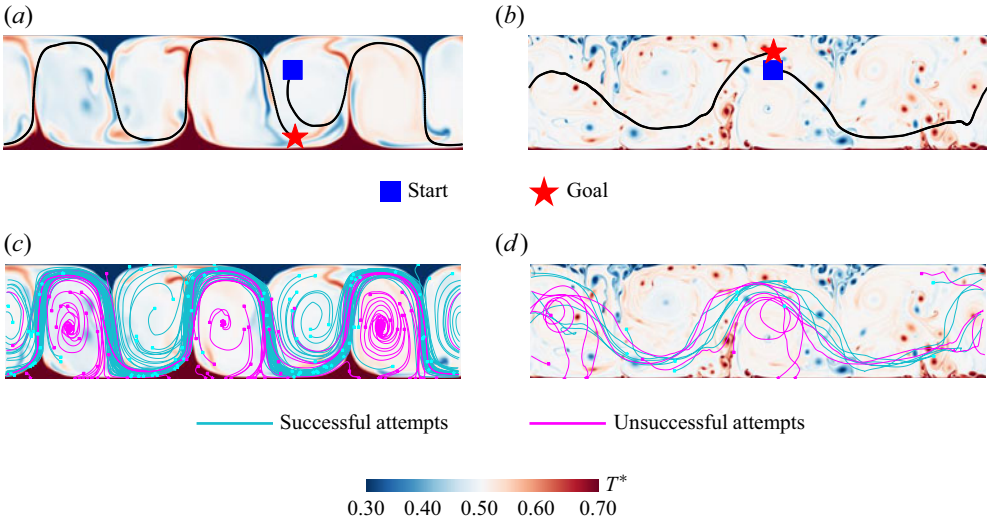


Figure 2. Representative navigation trajectories for the fixed-displacement task. (*a*,*b*) Representative single trajectories and (*c*,*d*) trajectory ensembles, for (*a*) $Ra = 10^8$ with $\mathcal{A}_{max} = 1.5$, (*b*) $Ra = 10^{10}$ with $\mathcal{A}_{max} = 5.0$, (*c*) $Ra = 10^8$ with $\mathcal{A}_{max} = 1.0$ and (*d*) $Ra = 10^{10}$ with $\mathcal{A}_{max} = 3.0$. The background contours indicate the instantaneous dimensionless temperature T^*. The prescribed horizontal displacement is $\ell = 4$, equal to one full domain width ($\Gamma = 4$), while the final vertical position is unconstrained. Because the lateral boundaries are periodic, the particle traverses the domain once. As a result, the start (blue square) and goal (red star) markers have the same horizontal coordinate in the plotted window; any apparent offset is due to the unconstrained vertical drift.

trajectories follow narrow pathways near roll edges. At $Ra = 10^{10}$, confinement weakens and trajectories become more space-filling; several successful examples exploit transient plume-assisted pathways to reach the target. These results suggest that increasing turbulent activity weakens coherent confinement and broadens the available pathways, thereby improving reachability under a fixed actuation limit.

Having described the qualitative differences in particle trajectories across flow regimes, we now quantify the macroscopic navigation performance in terms of reachability, completion time and energetic cost. In the following, the dimensionless maximum propulsive acceleration, $\mathcal{A}_{max}$, is used to characterise the actuation limit of the self-propelled particle. Figure 3(*a*) shows the success probability, estimated empirically as the success rate $S = N_1/N_0$, as a function of $\mathcal{A}_{max}$ for various Ra. Each curve is based on $N_0 = 10^4$ uniformly distributed particle releases under identical flow and boundary conditions, and N_1 denotes the number of trials that complete the fixed-displacement task. The success rate increases monotonically with $\mathcal{A}_{max}$ for all Ra, but the shape of the S curve varies with Ra. At moderate Rayleigh numbers ($Ra = 10^7$ and 10^8), the curves exhibit a sharp transition. Specifically, S remains zero for $\mathcal{A}_{max} \lesssim 1.0$, then rises steeply to about 0.8 over a narrow interval and approaches unity for $\mathcal{A}_{max} \gtrsim 1.5$. This behaviour suggests a finite acceleration threshold required for the particle to cross the boundaries between neighbouring large-scale rolls (Haller 2015; Schneide *et al.* 2019). As Ra increases to 10^9 and 10^{10}, the transition shifts to larger $\mathcal{A}_{max}$ and becomes more gradual. At $Ra = 10^{11}$, the success rate remains low (about 30 %) for $\mathcal{A}_{max} \lesssim 3$ and reaches about 80 % only when $\mathcal{A}_{max} \gtrsim 4$. The rightward shift and progressive smoothing of the transition with increasing Ra are consistent with the weakening of coherent roll barriers and the strengthening of intermittent motions, which require stronger and more sustained propulsive acceleration to keep trajectories aligned with the target direction.

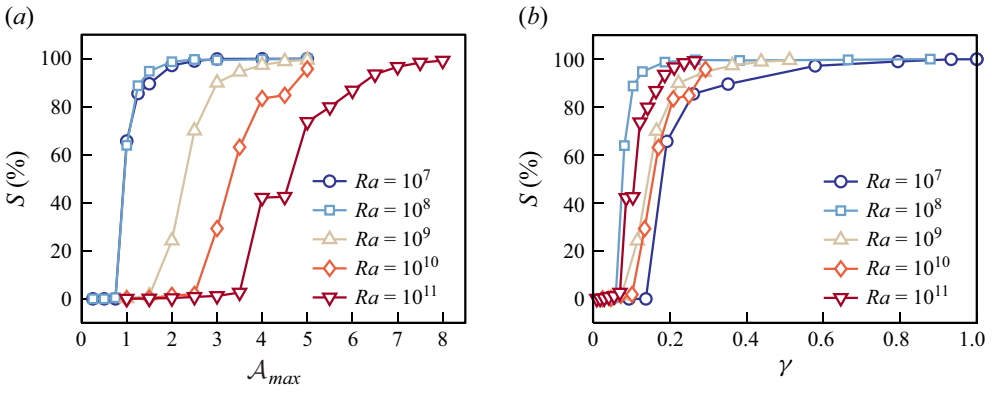


Figure 3. Success rate $S = N_1/N_0$ as a function of (*a*) the maximum dimensionless propulsive acceleration, $\mathcal{A}_{max} = \max(\|\boldsymbol{a}_{propel}\|/g)$, and (*b*) the navigable-area fraction, γ, for Rayleigh numbers $10^7 \leq Ra \leq 10^{11}$. Here, γ denotes the spatiotemporal mean fraction of the flow domain in which the particle's active terminal velocity $U^*_{t,max}$ exceeds the local fluid speed $\|\boldsymbol{u}^*_f\|$.

To examine the relation between the rightward shift and the increasing flow intensity at higher Ra, we compare the agent's kinematic capability with the Eulerian flow statistics. We define the particle's dimensionless active terminal velocity in a quiescent fluid, determined by the balance between horizontal propulsion and Stokes drag, as $U^*_{t,max} = St\mathcal{A}_{max}\Lambda$. We then calculate the navigable-area fraction γ, defined as the spatiotemporal average fraction of the domain in which the particle's terminal speed exceeds the local fluid speed, i.e. $U^*_{t,max} > \|\boldsymbol{u}^*_f\|$. As shown in figure 3(*b*), plotting the success rate against the navigable-area fraction γ largely collapses the curves for all Rayleigh numbers. The success rate exhibits a sharp increase over a narrow interval, $0.05 \leq \gamma \leq 0.20$. This result suggests that the higher actuation required at large Ra is predominantly associated with the shrinking fraction of kinematically navigable regions. Thus, γ provides a useful reduced parameter for organising the baseline reachability data within the present convective-flow setting. However, γ alone does not fully determine the success probability, because it measures local kinematic accessibility but does not encode the finite-time connectivity of the navigable regions. Different flows, or different time windows, with similar γ may therefore yield different success probabilities if their LCS connectivity differs. The slight residual spread among the collapsed curves in figure 3(*b*) is likely related to this effect. In particular, as Ra increases, coherent roll barriers become increasingly fragmented and intermittent plume-assisted pathways emerge, modifying the transport connectivity even at comparable values of γ.

We next examine the dimensionless completion time t^*_{comp} for the fixed-displacement task. Figure 4(*a*) shows that t^*_{comp} decreases monotonically with the maximum propulsive acceleration $\mathcal{A}_{max}$ for all Rayleigh numbers. The strongest reduction occurs for $1 \lesssim \mathcal{A}_{max} \lesssim 3$; beyond this range, the curves gradually flatten, indicating diminishing time savings as the acceleration is increased further. This behaviour is consistent with the presence of large-strain boundaries and plume channels that act as finite-time transport barriers. Once the applied acceleration is sufficient to cross these barriers, further increases in $\mathcal{A}_{max}$ yield only modest gains (Haller 2015; Schneide *et al.* 2019). Figure 4(*b*) provides a complementary perspective. For each actuation level, t^*_{comp} increases approximately monotonically with Ra. The separation between the curves widens with Ra, but their ordering with respect to $\mathcal{A}_{max}$ remains unchanged. Larger $\mathcal{A}_{max}$ consistently corresponds to shorter t^*_{comp} at a given Ra.

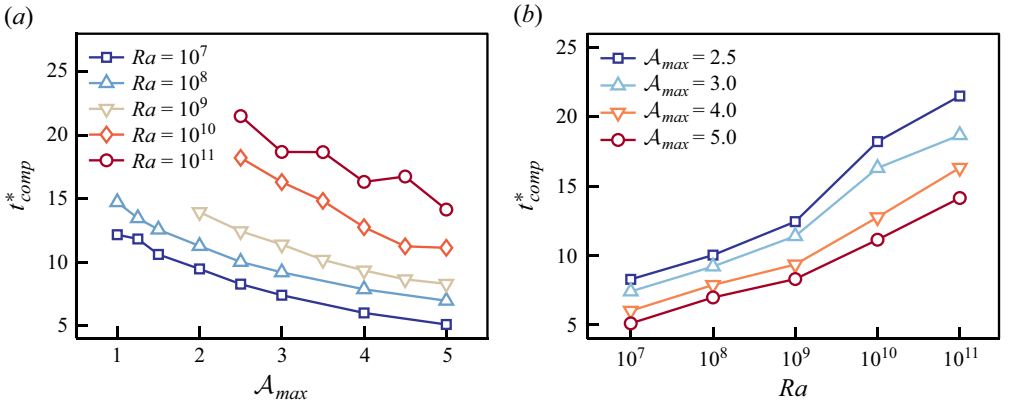


Figure 4. Dimensionless completion time t^*_{comp} for the fixed-displacement task as a function of (*a*) the maximum dimensionless propulsive acceleration $\mathcal{A}_{max}$ for different Ra and (*b*) the Rayleigh number Ra for different $\mathcal{A}_{max}$.

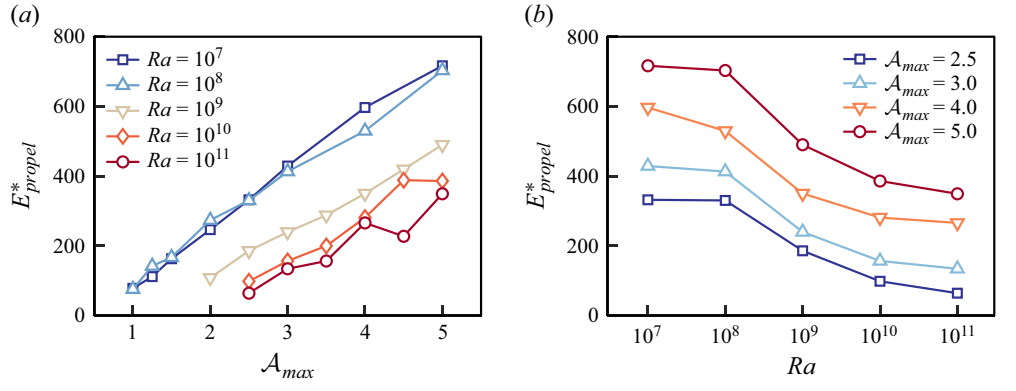


Figure 5. Dimensionless propulsion energy E^*_{propel} for the fixed-displacement task as a function of (*a*) the maximum dimensionless propulsive acceleration $\mathcal{A}_{max}$ for different Ra and (*b*) the Rayleigh number Ra for different $\mathcal{A}_{max}$.

