Soft Matter

PAPER

View Article Online
View Journal | View Issue

Cite this: *Soft Matter*, 2024, **20**, 2008

Learning how to find targets in the micro-world: the case of intermittent active Brownian particles

Michele Caraglio,* Harpreet Kaur, Lukas J. Fiderer, Andrea López-Incera, Hans J. Briegel, Thomas Franosch and Gorka Muñoz-Gil

Received 11th December 2023,
Accepted 29th January 2024

DOI: 10.1039/d3sm01680c

rsc.li/soft-matter-journal

Finding the best strategy to minimize the time needed to find a given target is a crucial task both in nature and in reaching decisive technological advances. By considering learning agents able to switch their dynamics between standard and active Brownian motion, here we focus on developing effective target-search behavioral policies for microswimmers navigating a homogeneous environment and searching for targets of unknown position. We exploit projective simulation, a reinforcement learning algorithm, to acquire an efficient stochastic policy represented by the probability of switching the phase, *i.e.* the navigation mode, in response to the type and the duration of the current phase. Our findings reveal that the target-search efficiency increases with the particle's self-propulsion during the active phase and that, while the optimal duration of the passive case decreases monotonically with the activity, the optimal duration of the active phase displays a non-monotonic behavior.

1 Introduction

Target search is a universal problem occurring in various fields and at several length scales.[1] Examples range from animals searching for food, mate, or shelter[2–5] to castaway rescue operations,[6] and to proteins binding to specific DNA sequences.[7,8] At the micro-scale, further paradigmatic examples include bacteria foraging nourishment,[9,10] phagocytes of the immune system performing chemotactic motion during infection,[11,12] and sperm cells on their way to the egg.[13] Furthermore, artificial and biohybrid microswimmers[14–16] with good target-search skills have been envisaged as revolutionary smart materials able to perform assisted fertilization,[17] targeted drug delivery,[18,19] or environmental remediation.[20]

In many relevant circumstances, the agent has no *a priori* knowledge of the target location and has to develop effective stochastic strategies that allow minimizing, at least on average, the search time in an environment with randomly distributed targets. Motivated by observational data, physical intuition, and analytical tractability, Lévy walks[21–23] and intermittent searches[1,24–26] are among the statistical strategies that have received major attention in the past. In the former, the agent undergoes straight runs at constant speed with run lengths l drawn from a Lévy distribution $p(l) \propto l^{-(\alpha+1)}$, with $0 < \alpha < 2$ and the target is detected if the searcher transits closer than a threshold distance, which also acts as a small-lengths cut-off allowing to normalize the Lévy distribution. The optimal value of α depends sensitively on model details such as the revisitability and mobility of the targets or the complexity of the environment.[22,27–29] Intermittent-search strategies have been proposed based on the observation that fast movements allow exploring quickly the whole environment but may, on the other hand, significantly degrade perception abilities.[3] In these strategies, phases of diffusive motion permitting target detection are alternated with phases of ballistic motion which allow quick relocation to different positions at the cost of not being able to detect the target. In the simplest version of the model, the agent switches from one phase to the other with a fixed rate leading to exponentially distributed phase durations, but other distributions have also been considered.[30,31] The mean search time of these strategies can be minimized under broad conditions. In particular, it has been shown that there is an optimal duration of the ballistic nonreactive phase which depends only on the dimensionality of the system and is independent of the details of the slow reactive phase.[26] Intermittent-search strategies remain robust also in the cases of different target distributions such as patches[32] and Poissonian distributions in one dimensions.[33]

In the past decade, machine learning has emerged as a revolutionary tool helping to elucidate various aspects of active matter systems.[34] In particular, reinforcement learning (RL)[35] and genetic algorithms[36] have proved to be powerful tools able to identify successful swimming strategies improving the navigation performances of microswimmers and their odds of reaching a target. Promising and worthy results have been obtained