We next examine the dimensionless propulsion energy E^*_{propel} for the fixed-displacement task. To enable a meaningful comparison across parameter values, the dimensional energy $E_{propel} = \int \boldsymbol{F}_{propel} \cdot \mathrm{d}\boldsymbol{s}$ is non-dimensionalised by the reference mechanical work $m_p g \beta \Delta_T H$, giving

$$E^*_{propel} = \frac{E_{propel}}{m_p g \beta \Delta_T H} = \int \boldsymbol{a}^*_{propel} \cdot \mathrm{d}\boldsymbol{s}^*, \tag{3.1}$$

where $\boldsymbol{s}^* = \boldsymbol{s}/H$ is the dimensionless path coordinate. The integral is evaluated only for successful trajectories and then averaged over them. Figure 5(*a*) plots E^*_{propel} versus the maximum dimensionless propulsive acceleration $\mathcal{A}_{max}$ for different Ra. For all Ra, E^*_{propel} increases with $\mathcal{A}_{max}$. At moderate $Ra = 10^7$ and 10^8, the increase is nearly linear and the curves have similar slopes. This linear scaling is consistent with traversal across successive coherent LSC cells. In this regime, each increment in actuation produces a comparable increment in mechanical work, as also reported for active-particle transport in laminar or weakly chaotic flows (Piro *et al.* 2024). As Ra rises to 10^9 and above, the curves

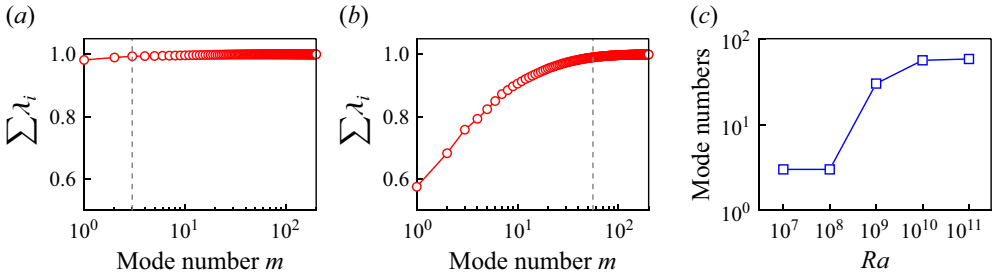


Figure 6. Normalised cumulative modal energy $\sum_{i=1}^{m}\lambda_i$ as a function of mode number m for (*a*) $Ra = 10^8$ and (*b*) $Ra = 10^{10}$. The vertical dashed lines indicate the number of modes required to capture 99 % of the total fluctuating kinetic energy. (*c*) The required number of modes as a function of Ra.

bend downward and shift to lower values. For a given $\mathcal{A}_{max}$, the required propulsion energy is smaller and less sensitive to $\mathcal{A}_{max}$, suggesting a larger contribution from background plume motions and small-scale intermittency. Similar reductions in propulsion cost with increasing environmental fluctuations have been reported for microswimmers in turbulence and active suspensions (Yang *et al.* 2019). Figure 5(*b*) shows E^*_{propel} versus Ra at fixed $\mathcal{A}_{max}$. For each $\mathcal{A}_{max}$, E^*_{propel} generally decreases with Ra, with only mild non-monotonic variations. The narrowing of the spacing between curves at high Ra suggests that the dependence of E^*_{propel} on $\mathcal{A}_{max}$ weakens; further increases in $\mathcal{A}_{max}$ then produce only small changes in E^*_{propel}. Taken together with the success-rate trends, these results suggest that more vigorous flows at higher Ra require larger $\mathcal{A}_{max}$ to achieve reachability, but that once reachability is attained, the energetic cost of traversal decreases. This trade-off between flow-assisted transport and actuation cost is consistent with observations for active agents in complex flows (Ishikawa 2025).

3.2. *Flow coherence and policy generalisability*

To further examine the Ra dependence of the success-rate curves, we analyse the organisation of the carrier flow. Specifically, we apply POD to extract the dominant coherent structures. The POD method has been widely used to study LSC dynamics in convection cells (Podvin & Sergent 2015; Castillo-Castellanos *et al.* 2019; Soucasse *et al.* 2019). The spatiotemporal velocity field $\boldsymbol{u}(\boldsymbol{x}, t)$ is expressed as a superposition of empirical orthogonal eigenfunctions $\boldsymbol{\phi}_i(\boldsymbol{x})$ with time-dependent amplitudes $a_i(t)$:

$$\boldsymbol{u}(\boldsymbol{x}, t) = \sum_{i=1}^{\infty} a_i(t)\,\boldsymbol{\phi}_i(\boldsymbol{x}). \tag{3.2}$$

Here, $\boldsymbol{u}(\boldsymbol{x}, t) = [u(\boldsymbol{x}, t),\ v(\boldsymbol{x}, t)]^T$ is the two-dimensional velocity field, $\boldsymbol{\phi}_i(\boldsymbol{x}) = [\phi_i^u(\boldsymbol{x}),\ \phi_i^v(\boldsymbol{x})]^T$ are the spatial eigenfunctions (the POD modes) and $a_i(t)$ are the temporal coefficients. Each orthonormal spatial mode $\boldsymbol{\phi}_i(\boldsymbol{x})$ has an eigenvalue λ_i representing the fraction of fluctuation energy captured by mode i. The eigenvalues are normalised such that $\sum_i \lambda_i = 1$. We report the cumulative energy $\sum_{i=1}^{m}\lambda_i$ as a function of the number of retained modes m, which indicates how many modes are required to represent the dominant dynamics. Rapid saturation of this curve indicates strong flow coherence, whereas slow saturation implies that the energy is distributed over many modes.

At moderate $Ra = 10^8$ (see figure 6*a*), only a few modes capture most of the energy, consistent with the presence of coherent multiple-roll LSCs. The edges of adjacent rolls, together with the plume-ejection and plume-impingement zones, act as transport barriers.

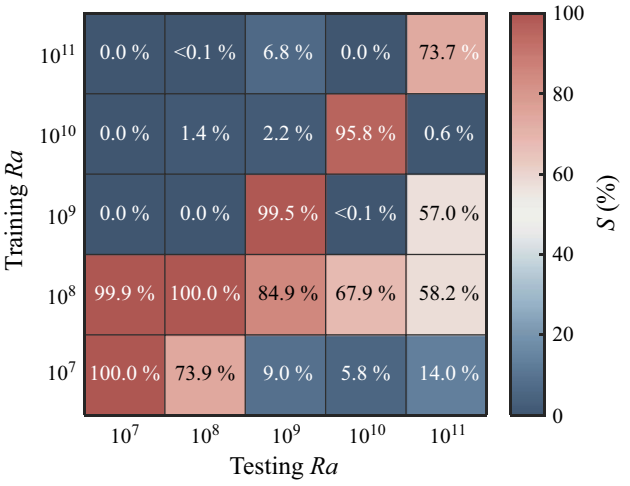


Figure 7. Generalisability of the learned policies across Rayleigh numbers. The heatmap shows the success rate S of policies trained at a given Ra (rows) and evaluated at different testing Ra (columns). All cases are evaluated at $\mathcal{A}_{max} = 5.0$.

Crossing these barriers requires additional propulsive acceleration, which is consistent with the sharp onset of success once $\mathcal{A}_{max}$ exceeds a threshold value. At higher $Ra = 10^{10}$ (see figure 6*b*), the energy is distributed over more modes and the coherence weakens. The large-scale barriers become more fragmented and drift in time, forming short-lived pathways. As a result, $S(\mathcal{A}_{max})$ increases more gradually, and its onset shifts to higher $\mathcal{A}_{max}$. Figure 6(*c*) further summarises the number of modes required to capture 99 % of the total energy as a function of Ra. These trends suggest that the success-rate curves evolve from step-like to smoother as the flow changes from a low-order, barrier-dominated regime to a more intermittent regime at higher Ra.

To assess the robustness and generalisability of the learned strategies across different turbulence intensities, we conducted a cross-evaluation study. Policies trained at one Rayleigh number were then deployed in environments at other Rayleigh numbers. All evaluations were performed with $\mathcal{A}_{max} = 5.0$. The resulting success rates are shown in the cross-evaluation heatmap in figure 7, where the diagonal elements represent the baseline success rates when training and testing are performed in the same flow environment. Deploying a policy across full orders of magnitude in Ra represents a stringent regime shift. When evaluated under more moderate environmental variations closer to the training domain, the learned policies transfer well without any retraining. For example, a policy trained at $Ra = 10^8$ achieves a 100.0 % success rate when deployed at $Ra = 1.5 \times 10^8$, and a policy trained at $Ra = 10^9$ maintains a 99.5 % success rate when evaluated at $Ra = 1.5 \times 10^9$. A notable feature in figure 7 is the asymmetry in generalisability across large Ra gaps. Policies trained at lower Ra, for example $Ra = 10^8$, transfer reasonably well to higher-Ra flows, achieving a success rate of 67.9 % at $Ra = 10^{10}$. By contrast, policies trained at higher Ra show much lower success rates when deployed at lower Ra; for example, training at $Ra = 10^{10}$ yields a success rate of only 1.4 % at $Ra = 10^8$. Because the same hyperparameter set was used across all training cases, this asymmetry does not appear to be primarily caused by case-specific hyperparameter tuning. Rather, it is consistent with changes in the underlying flow topology. This interpretation is also consistent with recent findings in active navigation (Jiao *et al.* 2025), where policies trained at low Reynolds numbers generalise to high Reynolds numbers more effectively

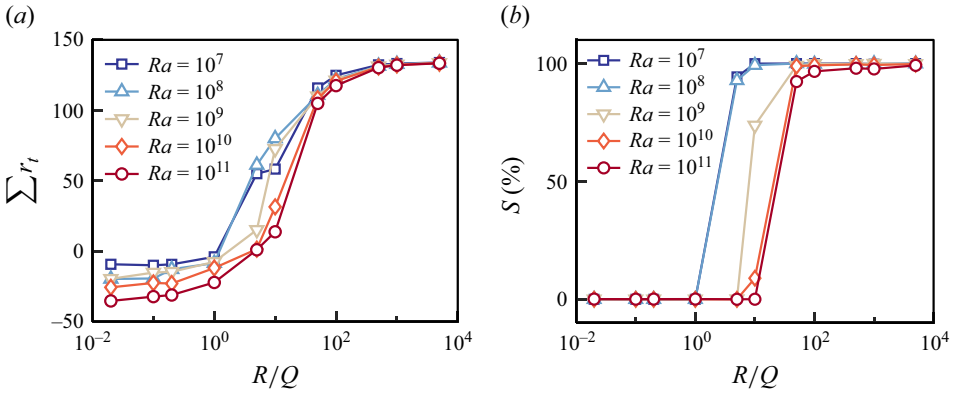


Figure 8. (*a*) Average cumulative reward $\langle \sum r_t \rangle$ and (*b*) success rate S as functions of the reward-weighting ratio R/Q for different Rayleigh numbers Ra.

than in the reverse direction. At moderate Ra, the flow is dominated by coherent LSC rolls and persistent transport barriers, so the policy tends to learn a relatively direct and thrust-intensive strategy for crossing them. When transferred to a higher-Ra environment, this barrier-crossing logic can remain effective because the corresponding barriers are more fragmented and transient. Conversely, a policy trained at high Ra learns a more opportunistic strategy, mainly exploiting transient openings and plume-assisted pathways to reduce propulsion. When placed in a moderate-Ra flow, such transient openings are less common, and the policy does not appear to have learned a sufficiently robust response for crossing the more persistent transport boundaries. This leads to a substantially lower success rate.

3.3. *Energy–time trade-offs in the reward formulation*

The reward function balances time-based rewards, represented by the term $RV^*_{eff}(t)$, and energy-based penalties, represented by $-Q\|\boldsymbol{a}^*_{propel}(t)\|$. The ratio R/Q sets this trade-off and therefore influences both the temporal efficiency and the energetic cost of the self-propelled particle. When R/Q is small, the propulsion penalty dominates; in the limit $R/Q \to 0$, the reward reduces to $r_t = -\|\boldsymbol{a}^*_{propel}(t)\|$. When R/Q is large, the reward for forward progress dominates; in the limit $R/Q \to \infty$, the reward reduces to $r_t = V^*_{eff}(t)$. This formulation reflects the classical compromise between energy expenditure and travel time in locomotion problems. The hydrodynamics of low-Reynolds-number swimmers naturally involves a balance between mechanical power input and translational speed (Lauga & Powers 2009), and similar energy–time trade-offs have been discussed for active particles operating under finite energy budgets (Bechinger *et al.* 2016). Reinforcement-learning control frameworks likewise often use composite reward functions that encourage rapid yet energy-efficient motion, as demonstrated for numerical swimmers and flow-navigation tasks (Reddy *et al.* 2016; Colabrese *et al.* 2017; Verma, Novati & Koumoutsakos 2018). In the present context, the ratio R/Q plays an analogous role by determining whether the agent prioritises rapid progress or economical propulsion.

Figure 8(*a*) shows how the reward weighting R/Q influences the learned policy. The averaged cumulative reward $\langle \sum r_t \rangle$ increases with R/Q. When $R/Q \lesssim 10$, the penalty term dominates, and the return remains small or negative for all Ra. As R/Q increases to $\mathcal{O}(10)$–$\mathcal{O}(100)$, the return rises steeply before reaching a plateau, indicating convergence towards policies that prioritise rapid progress. The success rate in figure 8(*b*) exhibits a

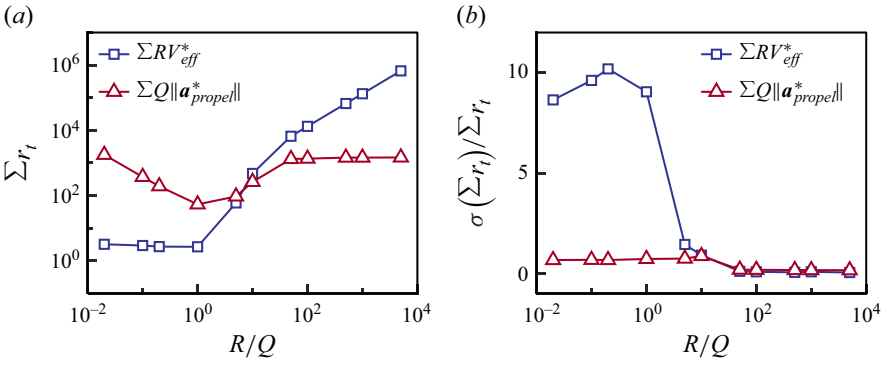


Figure 9. (*a*) Average cumulative reward components and (*b*) their relative variations, $\sigma(\sum r_t)/\langle\sum r_t\rangle$. The curves show how the total reward decomposes into the time-efficiency term $\sum RV^*_{eff}$ (squares) and the actuation-penalty term $\sum Q\|\boldsymbol{a}^*_{propel}\|$ (triangles), as functions of the reward-weighting ratio R/Q at $Ra = 10^9$.

similar transition. Success is negligible at small R/Q, increases rapidly once a threshold ratio $(R/Q)_c$ is exceeded and then saturates slightly below unity at $Ra = 10^{11}$. As Ra increases, the success rate at fixed R/Q decreases and the threshold value $(R/Q)_c$ shifts to larger values. This trend suggests that stronger emphasis on forward progress is required in more vigorous convective flows. Because the energy-based penalty discourages large control effort, a larger R/Q is required at high Ra to favour forward progress and enable barrier crossing. For example, at $R/Q = 10$, the success rate is about 70 % for $Ra = 10^9$, but drops to about 10 % at $Ra = 10^{10}$.

To clarify the competition between the time reward and the actuation penalty, figure 9 decomposes the averaged reward at $Ra = 10^9$ into two unnormalised cumulative contributions, namely the time-reward term $\sum RV^*_{eff}$ and the actuation-penalty magnitude $\sum Q\|\boldsymbol{a}^*_{propel}\|$. In figure 9(*a*), when $R/Q \lesssim 10$, the penalty contribution dominates, consistent with the near-zero success rate observed in figure 8(*b*). Thus, a small R/Q penalises propulsion, discourages energetically costly actions and can hinder the exploitation of transport pathways in convective flows. As R/Q increases to $\mathcal{O}(10)$–$\mathcal{O}(100)$, the term $\sum RV^*_{eff}$ grows rapidly, whereas $\sum Q\|\boldsymbol{a}^*_{propel}\|$ remains nearly constant. The improvement in success therefore arises mainly from an increase in the time-reward contribution rather than from a reduction in the actuation penalty. Figure 9(*b*) further shows the relative variation of the total reward, $\sigma(\sum r_t)/\langle\sum r_t\rangle$. The variation of $\sum RV^*_{eff}$ is large at small R/Q but drops sharply once R/Q exceeds $\mathcal{O}(10)$, whereas the variation of $\sum Q\|\boldsymbol{a}^*_{propel}\|$ remains small and depends only weakly on R/Q. These trends suggest that higher success rates are associated with more stable policies, because large reward variability reflects a highly stochastic policy that often fails to complete the task. Overall, this multi-objective optimisation problem exhibits a threshold weighting $(R/Q)_c$ above which training yields consistently successful policies. This threshold marks the point at which optimising forward progress begins to outweigh minimising actuation expenditure. Beyond it, the policy stabilises and both reward components vary only weakly. Accordingly, in the primary tests we use a sufficiently large value, namely $R/Q = 5000$.

To clarify the limiting behaviours implied by the reward decomposition, figure 10 compares trajectories in the two limiting cases: time reward only ($R/Q \to \infty$) and energy penalty only ($R/Q \to 0$). In figures 10(*a*) and 10(*b*), the particle prioritises rapid progress, maintains a higher instantaneous speed (as indicated by the trajectory colouring) and follows plume-assisted pathways that cut through roll boundaries and

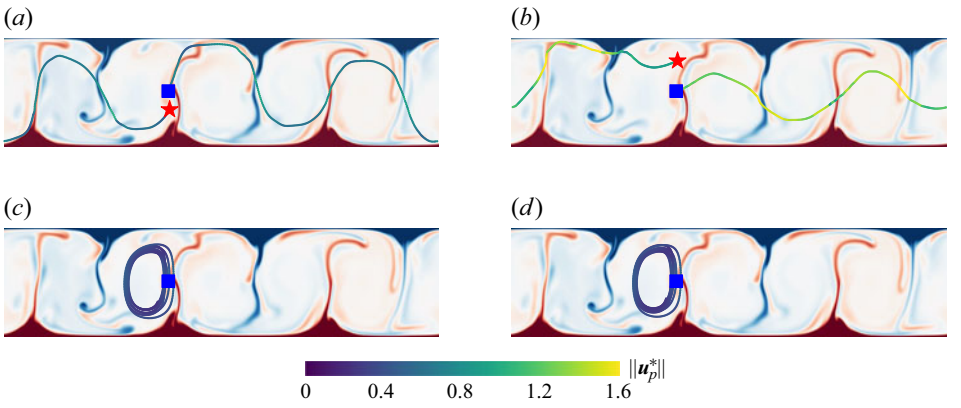


Figure 10. Trajectories illustrating the limiting behaviours of the reward formulation for two actuation bounds. (*a*,*b*) The time-reward-only limit ($R/Q \to \infty$) and (*c*,*d*) the energy-penalty-only limit ($R/Q \to 0$), for (*a*,*c*) $\mathcal{A}_{max} = 2.5$ and (*b*,*d*) $\mathcal{A}_{max} = 10$. The background contours indicate the instantaneous dimensionless temperature T^*, and the colour along each trajectory indicates the dimensionless particle speed $\|\boldsymbol{u}_p^*\|$.

near-wall shear layers. With $\mathcal{A}_{max} = 2.5$, the trajectory already threads transient openings in the LSC. Increasing the bound to $\mathcal{A}_{max} = 10$ mainly straightens the trajectory and reduces residence near stagnation regions, thereby shortening the traversal time without altering the overall strategy. While our principal quantitative analyses focus on $\mathcal{A}_{max} \in [0, 5]$ to study flow-assisted transport, the larger bound $\mathcal{A}_{max} = 10$ is used here only for illustration. It demonstrates the limiting behaviour when the control amplitude is sufficiently large, suggesting that the main features of these reward-driven strategies persist even when the agent can readily overpower the local flow. Figures 10(*c*) and 10(*d*) show the complementary condition in which only the energy penalty is applied. In this case, the RL-controlled self-propelled particle minimises its propulsive acceleration, aligns with the local circulation of the large-scale roll and is advected onto nearly closed orbits that avoid plume-ejection and plume-impingement regions. The particle repeatedly recirculates within a single convection cell, lingering near separatrices and stagnation saddles rather than using propulsive effort to cross them. Even when the actuation bound is increased from $\mathcal{A}_{max} = 2.5$ (see figure 10*c*) to $\mathcal{A}_{max} = 10$ (see figure 10*d*), the motion remains confined to the roll and the prescribed horizontal displacement is not completed within the allowed time. These examples show that, when rapid progress is not encouraged, energy-saving behaviour leads to attachment to the LSC and poor reachability, whereas time-oriented control exploits intermittent transport pathways to cross barriers. A higher $\mathcal{A}_{max}$ then acts mainly to accelerate crossing rather than to change the underlying navigation logic.

3.4. *Navigation mechanisms: kinematics and Lagrangian coherent structures*

Figure 11 compares the learned RL policy with a constant-heading baseline for the fixed-displacement task at $Ra = 10^{10}$ and $\mathcal{A}_{max} = 5.0$. Details of the baseline configuration are given in § 2.5. The RL agent follows plume-assisted pathways near roll edges and avoids recirculation zones (see figure 11*a*), whereas the baseline agent maintains a constant heading towards the target and repeatedly encounters adverse cross-stream currents (see figure 11*b*). To aid physical interpretation, side-by-side visualisations of these distinct navigation strategies at moderate and high Rayleigh numbers are provided

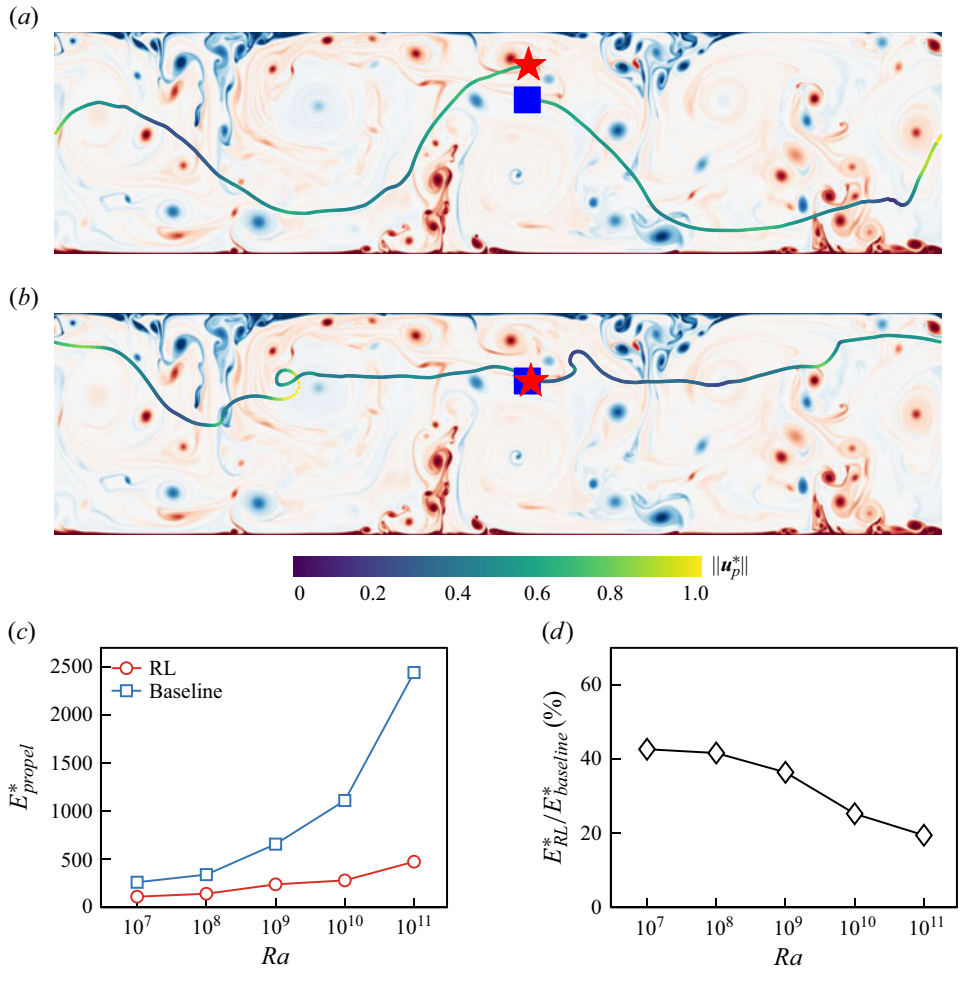


Figure 11. Comparison of the learned RL policy with the constant-heading baseline for the fixed-displacement task. Trajectories obtained with (*a*) the RL policy and (*b*) the constant-heading baseline at $Ra = 10^{10}$ and $\mathcal{A}_{max} = 5.0$. The background contours indicate the instantaneous dimensionless temperature T^*, and the colour along each trajectory indicates the dimensionless particle speed $\|\boldsymbol{u}_p^*\|$. As in figure 2, the start marker (blue square) and goal marker (red star) have the same horizontal coordinate because of the periodic lateral boundaries; in (*b*), they overlap visually because the particle returns to its initial vertical position on completing the task. (*c*) The dimensionless propulsion energy E^*_{propel} versus Ra for both strategies. (*d*) The relative propulsion cost $E^*_{RL}/E^*_{baseline}$ as a function of Ra.

in supplementary movie 1 ($Ra = 10^8$) and supplementary movie 2 ($Ra = 10^{10}$) available at https://doi.org/10.1017/jfm.2026.11802.

These trajectory differences are reflected in the propulsion energetics (see figure 11*c*). For a fair comparison, we define a reference actuation level $\mathcal{A}_{max}$ for each Ra as the value that yields a success rate of 80 % in the convective flow. A sensitivity analysis, not shown for brevity, indicates that this energetic advantage remains evident under a stricter success criterion, for example 90 %. The propulsion energy E^*_{propel} of the RL policy increases only slightly with Ra, whereas that of the constant-heading baseline rises sharply. Consequently, the ratio $E^*_{RL}/E^*_{baseline}$ decreases from about 45 % at $Ra = 10^7$ to 17 % at $Ra = 10^{11}$ (see figure 11*d*), suggesting that the relative advantage of the learned

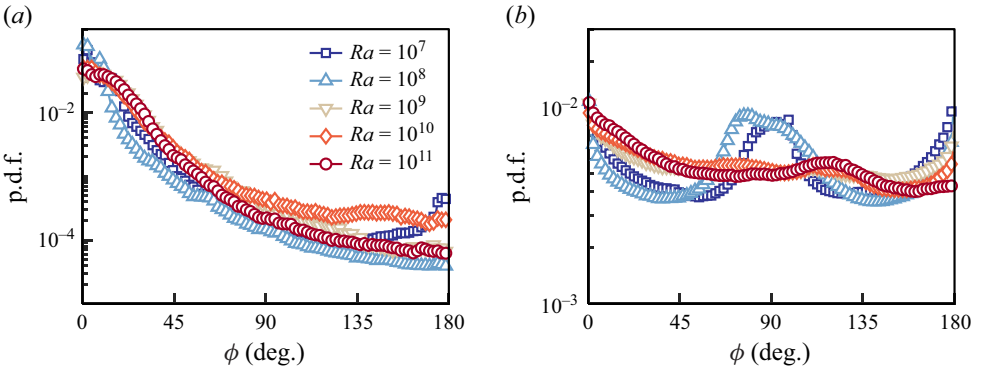


Figure 12. The p.d.f.s of the instantaneous alignment angle ϕ between the particle's propulsive vector and the local fluid-velocity vector. The data are computed over all successful trajectories obtained with (*a*) the learned RL policy and (*b*) the constant-heading baseline for different Rayleigh numbers Ra.

policy becomes more pronounced with increasing Ra. This occurs because the learned policy preferentially aligns the particle with favourable flow structures, allowing part of the mechanical work to be supplied by the background flow through drag. By contrast, the constant-heading baseline encounters strong counterflows more frequently as Ra increases, so the opposing streamwise velocity and enhanced shear lead to substantially higher propulsion energy.