Institut für Theoretische Physik, Universität Innsbruck, Technikerstraße 21A, A-6020, Innsbruck, Austria. E-mail: michele.caraglio@uibk.ac.at

in several situations including simple energy landscapes,[37] viscous surroundings,[38–40] complex motility fields,[41] and steady or turbulent flows.[42–46] However, previous literature has mainly focused on increasing the net flux of particles in a certain direction or on optimizing point-to-point navigation towards a target whose position is fixed and then implicitly learned during the learning process. Thus, notwithstanding the increasing popularity of machine-learning algorithms in the active matter field[34,47] and the previously mentioned seminal works, investigation of stochastic target-search problems with randomly distributed targets *via* machine-learning approaches remains largely unexplored, with only a couple of very recent exceptions.[48,49]

Muñoz-Gil *et al.*[48] have applied RL methods to learn optimal foraging strategies outperforming the efficiency of Lévy walks in the case of revisitable, sparsely distributed targets. In their setup, the learning agent performs a stepwise motion with constant velocity and, at each step, decides if maintaining the current direction or turning in a new random one, this choice being based only on the length of the current straight segment. Whenever the agent detects a target, it receives a reward and, through several trials, it optimizes its policy, learning an efficient distribution of the length of the straight segments. The former approach, with respect to traditional analytical ones, has the advantage of not being restricted to a specific ansatz of the straight-segment length distribution. However, it remains focused on investigating known idealized scenarios, which are not entirely apt for describing the behavior of real microswimmers.

On the other hand, Kaur *et al.*[49] rely on the active Brownian particle (ABP),[50] which well describes the behavior of artificial microswimmers, and show that genetic algorithms manage to address the problem of finding targets of unknown positions for particles able to decide if and when switching their behavior between standard passive Brownian diffusion and directed ABP motion. In particular, they use the algorithm NeuroEvolution of Augmenting Topologies[51] to evolve an initial population of particles taking random decisions towards a population in which the majority of particles are optimized to solve the target-search problem. However, their findings are limited by the fact that, in their setup, a given individual particle acts deterministically in the sense that it always selects the same duration for each phase from a set of predetermined durations.

In the present manuscript, we combine the two former approaches: While, as in ref. 49, we resort again on agents able to switch their behavior between passive and active Brownian motion, we here exploit the powerful RL framework employed in ref. 48, thus allowing our agents to learn a distribution of durations for each of the two phases. This results in a stochastic strategy maximizing the foraging efficiency in a homogeneous environment, which can eventually be tested in experiments with artificial Janus particles[52,53] where the activity is controlled by an external illuminating system.[38]

2 Model

With intermittent-search strategies in mind, we design our agent as a particle switching between two different phases ϕ and able to keep track of the current phase duration ω. More specifically, the particle can perform either standard Brownian diffusion ($\phi = 0$) enabling target detection, or active Brownian motion ($\phi = 1$), which does not allow to sense the target but, depending on its self-propulsion, may quickly relocate the particle to a different region. In the following, the two navigation modes are also referred to in short as the Brownian Particle (BP) phase and the ABP phase, respectively. At each time t, the *state* of the agent s_t is then characterized by the tuple $s_t = (\phi_t,\omega_t)$, with ϕ_t representing the current phase and ω_t its time duration since the last switching event. As customary in the RL framework,[35] given its current state s_t, the agent replies with an *action* a_t, and gains a *reward* if this action leads to a benefit for the agent. In our case, the action corresponds to making a decision on whether to maintain the current phase or switch to the other one. This choice follows a probabilistic rule, with p_t the probability of switching phase. We highlight the fact that, in our approach, p_t is not a constant but it depends on the current state s_t of the agent. The full set of these probabilities (one for each state s_t) constitutes the *policy* of the agent. During the learning process, such a policy is constantly updated with the goal of maximizing the total reward (see Methods section for more details).