To further compare propulsive behaviour during navigation, we examine the probability density functions (p.d.f.s) of the instantaneous alignment angle ϕ between the particle's propulsive vector and the local fluid-velocity vector, computed over all successful trials. A value of $\phi = 0°$ indicates alignment with the carrier flow, whereas $\phi = 180°$ indicates opposition. For the learned strategy, the p.d.f. exhibits a pronounced peak near 0° and then decreases monotonically (see figure 12*a*). This pattern suggests that the RL agent tends to align its propulsion direction with the local current and thereby exploits the carrier flow to progress towards the target. The tail broadens slightly as the Rayleigh number increases, consistent with increasingly intermittent reorientation of the local flow. The constant-heading baseline yields a comparatively flat distribution with a small secondary peak near 90° (see figure 12*b*), reflecting frequent cross-stream events. The baseline agent therefore often accelerates nearly orthogonally to the carrier flow rather than following it, which leads to higher control effort and is consistent with the larger propulsive-energy expenditure reported in figure 11.

To better understand the physical origin of this energetic advantage, we move from an Eulerian kinematic description to a finite-time Lagrangian view of transport barriers and pathways. In unsteady flows, such structures can be characterised using LCS, which are identified here as ridges of the finite-time Lyapunov exponent (FTLE) fields (Krishna *et al.* 2022, 2023). To relate the learned trajectories to these structures, we compute both forward-time and backward-time FTLE fields for the convective flow. For visualisation, the LCS fields are extracted by thresholding the FTLE fields: regions exceeding 70 % of the maximum forward- and backward-time FTLE values are classified as repelling and attracting LCS, respectively. As shown in figure 13, at moderate Rayleigh number ($Ra = 10^8$), the repelling LCS, which act as finite-time transport barriers, and the attracting LCS, which indicate preferred transport pathways, show substantial spatial overlap. Together they outline closed or nearly closed boundaries around the LSC cells. This topological organisation is consistent with the trapping observed for the energy-penalty-only policy: because this policy applies little propulsive thrust, it tends to remain

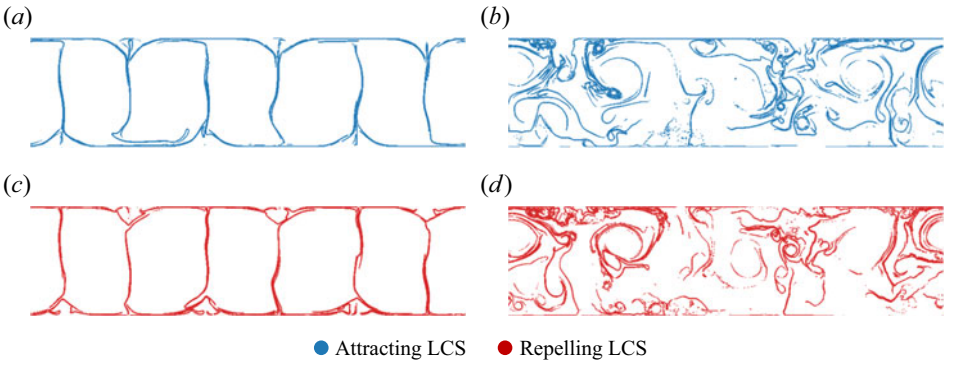


Figure 13. Instantaneous snapshots of the LCS at $t^* = 9.00$ for (*a*,*c*) $Ra = 10^8$ and (*b*,*d*) $Ra = 10^{10}$. (*a*,*b*) The attracting LCS (blue regions). (*c*,*d*) The repelling LCS (red regions).

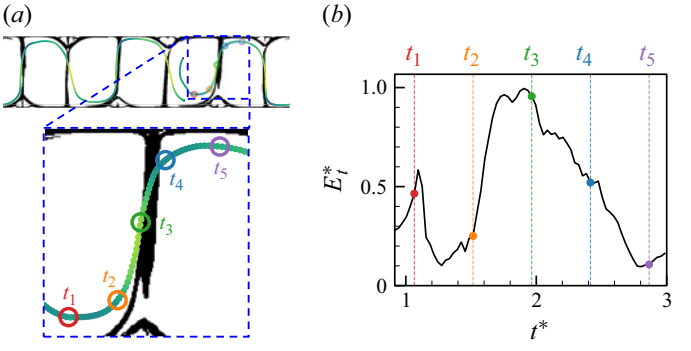


Figure 14. (*a*) Representative trajectory of the RL agent at $Ra = 10^8$ with $\mathcal{A}_{max} = 1.25$, corresponding to the learned policy used for this regime. The upper panel shows the repelling LCS over the full domain, whereas the lower panel provides a magnified view of the dashed region and highlights a barrier-crossing event. Coloured markers indicate five representative instants, t_1–t_5, along the trajectory. (*b*) Corresponding time history of the dimensionless instantaneous propulsion cost E_t^* as a function of dimensionless time t^*.

within the same circulation cell rather than crossing the repelling FTLE ridges. At higher Rayleigh number ($Ra = 10^{10}$), this overlap weakens, and the repelling LCS no longer form comparably closed structures. The associated weakening and fragmentation of the barrier network provide more intermittent pathways through which the agent can traverse the domain.

Furthermore, because crossing an LCS is a dynamic and continuous process, comparing a trajectory with a single instantaneous FTLE snapshot is insufficient. To analyse the barrier-crossing events, we therefore compute a time-averaged FTLE field over the corresponding crossing interval, for example from t_1 to t_5. Applying a threshold of 60 % of the maximum value to this time-averaged field is used to extract the persistent LCS. Figures 14 and 15 show a representative learned particle trajectory overlaid on the corresponding time-averaged FTLE ridge during a barrier-crossing event. The corresponding instantaneous propulsion cost E_t^* is plotted alongside. As the particle approaches and actively crosses the repelling FTLE ridge, i.e. the barrier, between t_2 and t_4, the propulsion cost exhibits a pronounced local increase. This suggests that the agent applies large propulsive effort near this topological bottleneck. Immediately after crossing the barrier, the particle aligns with the attracting LCS, i.e. the transport pathway, and

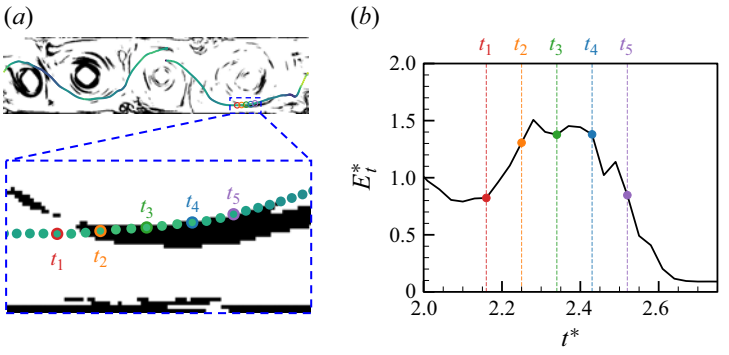


Figure 15. (*a*) Representative trajectory of the RL agent at $Ra = 10^{10}$ with $\mathcal{A}_{max} = 4.0$, corresponding to the learned policy used for this regime. The upper panel shows the repelling LCS over the full domain, whereas the lower panel provides a magnified view of the dashed region and highlights a typical event in which the trajectory crosses a repelling LCS. Coloured markers indicate five representative instants, t_1–t_5, along the trajectory. (*b*) Corresponding time history of the dimensionless instantaneous propulsion cost E_t^* as a function of dimensionless time t^*.

its propulsion cost decreases as it follows favourable downstream currents. These results suggest that the RL policy exploits aspects of the underlying Lagrangian topology of the turbulent convective flow. We acknowledge, however, that the representative trajectory overlays and time-averaged FTLE fields, although supportive of the barrier-crossing interpretation, do not yet constitute a full event-level statistical census of all crossings. Such a quantitative analysis would require automated identification of individual trajectory intersections with repelling FTLE ridges, followed by correlation with local propulsion-cost peaks. This is non-trivial in the high-Ra regimes, because the LCS geometry is highly transient and the topological extraction can be sensitive to the chosen finite-time integration window. We therefore regard a rigorous automated event-level analysis of LCS crossings as an important direction for future research. Furthermore, although FTLE analysis provides a mathematically well-defined description of finite-time transport barriers, its computation requires non-local, time-resolved flow information over a finite integration window. By contrast, the RL agent navigates these dynamic barriers using only instantaneous, locally measurable cues. To understand how the policy achieves this, we next translate these global Lagrangian structures into local Eulerian observables.

3.5. *Distilling an interpretable heuristic strategy from Eulerian flow topology*

Although the RL policy achieves lower propulsion cost than the constant-heading baseline, the physical logic underlying this black-box optimisation still warrants clearer interpretation. To elucidate how the agent exploits the observed flow information, we analyse the particle behaviour relative to the local flow structures, focusing on representative cases at $Ra = 10^8$ and $Ra = 10^{10}$ under their respective reference actuation bounds.

Our analysis shows that the learned policy exhibits a consistent behavioural pattern. To highlight the agent's spatial preferences, we compare the distribution of active navigators with that of passive tracers released under identical conditions. The instantaneous Voronoi tessellations in figure 16(*a–d*) show that the passive tracers (figures 16*a* and 16*b*) are broadly dispersed by the background turbulence and often become trapped within recirculating vortex cores. By contrast, the active navigators (figures 16*c* and 16*d*) preferentially avoid these cores and instead accumulate near the outer regions of vortical

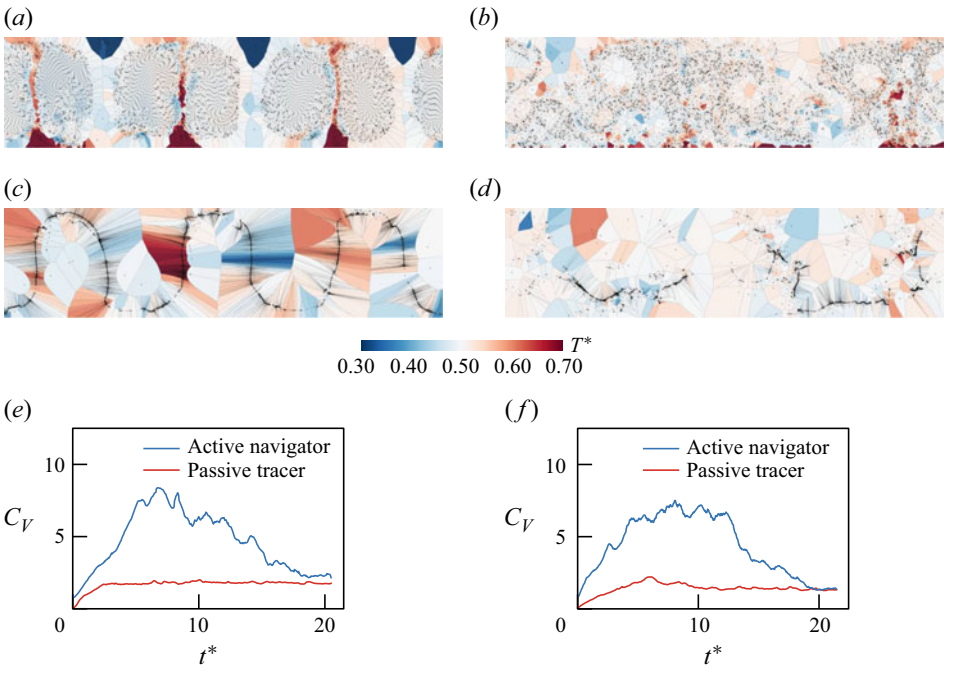


Figure 16. Instantaneous Voronoi tessellations of the distributions of (*a*,*b*) passive tracers and (*c*,*d*) active navigators at $t^* = 9.00$. (*e*,*f*) The corresponding temporal evolution of the clustering index C_V for the reference actuation conditions. The Voronoi cells are coloured by the instantaneous background temperature field T^*. (*a*,*c*,*e*) Rayleigh number $Ra = 10^8$ with $\mathcal{A}_{max} = 1.25$ and (*b*,*d*,*f*) $Ra = 10^{10}$ with $\mathcal{A}_{max} = 4.0$.

structures and along roll edges, where pathways for horizontal traversal are more favourable. To quantify this preferential clustering, we introduce a clustering index C_V, defined as the normalised standard deviation of the Voronoi-cell areas:

$$C_V = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{A_i}{\bar{A}} - 1\right)^2}, \tag{3.3}$$

where A_i denotes the area of the ith Voronoi cell, $\bar{A}$ is the mean Voronoi-cell area and M is the total number of particles. As shown in figures 16(*e*) and 16(*f*), C_V for the active navigators remains consistently higher than for the passive tracers throughout the migration process. This larger variance in cell areas suggests that the learned policy steers particles away from broad, relatively featureless flow regions, leaving a few large Voronoi cells, while concentrating them within narrow and dynamically favourable pathways, thereby producing dense clusters of small cells. In this way, rather than being passively mixed by the flow, the active navigators exploit the spatial heterogeneity of the turbulence to their advantage.

To characterise the nature of these favourable pathways, we examine the coupling between particle distribution and local flow topology by plotting the joint distribution of the Voronoi-cell area A and the local Q-value, as shown in figure 17. The Q-criterion distinguishes rotation-dominated regions (vortex cores, $Q > 0$) from strain-dominated regions (shear layers and transport pathways, $Q < 0$). The scatter plots show that, at a moderate Rayleigh number ($Ra = 10^8$), passive tracers are mainly trapped in rotation-dominated cores, with only 15.6 % located in the $Q < 0$ regime (see figure 17*a*).

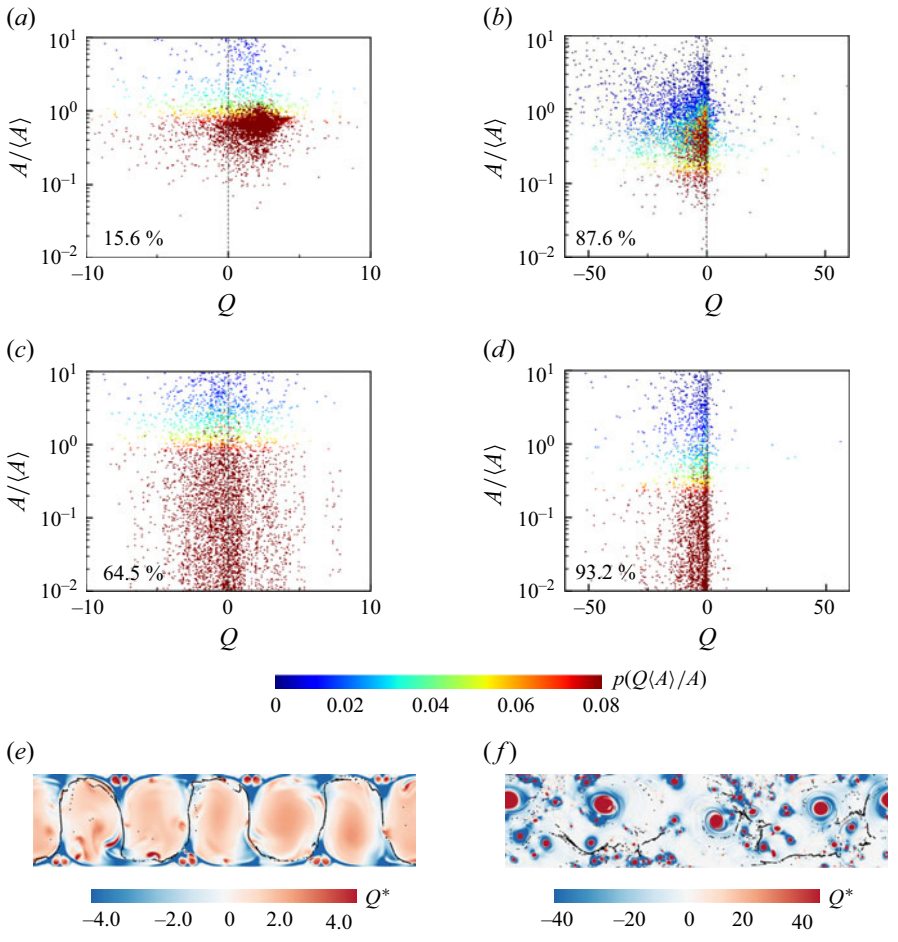


Figure 17. Joint probability distributions of the Voronoi-cell area A and the local Q-value at $t^* = 9.00$ for the reference actuation conditions. Passive tracers at (*a*) $Ra = 10^8$ and (*b*) $Ra = 10^{10}$; (*c*,*d*) active navigators at the corresponding Rayleigh numbers. The colour scale in (*a–d*) denotes the joint probability density $p(Q, A/\langle A\rangle)$, where $\langle A\rangle$ is the mean Voronoi-cell area. The corresponding instantaneous particle distributions at $t^* = 9.00$, overlaid on the instantaneous Q^* field for (*e*) $Ra = 10^8$ and (*f*) $Ra = 10^{10}$.

By contrast, active navigators are redistributed away from these traps, shifting a large fraction of their distribution (64.5 %) into the strain-dominated $Q < 0$ regions (see figure 17*c*). Furthermore, the high-probability-density region (red) simultaneously shifts towards smaller Voronoi-cell areas, $A/\langle A\rangle < 1$, suggesting that active navigators not only avoid vortical regions but also cluster within narrow and favourable pathways. Even at $Ra = 10^{10}$, where the highly turbulent background flow already scatters most passive tracers into $Q < 0$ regions (87.6 %; see figure 17*b*), the active navigators further increase this fraction to 93.2 % (see figure 17*d*). More importantly, comparison of figures 17(*b*) and 17(*d*) shows that the RL policy transforms a dispersed passive distribution into a more concentrated cluster at low $A/\langle A\rangle$, suggesting pronounced spatial agglomeration within strain-dominated regions. By actively avoiding $Q > 0$ regions (see figures 17*e* and 17*f*), the active navigators reduce closed-orbit recirculation and remain in the strain-dominated outer layers, which are favourable for sustained downstream migration. This interpretation

is consistent with our trajectory analysis, which showed that the learned strategy tends to follow plume-assisted pathways and remain clustered near fragmented transport barriers.

To connect the Eulerian and Lagrangian interpretations, it is useful to clarify how the instantaneous Q-criterion relates to the FTLE fields. These two diagnostics are complementary but not equivalent. The Q-criterion provides a local and instantaneous Eulerian classification of the flow topology, with $Q > 0$ corresponding to rotation-dominated vortex cores and $Q < 0$ to strain-dominated regions. By contrast, FTLE ridges identify finite-time LCS that organise material transport over a finite integration window. In the present flow, the $Q > 0$ vortex cores are often located inside regions enclosed or constrained by repelling LCS, whereas the surrounding $Q < 0$ regions are more closely associated with stretching and inter-roll transport. This correspondence provides a physical link between the two analyses: by avoiding $Q > 0$ regions, the active navigators reduce the likelihood of remaining trapped in recirculating vortex cores; by staying closer to strain-dominated regions, they are more likely to access transport pathways identified in the LCS analysis. Importantly, computing FTLE fields requires non-local, time-resolved velocity information over a finite integration window, which is not directly available to an autonomous agent making real-time decisions. The agent instead relies on instantaneous, locally measurable Eulerian cues, including the velocity-gradient information from which Q-like topology can be inferred. The consistency among the learned trajectories, the FTLE/LCS fields and the Voronoi statistics therefore suggests that local Eulerian flow topology may serve as a practical proxy for aspects of the finite-time Lagrangian transport structure.

Motivated by these physical observations, we distil a simple heuristic strategy from the behaviour of the learned policy. The key idea is that the particle switches between two modes: a drift mode, in which horizontal propulsion is switched off to avoid inefficient thrust expenditure when the particle is located in a locally adverse but weakly constrained flow region; and a thrust mode, in which stronger horizontal propulsion is applied to escape confinement, cross barriers or maintain progress through unfavourable flow structures. For the horizontal-migration task, we formalise this heuristic strategy using two threshold parameters. Because the absolute magnitudes of the fluid velocity and velocity gradients vary substantially across the Rayleigh numbers considered here, using dimensional thresholds would require case-dependent calibration. For example, the magnitude of the Q-criterion differs by about one order of magnitude between $Ra = 10^8$ and $Ra = 10^{10}$, as shown in figures 17(*e*) and 17(*f*). To avoid threshold tuning for each Ra, we define the trigger conditions using locally normalised, dimensionless variables, namely the normalised Q-criterion $\widetilde{Q} = Q/\|\nabla \boldsymbol{u}\|_F^2$ and the normalised vertical velocity $\widetilde{u}_y = u_y/\|\boldsymbol{u}\|_2$. These bounded quantities, $\widetilde{Q} \in [-0.5, 0.5]$ and $\widetilde{u}_y \in [-1, 1]$, characterise the relative dominance of local rotation and vertical motion rather than their absolute magnitudes. A single fixed set of thresholds is used for all Rayleigh numbers considered in this study:

$$\widetilde{Q}_{th} = 0.3, \quad \widetilde{u}_{y,th} = 0.2. \tag{3.4}$$

The active control acceleration $\boldsymbol{a}_{propel} = (a_{propel,x}, a_{propel,y})$ is then defined as

$$a_{propel,x} = \begin{cases} a_{drift}, & \text{if } \widetilde{Q} < \widetilde{Q}_{th},\ |\widetilde{u}_y| < \widetilde{u}_{y,th} \text{ and } u_x < 0, \\ a_{thrust}, & \text{otherwise,} \end{cases} \tag{3.5}$$

$$a_{propel,y} = \frac{\rho_p - \rho_f}{\rho_p} g, \tag{3.6}$$

where $a_{drift} = 0$, and the thrust-mode acceleration is

$$a_{thrust} = \sqrt{\|\boldsymbol{a}_{propel}\|_{max}^2 - a_{propel,y}^2}. \tag{3.7}$$

The drift condition $\widetilde{Q} < \widetilde{Q}_{th}$, $|\widetilde{u}_y| < \widetilde{u}_{y,th}$ and $u_x < 0$ corresponds to a state in which the particle is outside strongly rotation-dominated vortex cores, does not encounter strong vertical motion, but is embedded in a locally adverse horizontal flow. In this situation, applying strong positive horizontal thrust would mainly oppose the local current and lead to inefficient energy expenditure. The heuristic therefore sets $a_{drift} = 0$, allowing the particle to drift until the local flow carries it out of the adverse region. In all other cases, including proximity to rotation-dominated regions, strong vertical motion or situations requiring active barrier crossing, the particle switches to the thrust mode and applies constant horizontal propulsion. The thresholds used in this heuristic strategy were calibrated only in the lower-Ra, more coherent cases, namely $Ra = 10^7$ and $Ra = 10^8$, through parameter sweeps of $\widetilde{Q}_{th}$ and $\widetilde{u}_{y,th}$. The same values were then applied without further tuning to the higher-Ra cases, $Ra = 10^9$, 10^{10} and 10^{11}. Thus, the high-Ra performance can be interpreted as a transfer test of the heuristic strategy outside its calibration regime.

Representative trajectories obtained from this distilled heuristic strategy are shown in figures 18(*a*) and 18(*b*), illustrating the switching between the green drift mode and the magenta thrust mode. To contextualise the performance of the learned RL policy, we compare its energetics with those of this rule-based heuristic strategy, as shown in figures 18(*c*) and 18(*d*). At high Rayleigh numbers ($Ra \geqslant 10^9$), the learned RL policy requires less propulsion energy than the heuristic baseline, suggesting that the neural-network policy can exploit highly intermittent and spatiotemporally chaotic flow structures more effectively. Interestingly, at moderate Rayleigh number ($Ra = 10^8$), the heuristic strategy occasionally consumes less propulsion energy than the RL policy. However, this energetic advantage comes at the cost of longer completion times (see figure 18*e*). The rule-based heuristic agent can be overly conservative, switching off propulsion and becoming temporarily trapped within large, coherent LSC vortices. Because the RL reward function is calibrated to prioritise time efficiency and barrier crossing, the RL policy chooses to expend slightly more energy to escape these vortex traps more actively, thereby completing the task much faster. Furthermore, as shown in figure 18(*f*), although this heuristic strategy is much simpler than the learned RL policy and relies on only three interpretable criteria, it reproduces the main transport logic of the learned policy. It should be noted that the distilled heuristic is intended primarily as an interpretability tool and a simple rule-based controller derived from the flow-topology analysis of the learned policy, rather than as a completely independent baseline or a universally optimal controller. The comparison shows that key features of the learned behaviour can be related to local Eulerian observables such as the Q-criterion and the vertical velocity. The fact that the heuristic retains high success rates at higher Rayleigh numbers suggests that the normalised local topology and velocity cues capture features of the navigation problem that remain useful across the range of flow regimes considered here. Thus, the RL policy is not merely an opaque numerical optimiser, but can provide a route for extracting interpretable navigation rules with reasonable transferability across the present Ra range.

4. Conclusion

In this work, we investigated the navigation of a self-propelled inertial particle in two-dimensional RB convection over the range $10^7 \leq Ra \leq 10^{11}$ at $Pr = 0.71$ and $\Gamma = 4$. An

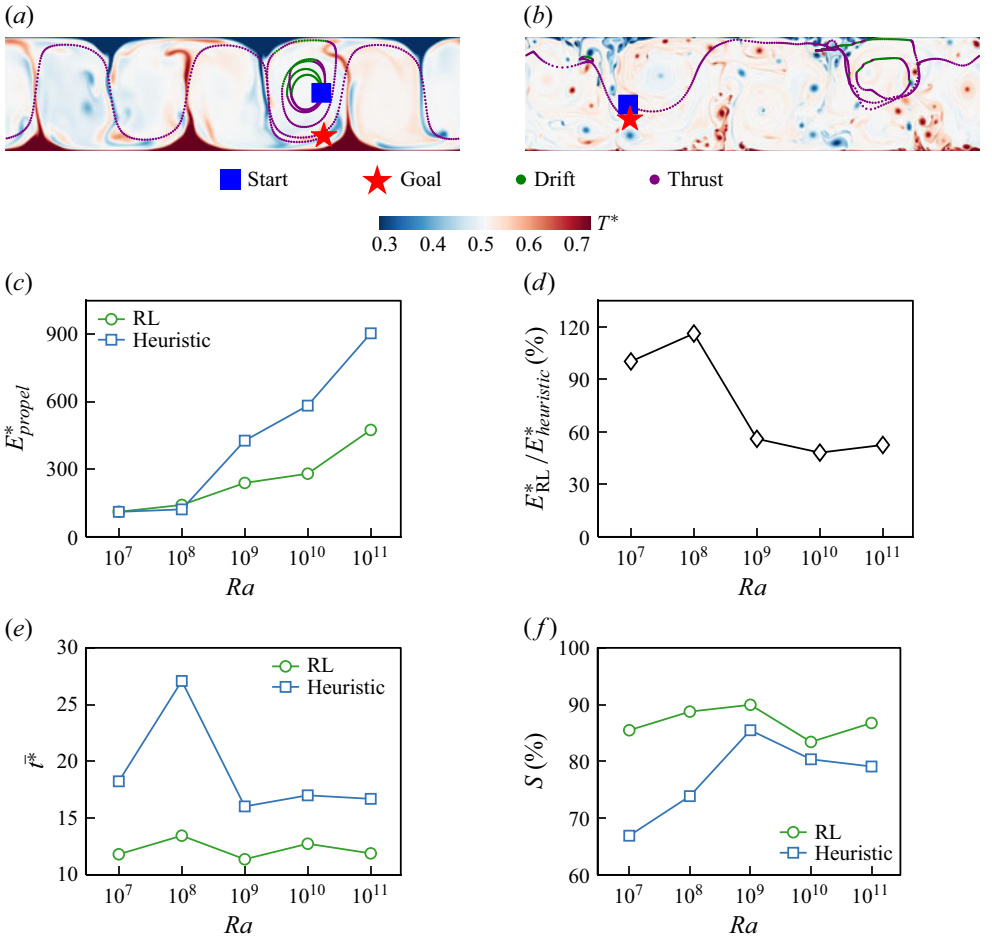


Figure 18. Comparison of representative trajectories and performance between the distilled heuristic strategy and the learned RL policy. (*a*,*b*) Representative trajectories of the heuristic strategy at the reference actuation bounds: (*a*) $Ra = 10^8$ with $\mathcal{A}_{max} = 1.25$ and (*b*) $Ra = 10^{10}$ with $\mathcal{A}_{max} = 4.0$. The background colour indicates the instantaneous temperature field T^*. The blue square and red star mark the start and goal locations, respectively. Green solid lines denote the drift mode, whereas magenta dotted lines denote the thrust mode. (*c*) The propulsion energy E^*_{propel} as a function of Ra for both the RL policy and the heuristic strategy. (*d*) The corresponding energy ratio $E^*_{RL}/E^*_{Heuristic}$. (*e*) The mean completion time $\bar{t}^*$ and (*f*) the success rate S as functions of Ra. The heuristic strategy uses a single fixed set of normalised thresholds for all Rayleigh numbers considered here, with $\widetilde{Q}_{th} = 0.3$ and $\widetilde{u}_{y,th} = 0.2$.

RL controller selected the propulsive acceleration, subject to an upper bound, to complete a fixed-displacement task. We evaluated macroscopic navigation performance in terms of reachability, completion time and propulsion energy, and interpreted the underlying physical mechanisms using Eulerian flow topology and LCS.

Our main findings on macroscopic performance are as follows. The success rate increases monotonically with the actuation limit $\mathcal{A}_{max}$. At moderate Ra, this increase is abrupt, whereas at higher Ra the transition becomes more gradual and shifts to larger $\mathcal{A}_{max}$. Although higher-Ra flows require stronger actuation to achieve reachability and lead to longer completion times, the energetic cost of traversal decreases once the task becomes feasible. Proper orthogonal decomposition indicates that these trends are associated with reorganisation of the carrier flow. At moderate Ra, kinetic energy is concentrated in a

few modes, and the multi-roll LSC creates persistent transport barriers that require a finite thrust surplus to cross. At higher Ra, energy is distributed across many modes, causing the barriers to fragment and intermittent plume-assisted pathways to emerge.

We further identified the kinematic and topological strategies that enable flow-assisted navigation. The learned policy balances progress and energy expenditure by aligning its thrust with favourable local currents. Lagrangian analysis suggests that the agent modulates its propulsion to cross repelling LCS barriers and then follow attracting transport pathways. By mapping this behaviour onto the local Eulerian flow fields via Voronoi tessellation and the Q-criterion, we found that the agent preferentially avoids rotation-dominated vortex cores ($Q > 0$) and exploits strain-dominated regions ($Q < 0$). Based on these insights, we distilled an interpretable, physics-based heuristic strategy that reproduces the main barrier-crossing logic of the learned policy, retaining good navigability using only instantaneous and locally measurable cues.

It is important to note that this flow-assisted navigation strategy is tied to the heavy-particle, low-Stokes-number regime considered here, with $St \sim \mathcal{O}(10^{-3})$. In this regime, the particle rapidly equilibrates with the local flow, and unsteady hydrodynamic effects such as added mass and Basset history forces remain small. For particles with larger inertia, for example $St \sim \mathcal{O}(1)$, finite-size effects, added mass and history forces may become important. The particle would then respond less directly to the local flow, and the navigation logic may shift from the present flow-assisted, topology-sensitive behaviour towards a more inertia-dominated crossing strategy. Similarly, when rotational dynamics is important, additional orientational control would be required to counteract flow-induced torques, which could modify the energy–time trade-off.

Although the present framework examines flow-assisted translational navigation, several idealisations and scope limitations should be noted. First, the learned policy is direction-specific by construction. Because the convective flow structures, including updrafts, downdrafts and large-scale rolls, as well as the domain boundaries, are anisotropic, a policy trained only for horizontal traversal is not expected to retain the same performance when deployed in arbitrary directions. Second, the use of absolute position to track fixed-displacement progress limits the expected generalisability of the policy across modified domain geometries. The spatiotemporal variability of the flow reduces the possibility of simply memorising static flow features, but changes in the geometric configuration may still affect performance. For example, in domains with different aspect ratios, the local gradient-based navigation logic may remain partly useful, whereas the modified organisation of the LSC could change the preferred transport pathways and reduce the overall success rate. In non-periodic horizontal domains with solid lateral walls, additional effects such as lateral boundary layers and corner recirculation would arise; a policy trained only in a periodic domain may not have learned the behaviours needed to navigate such boundaries.

The two-dimensionality of the present set-up also frames the scope of the transport-barrier interpretation. In three-dimensional RB convection, the LSC may exhibit spanwise deformation, torsion and more complex plume dynamics. The corresponding boundary of LCS would also change from one-dimensional repelling curves to two-dimensional material surfaces. As a result, an active agent in a three-dimensional flow may use out-of-plane motion to move around or along such surfaces, rather than crossing transport barriers in the same manner as in the present two-dimensional setting. Therefore, the LCS-based barrier-crossing picture should be interpreted as a mechanistic explanation for navigation under two-dimensional topological constraints, rather than as a direct quantitative prediction for three-dimensional flows.

Finally, the present particle model focuses on translational dynamics. Incorporating rotational degrees of freedom, orientational response times and the simultaneous control of translational and angular accelerations would be a useful extension towards more realistic models of artificial swimmers. Beyond these kinematic and dynamical extensions, another open direction concerns the thermodynamics of active navigation. Although our current objective function penalises only the agent's mechanical propulsive effort, from a stochastic-thermodynamic perspective the continuous processes of sensing the turbulent environment, processing information and executing a control policy may also incur thermodynamic costs. Recent studies have shown that autonomous localisation and self-steering are subject to dissipation–accuracy trade-offs (Cocconi, Mahault & Piro 2025). Accounting for the full thermodynamic dissipation associated with active control in complex convective flows therefore represents an important direction for future research, linking optimal navigation strategies to the physical limits of non-equilibrium active matter.

Overall, the present study suggests that reinforcement learning can be used not only as an optimisation tool, but also as a means of extracting interpretable physical strategies from complex flow environments. By connecting navigation performance with the reorganisation of turbulent transport barriers and pathways, our results provide a framework for understanding and designing energy-efficient autonomous navigation in buoyancy-driven flows.

Supplementary movies. Supplementary movies are available at https://doi.org/10.1017/jfm.2026.11802.

Funding. This work was supported by the National Natural Science Foundation of China (NSFC) through grant nos. 12272311, 12388101, 12125204; the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001); the Fundamental Research Funds for the Central Universities (no. D5000260269); and the 111 project of China (project no. B17037). The authors acknowledge the Beijing Beilong Super Cloud Computing Co. Ltd for providing HPC resources that have contributed to the research results reported within this paper (http://www.blsc.cn/).

Declaration of interests. The authors report no conflict of interest.

Appendix A. Soft actor–critic reinforcement-learning algorithm

The SAC algorithm is an off-policy actor–critic deep-reinforcement-learning algorithm formulated within the maximum-entropy RL framework. The key idea of SAC is to augment the standard RL objective with an entropy term, thereby encouraging exploration while optimising the expected return. The SAC objective can be written as

$$\pi_\theta^* = \arg\max_{\pi_\theta} \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\left\{r_t(s_t, a_t, s_{t+1}) + \alpha\mathcal{H}\left[\pi_\theta(\cdot|s_t)\right]\right\}\right], \tag{A1}$$

where π_θ denotes the policy, θ denotes the neural-network parameters of the policy and π_θ^* denotes the optimal policy. Furthermore, r_t is the reward at time step t, s_t and a_t are the state and action at time step t, γ is the discount factor, α is the temperature parameter that controls the trade-off between reward maximisation and entropy maximisation and $\mathcal{H}[\pi_\theta(\cdot|s_t)]$ is the entropy of the policy at state s_t.

The SAC consists of three main components:

(i) ***Q*-networks.** The Q-networks are neural networks that approximate the action–value function $Q(s, a)$. The SAC employs two Q-networks to reduce overestimation bias. The corresponding loss functions are

$$L(Q_{\phi,i}) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim\mathcal{D}}\left[\left(Q_{\phi,i}(s_t, a_t) - y_t\right)^2\right], \tag{A2}$$

where

$$y_t = r_t + \gamma\Big[\min_{i=1,2} Q_{\phi^{target},i}(s_{t+1}, a'_{t+1}) - \alpha \log \pi_\theta(a'_{t+1}|s_{t+1})\Big] \tag{A3}$$

is the target value. Here, $\mathcal{D}$ is the replay buffer, $Q_{\phi^{target},i}$ denotes the ith target Q-network, ϕ and ϕ^{target} are the neural-network parameters of the Q-networks and target Q-networks, respectively, and $a'_{t+1} \sim \pi_\theta(\cdot|s_{t+1})$ is the action sampled from the current policy at the next state.

(ii) **Target Q-networks.** The target Q-networks are used to stabilise the training of the Q-networks. The SAC maintains target Q-networks that are updated through the soft-update mechanism:

$$\phi^{target} \leftarrow \tau\phi + (1-\tau)\phi^{target}, \tag{A4}$$

where τ is a small constant that controls the update rate. This procedure reduces rapid variations in the target values and improves training stability.

(iii) **Policy network.** The policy network represents the policy π_θ. It is trained to maximise the expected return together with the policy entropy. The corresponding loss function is

$$L(\pi_\theta) = \mathbb{E}_{s_t\sim\mathcal{D},\,\xi\sim\mathcal{N}(0,I)}\Big[\alpha \log \pi_\theta(\tilde{a}_\theta(s_t, \xi)|s_t) - \min_{i=1,2} Q_{\phi,i}(s_t, \tilde{a}_\theta(s_t, \xi))\Big], \tag{A5}$$

where

$$\tilde{a}_\theta(s, \xi) = \tanh\left(\mu_\theta(s) + \sigma_\theta(s) \odot \xi\right) \tag{A6}$$

is the reparametrisation used to sample bounded actions from the Gaussian policy, with $\xi \sim \mathcal{N}(0, I)$. Here, $\mu_\theta(s)$ and $\sigma_\theta(s)$ are the mean and standard deviation predicted by the policy network, and $\odot$ denotes elementwise multiplication.

Moreover, we implement the automatic entropy-adjustment mechanism in SAC to adaptively tune the temperature parameter α during training. The loss function for the temperature parameter is

$$L(\alpha) = \mathbb{E}_{s_t\sim\mathcal{D},\,a_t\sim\pi_\theta}\left[-\alpha\left(\log \pi_\theta(a_t|s_t) + \bar{\mathcal{H}}\right)\right], \tag{A7}$$

where $\bar{\mathcal{H}}$ is the target entropy.

The SAC training procedure consists of the following steps:

(i) Collect experience by interacting with the environment using the current policy.
(ii) Store the collected experience in a replay buffer.
(iii) Sample a mini-batch of experiences from the replay buffer.
(iv) Update the temperature parameter α by minimising the loss function in (A7).
(v) Update the Q-networks by minimising the loss functions in (A2) using the sampled mini-batch.
(vi) Update the policy network by minimising the loss function in (A5) using the sampled mini-batch.
(vii) Update the target Q-networks using the soft-update mechanism in (A4).

The hyperparameters used to train the SAC agent in this work are listed in table 2. The entries for the number of hidden layers and the number of units per layer specify the architecture of all neural networks used in the SAC agent, including the Q-networks, target Q-networks and the policy network.

Hyperparameter	Value
Learning rate	3×10^{-4}
Discount factor γ	0.99
Replay-buffer size	10^6
Batch size	256
Target update rate τ	0.005
Target entropy $\bar{\mathcal{H}}$	-2
Hidden layers	2
Hidden units per layer	256
Activation function	ReLU

Table 2. Hyperparameters used to train the SAC agent.

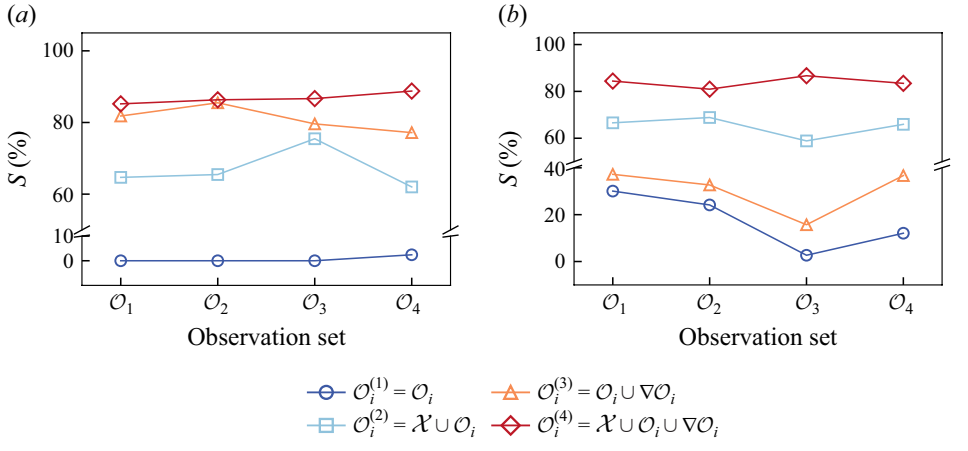


Figure 19. Success rate S for different observation state spaces at (*a*) $Ra = 10^8$ and (*b*) $Ra = 10^{10}$, evaluated at the corresponding optimal values of $\mathcal{A}_{max}$. Four baseline observation sets are considered: $\mathcal{O}_1 = \{\boldsymbol{u}_f^*\}$, $\mathcal{O}_2 = \{\boldsymbol{u}_f^*, T_f^*\}$, $\mathcal{O}_3 = \{\boldsymbol{u}_f^*, \boldsymbol{u}_p^*\}$ and $\mathcal{O}_4 = \{\boldsymbol{u}_f^*, \boldsymbol{u}_p^*, T_f^*\}$. For each baseline set $\mathcal{O}_i$, three augmentations are considered: the inclusion of spatial information $\mathcal{X}$, the inclusion of gradient information $\nabla\mathcal{O}_i$ and the inclusion of both. This yields four configurations per baseline: $\mathcal{O}_i^{(1)} = \mathcal{O}_i$, $\mathcal{O}_i^{(2)} = \mathcal{X} \cup \mathcal{O}_i$, $\mathcal{O}_i^{(3)} = \mathcal{O}_i \cup \nabla\mathcal{O}_i$ and $\mathcal{O}_i^{(4)} = \mathcal{X} \cup \mathcal{O}_i \cup \nabla\mathcal{O}_i$, giving 16 configurations in total.

Appendix B. Ablation study on the observation state space

To assess the contribution of different sensory inputs, we conducted a systematic comparison across multiple observation sets. Specifically, we considered four baseline sets:

$$\mathcal{O}_1 = \{\boldsymbol{u}_f^*\}, \quad \mathcal{O}_2 = \{\boldsymbol{u}_f^*, T_f^*\}, \quad \mathcal{O}_3 = \{\boldsymbol{u}_f^*, \boldsymbol{u}_p^*\} \quad \text{and} \quad \mathcal{O}_4 = \{\boldsymbol{u}_f^*, \boldsymbol{u}_p^*, T_f^*\}. \tag{B1}$$

For each baseline, we then considered three augmentations: the inclusion of spatial information $\mathcal{X} = \{\boldsymbol{x}_p^*\}$, the inclusion of gradient information $\nabla\mathcal{O}_i$ (including $\nabla\boldsymbol{u}_f$ and ∇T_f, where applicable) and the inclusion of both. This yields 16 combinations in total. The resulting success rates S, evaluated at the corresponding optimal values of $\mathcal{A}_{max}$ for $Ra = 10^8$ and 10^{10}, are shown in figure 19. These results show that both positional and gradient information improve navigation performance, in agreement with the findings of Jiao *et al.* (2025). The highest success rates are achieved when both are included

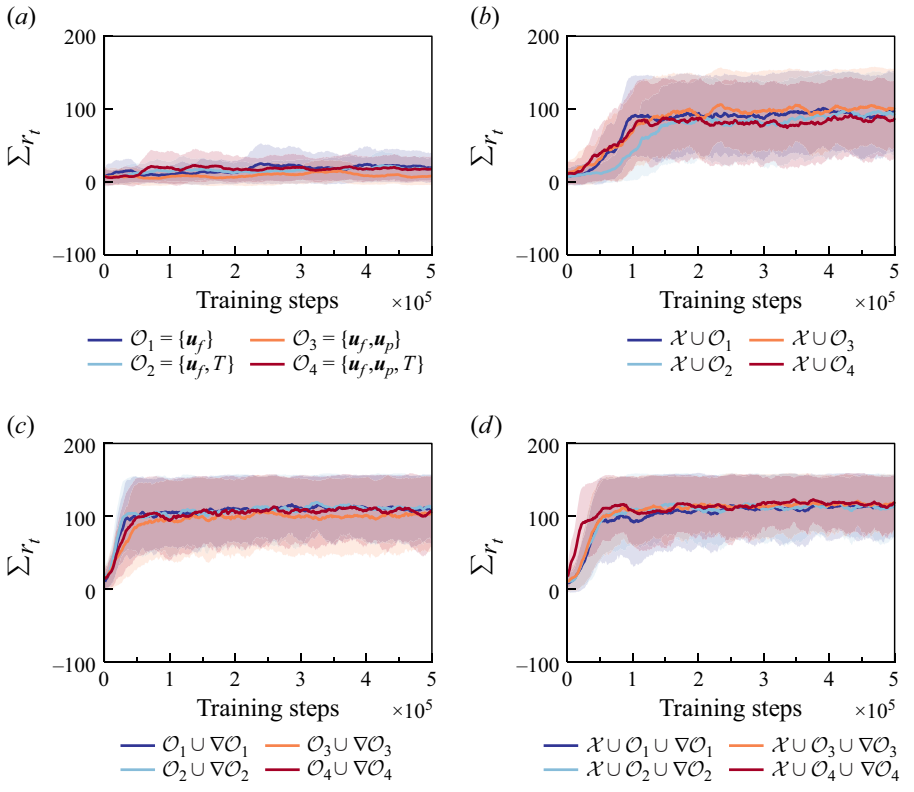


Figure 20. Sample learning curves showing the averaged cumulative reward $\sum r_t$ as a function of training time steps for $Ra = 10^8$. The panels compare training convergence for different observation sets: (*a*) the four baseline sets $\mathcal{O}_i$, (*b*) baselines augmented with spatial information $\mathcal{X}$, (*c*) baselines augmented with gradient information $\nabla\mathcal{O}_i$ and (*d*) baselines augmented with both $\mathcal{X}$ and $\nabla\mathcal{O}_i$. Shaded regions represent the standard deviation across training seeds.

simultaneously, indicating that gradient information is not redundant with the basic flow observations.

Furthermore, the relative importance of these observations depends on the flow regime. At lower $Ra = 10^8$, the flow is more coherent and spatially organised, so gradient-based local cues are more strongly correlated with the large-scale dynamics, yielding a clear performance gain. At higher $Ra = 10^{10}$, however, the flow becomes more intermittent, and local gradient information alone is less sufficient for the agent to infer its global progress towards the target. In this regime, explicit positional information ($\mathcal{X}$) becomes more valuable for maintaining global orientation. These results suggest that both spatial coordinates and fluid gradients contribute to effective autonomous navigation in spatiotemporally chaotic environments.

To assess algorithmic stability and verify that the performance variations observed in the ablation study are not artefacts of insufficient training time, we present sample learning curves in figure 20 ($Ra = 10^8$) and figure 21 ($Ra = 10^{10}$). These curves show the evolution of the averaged cumulative reward during training. For observation sets that provide sufficient physical context, for example configurations incorporating both spatial coordinates and local fluid gradients, the reward rises rapidly and reaches a stable high-value plateau within approximately 1×10^5 training steps. By contrast, policies restricted

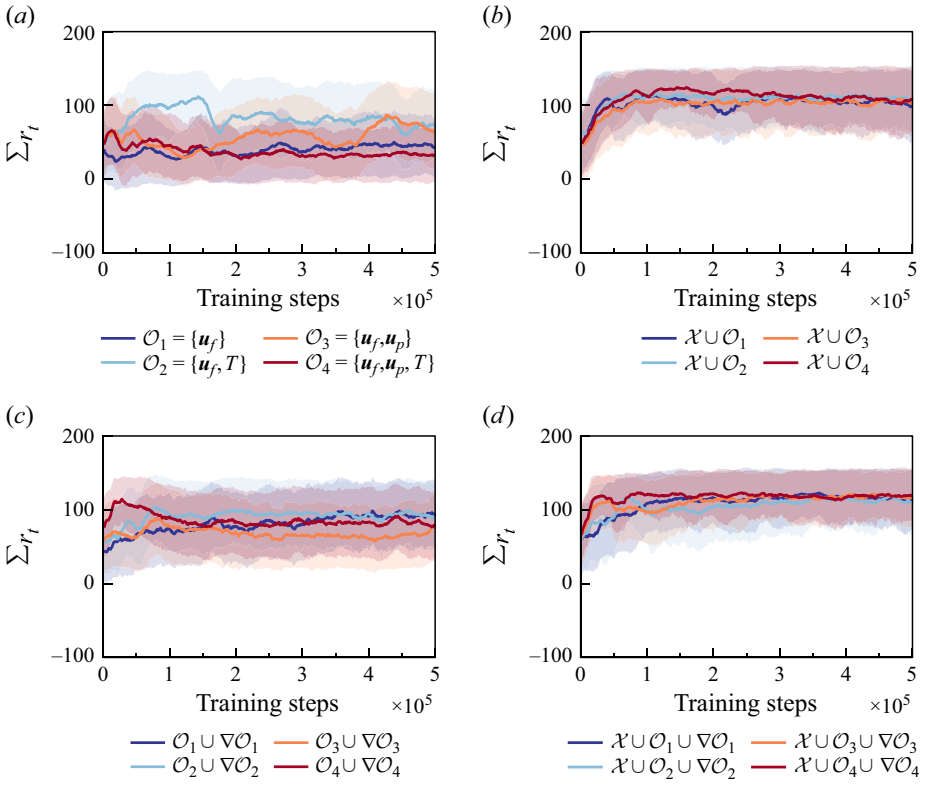


Figure 21. Sample learning curves showing the averaged cumulative reward $\sum r_t$ as a function of training time steps for $Ra = 10^{10}$. The panels compare training convergence for different observation sets: (*a*) the four baseline sets $\mathcal{O}_i$, (*b*) baselines augmented with spatial information $\mathcal{X}$, (*c*) baselines augmented with gradient information $\nabla\mathcal{O}_i$ and (*d*) baselines augmented with both $\mathcal{X}$ and $\nabla\mathcal{O}_i$. Shaded regions represent the standard deviation across training seeds.

to less informative observation sets, for example those using only local velocity, exhibit reward curves that remain at low values and fluctuate strongly, with no clear sign of convergence. Since all agents were trained for a total of 1×10^6 training steps, these curves indicate that the training duration was sufficient for the cases considered. The inability of the restricted agents to achieve high rewards therefore appears to arise from the absence of key physical information for navigating the spatiotemporally chaotic flow, rather than from premature termination of training.

References

Akos, Z., Nagy, M. & Vicsek, T. 2008 Comparing bird and human soaring strategies. *Proc. Natl Acad. Sci. USA* **105**, 4139–4143.

Bechinger, C., Di Leonardo, R., Löwen, H., Reichhardt, C., Volpe, G. & Volpe, G. 2016 Active particles in complex and crowded environments. *Rev. Mod. Phys.* **88**, 045006.

Bellemare, M.G., Candido, S., Castro, P.S., Gong, J., Machado, M.C., Moitra, S., Ponda, S.S. & Wang, Z. 2020 Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* **588**, 77–82.

Biferale, L., Bonaccorso, F., Buzzicotti, M., Clark di Leoni, P. & Gustavsson, K. 2019 Zermelo's problem: optimal point-to-point navigation in 2D turbulent flows using reinforcement learning. *Chaos* **29**, 103138.

BORRA, F., BIFERALE, L., CENCINI, M. & CELANI, A. 2022 Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number. *Phys. Rev. Fluids* **7**, 023103.
BRUNTON, S.L., NOACK, B.R. & KOUMOUTSAKOS, P. 2020 Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508.
CASTILLO-CASTELLANOS, A., SERGENT, A., PODVIN, B. & ROSSI, M. 2019 Cessation and reversals of large-scale structures in square Rayleigh–Bénard cells. *J. Fluid Mech.* **877**, 922–954.
CHILLÀ, F. & SCHUMACHER, J. 2012 New perspectives in turbulent Rayleigh–Bénard convection. *Eur. Phys. J. E* **35**, 58.
CICHOS, F., GUSTAVSSON, K., MEHLIG, B. & VOLPE, G. 2020 Machine learning for active matter. *Nat. Mach. Intell.* **2** (2), 94–103.
COCCONI, L., MAHAULT, B. & PIRO, L. 2025 Dissipation-accuracy tradeoffs in autonomous control of smart active matter. *New J. Phys.* **27** (1), 013002.
COLABRESE, S., GUSTAVSSON, K., CELANI, A. & BIFERALE, L. 2017 Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004.
EMRAN, M.S. & SCHUMACHER, J. 2010 Lagrangian tracer dynamics in a closed cylindrical turbulent convection cell. *Phys. Rev. E* **82**, 016303.
FISCHER, P.F. 1997 An overlapping Schwarz method for spectral element solution of the incompressible Navier–Stokes equations. *J. Comput. Phys.* **133** (1), 84–101.
GAO, Z.-Y., TAO, X., HUANG, S.-D., BAO, Y. & XIE, Y.-C. 2024 Flow state transition induced by emergence of orbiting satellite eddies in two-dimensional turbulent Rayleigh–Bénard convection. *J. Fluid Mech.* **997**, A54.
GUNNARSON, P., MANDRALIS, I., NOVATI, G., KOUMOUTSAKOS, P. & DABIRI, J.O. 2021 Learning efficient navigation in vortical flow fields. *Nat. Commun.* **12**, 7143.
GUSTAVSSON, K., BERGLUND, F., JONSSON, P.R. & MEHLIG, B. 2016 Preferential sampling and small-scale clustering of gyrotactic microswimmers in turbulence. *Phys. Rev. Lett.* **116**, 108104.
GUSTAVSSON, K., BIFERALE, L., CELANI, A. & COLABRESE, S. 2017 Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning. *Eur. Phys. J. E* **40**, 110.
HAARNOJA, T., ZHOU, A., ABBEEL, P. & LEVINE, S. 2018 Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1861–1870. PMLR.
HALLER, G. 2015 Lagrangian coherent structures. *Annu. Rev. Fluid Mech.* **47**, 137–162.
HANG, H., JIAO, Y., MEREL, J. & KANSO, E. 2026 Flow currents support simple and versatile trail-tracking strategies. *Phys. Rev. Res.* **8** (1), 013019.
HE, J.-C., BAO, Y. & CHEN, X. 2024 Turbulent boundary layers in thermal convection at moderately high Rayleigh numbers. *Phys. Fluids* **36** (2), 025140.
HEINONEN, R.A., BIFERALE, L., CELANI, A. & VERGASSOLA, M. 2025 Exploring Bayesian olfactory search in realistic turbulent flows. *Phys. Rev. Fluids* **10** (6), 064614.
ISHIKAWA, T. 2025 Transport phenomena in microswimmer suspensions: migration, collective motion, diffusion and rheology. *J. Fluid Mech.* **1016**, P1.
JIAO, Y., HANG, H., MEREL, J. & KANSO, E. 2025 Sensing flow gradients is necessary for learning autonomous underwater navigation. *Nat. Commun.* **16** (1), 3044.
KOOIJ, G.L., BOTCHEV, M.A., FREDERIX, E.M., GEURTS, B.J., HORN, S., LOHSE, D., VAN DER POEL, E.P., SHISHKINA, O., STEVENS, R.J. & VERZICCO, R. 2018 Comparison of computational codes for direct numerical simulations of turbulent Rayleigh–Bénard convection. *Comput. Fluids* **166**, 1–8.
KRISHNA, K., BRUNTON, S.L. & SONG, Z. 2023 Finite time lyapunov exponent analysis of model predictive control and reinforcement learning. *IEEE Access* **11**, 118916–118930.
KRISHNA, K., SONG, Z. & BRUNTON, S.L. 2022 Finite-horizon, energy-efficient trajectories in unsteady flows. *Proc. R. Soc. A* **478** (2258), 20210255.
LAUGA, E. & POWERS, T.R. 2009 The hydrodynamics of swimming microorganisms. *Rep. Prog. Phys.* **72** (9), 096601.
LAURENT, K.M., FOGG, B., GINSBURG, T., HALVERSON, C., LANZONE, M.J., MILLER, T.A., WINKLER, D.W. & BEWLEY, G.P. 2021 Turbulence explains the accelerations of an eagle in natural flight. *Proc. Natl Acad. Sci. USA* **118** (23), e2102588118.
LOHSE, D. & SHISHKINA, O. 2023 Ultimate turbulent thermal convection. *Phys. Today* **76** (11), 26–32.
LOHSE, D. & SHISHKINA, O. 2024 Ultimate Rayleigh–Bénard turbulence. *Rev. Mod. Phys.* **96** (3), 035001.
LOHSE, D. & XIA, K.Q. 2010 Small-scale properties of turbulent Rayleigh–Bénard convection. *Annu. Rev. Fluid Mech.* **42**, 335–364.
LOISY, A. & ELOY, C. 2022 Searching for a source without gradients: how good is infotaxis and how to beat it. *Proc. R. Soc. A* **478** (2262), 20220118.

MASMITJA, I., MARTIN, M., O'REILLY, T., KIEFT, B., PALOMERAS, N., NAVARRO, J. & KATIJA, K. 2023 Dynamic robotic tracking of underwater targets using reinforcement learning. *Sci. Robot.* **8** (80), eade7811.
MEHTA, P., BUKOV, M., WANG, C.H., DAY, A.G.R., RICHARDSON, C., FISHER, C.K. & SCHWAB, D.J. 2019 A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **810**, 1–124.
MONTHILLER, R., LOISY, A., KOEHL, M.A., FAVIER, B. & ELOY, C. 2022 Surfing on turbulence: a strategy for planktonic navigation. *Phys. Rev. Lett.* **129** (6), 064502.
MUIÑOS-LANDIN, S., FISCHER, A., HOLUBEC, V. & CICHOS, F. 2021 Reinforcement learning with artificial microswimmers. *Sci. Robot.* **6** (52), eabd9285.
NASIRI, M. & LIEBCHEN, B. 2022 Reinforcement learning of optimal active particle navigation. *New J. Phys.* **24** (7), 073042.
PANDEY, A., SCHEEL, J.D. & SCHUMACHER, J. 2018 Turbulent superstructures in Rayleigh–Bénard convection. *Nat. Commun.* **9**, 2118.
PIRO, L., HEINONEN, R.A., CENCINI, M. & BIFERALE, L. 2025 Many wrong models approach to localise an odour source in turbulence with static sensors. *J. Turbul.* **26** (5), 153–173.
PIRO, L., VILFAN, A., GOLESTANIAN, R. & MAHAULT, B. 2024 Energetic cost of microswimmer navigation: the role of body shape. *Phys. Rev. Res.* **6** (1), 013274.
PODVIN, B. & SERGENT, A. 2015 A large-scale investigation of wind reversal in a square Rayleigh–Bénard cell. *J. Fluid Mech.* **766**, 172–201.
QIU, J., MARCHIOLI, C. & ZHAO, L. 2022*a* A review on gyrotactic swimmers in turbulent flows. *Acta Mechanica Sin.* **38** (8), 722323.
QIU, J., MOUSAVI, N., GUSTAVSSON, K., XU, C., MEHLIG, B. & ZHAO, L. 2022*b* Navigation of micro-swimmers in steady flow: the importance of symmetries. *J. Fluid Mech.* **932**, A10.
QIU, J., MOUSAVI, N., ZHAO, L. & GUSTAVSSON, K. 2022*c* Active gyrotactic stability of microswimmers using hydromechanical signals. *Phys. Rev. Fluids* **7** (1), 014311.
REDDY, G., CELANI, A., SEJNOWSKI, T.J. & VERGASSOLA, M. 2016 Learning to soar in turbulent environments. *Proc. Natl Acad. Sci. USA* **113** (33), E4877–E4884.
REDDY, G., WONG-NG, J., CELANI, A., SEJNOWSKI, T.J. & VERGASSOLA, M. 2018 Glider soaring via reinforcement learning in the field. *Nature* **562** (7726), 236–239.
RIGOLLI, N., MAGNOLI, N., ROSASCO, L. & SEMINARA, A. 2022 Learning to predict target location with turbulent odor plumes. *eLife* **11**, e72196.
SAMUEL, R., SAMTANEY, R. & VERMA, M.K. 2022 Large-eddy simulation of Rayleigh–Bénard convection at extreme Rayleigh numbers. *Phys. Fluids* **34** (7), 075133.
SCHNEIDE, C., STAHN, M., PANDEY, A., JUNGE, O., KOLTAI, P., PADBERG-GEHLE, K. & SCHUMACHER, J. 2019 Lagrangian coherent sets in turbulent Rayleigh–Bénard convection. *Phys. Rev. E* **100** (5), 053103.
SCHNEIDER, E. & STARK, H. 2019 Optimal steering of a smart active particle. *EPL* **127** (6), 64003.
SENGUPTA, A., CARRARA, F. & STOCKER, R. 2017 Phytoplankton can actively diversify their migration strategy in response to turbulent cues. *Nature* **543** (7646), 555–558.
SHISHKINA, O. & LOHSE, D. 2024 Ultimate regime of Rayleigh–Bénard turbulence: subregimes and their scaling relations for the Nusselt vs Rayleigh and Prandtl numbers. *Phys. Rev. Lett.* **133** (14), 144001.
SIMONS, A.M. 2004 Many wrongs: the advantage of group navigation. *Trends Ecol. Evol.* **19** (9), 453–455.
SOUCASSE, L., PODVIN, B., RIVIÈRE, P. & SOUFIANI, A. 2019 Proper orthogonal decomposition analysis and modelling of large-scale flow reorientations in a cubic Rayleigh–Bénard cell. *J. Fluid Mech.* **881**, 23–50.
VERGASSOLA, M., VILLERMAUX, E. & SHRAIMAN, B.I. 2007 Infotaxis as a strategy for searching without gradients. *Nature* **445** (7126), 406–409.
VERMA, S., NOVATI, G. & KOUMOUTSAKOS, P. 2018 Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Natl Acad. Sci. USA* **115** (23), 5849–5854.
WANG, Q., VERZICCO, R., LOHSE, D. & SHISHKINA, O. 2020*b* Multiple states in turbulent large-aspect-ratio thermal convection: what determines the number of convection rolls? *Phys. Rev. Lett.* **125** (7), 074501.
WANG, B.F., ZHOU, Q. & SUN, C. 2020*a* Vibration-induced boundary-layer destabilization achieves massive heat-transport enhancement. *Sci. Adv.* **6** (21), eaaz8239.
WEIMERSKIRCH, H., BISHOP, C., JEANNIARD-DU DOT, T., PRUDOR, A. & SACHS, G. 2016 Frigate birds track atmospheric conditions over months-long transoceanic flights. *Science* **353** (6294), 74–78.
XIA, K.Q., CHONG, K.L., DING, G.Y. & ZHANG, L. 2025 Some fundamental issues in buoyancy-driven flows with implications for geophysical and astrophysical systems. *Acta Mechanica Sin.* **41** (1), 324287.
XIA, K.Q., HUANG, S.D., XIE, Y.C. & ZHANG, L. 2023 Tuning heat transport via coherent structure manipulation: recent advances in thermal turbulence. *Natl. Sci. Rev.* **10** (6), nwad012.
XU, A., SHI, L. & ZHAO, T.S. 2017 Accelerated lattice Boltzmann simulation using GPU and OpenACC with data management. *Intl J. Heat Mass Transfer* **109**, 577–588.

XU, A., WU, H.L. & XI, H.D. 2022 Migration of self-propelling agent in a turbulent environment with minimal energy consumption. *Phys. Fluids* **34** (3), 035117.

XU, A., WU, H.L. & XI, H.D. 2023 Long-distance migration with minimal energy consumption in a thermal turbulent environment. *Phys. Rev. Fluids* **8** (2), 023502.

YANG, J., DAVOODIIANIDALIK, M., XIA, H., PUNZMANN, H., SHATS, M. & FRANCOIS, N. 2019 Passive propulsion in turbulent flows. *Phys. Rev. Fluids* **4** (10), 104608.

ZHANG, Y., ZHOU, Q. & SUN, C. 2017 Statistics of kinetic and thermal energy dissipation rates in two-dimensional turbulent Rayleigh–Bénard convection. *J. Fluid Mech.* **814**, 165–184.

ZHU, G., FANG, W.Z. & ZHU, L. 2022 Optimizing low-Reynolds-number predation via optimal control and reinforcement learning. *J. Fluid Mech.* **944**, A3.

ZHU, Y., KANG, L., TONG, X., MA, J., TIAN, F. & FAN, D. 2025 Intermittent swimmers optimize energy expenditure with flick-to-flick motor control. *J. Fluid Mech.* **1006**, A27.

ZHU, X., MATHAI, V., STEVENS, R.J., VERZICCO, R. & LOHSE, D. 2018 Transition to the ultimate regime in two-dimensional Rayleigh–Bénard convection. *Phys. Rev. Lett.* **120** (14), 144502.