Including these notions into the standard ABP model[50] in a homogeneous environment results in the following set of Langevin equations, discretized according to Itô rule,

$$\phi_{t+\Delta t} = \begin{cases} \phi_t & \text{with probability } 1-p_t, \\ 1-\phi_t & \text{with probability } p_t, \end{cases} \tag{1}$$

$$\boldsymbol{r}_{t+\Delta t} = \boldsymbol{r}_t + v\boldsymbol{u}_t\phi_t\Delta t + \sqrt{2D\Delta t}\boldsymbol{\xi}_t, \tag{2}$$

$$\vartheta_{t+\Delta t} = \begin{cases} \vartheta_t + \sqrt{2D_\vartheta\Delta t}\eta_t & \text{if } \phi_{t+\Delta t} = \phi_t, \\ 2\pi\,\mathrm{rand} & \text{otherwise.} \end{cases} \tag{3}$$

Here Δt is the integration time step, $\boldsymbol{r}_t = (x_t,y_t)$ is the position at time t, and $\boldsymbol{u}_t = (\cos\vartheta_t,\sin\vartheta_t)$ denotes the instantaneous orientation of the self-propulsion velocity with constant modulus v. D and D_ϑ are the translational and rotational diffusion coefficients, respectively. Finally, the components of the vector noise $\boldsymbol{\xi}_t = (\xi_{x,t},\xi_{y,t})$ and of the scalar noise η_t are independent random variables, distributed according to a Gaussian with zero average and unit variance. Note that when the phase of the particle is that of a passive Brownian particle ($\phi_t = 0$), the spatial evolution is decoupled from the orientational diffusion of the self-propulsion.

Our homogeneous environment is modeled as a two-dimensional square box of size $L \times L$ with periodic boundary conditions. A circular target of radius $R = 0.05L$ is located randomly inside the box. Every time the agent finds this target (*i.e.* the distance between the center of the target and the particle position is smaller than the target radius R), it gets a positive reward, the target is destroyed, and a new target appears at a new random location inside the box. Due to the periodic boundary conditions, this environment is formally equivalent to an infinite domain with a lattice of targets. Being the reward given only when a target is found, in order to

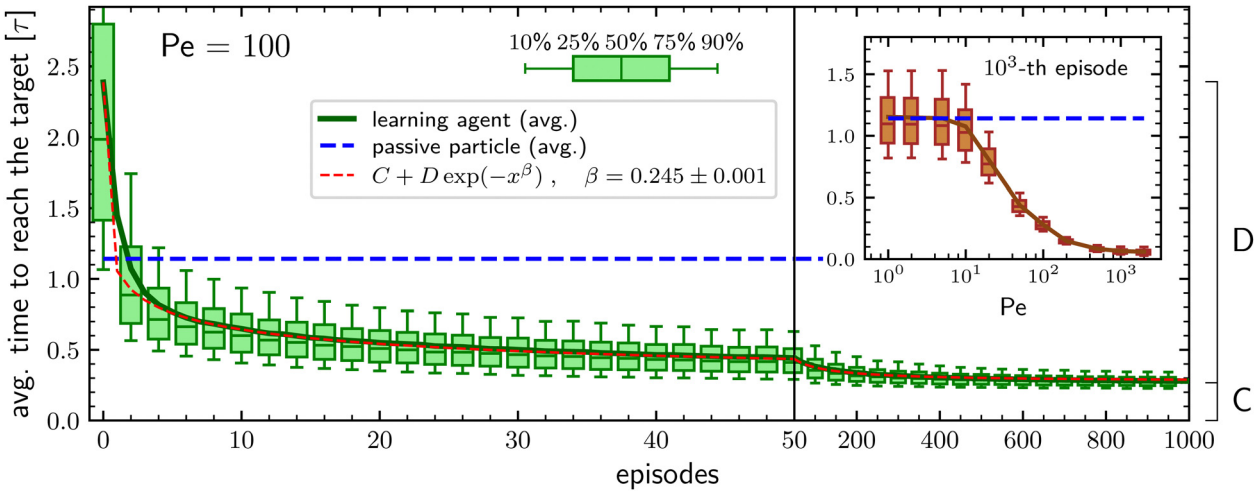

Fig. 1 Average time required to reach the target during an episode of duration 20τ as a function of the number of episodes for Pe = 100. The continuous green line represents the average over $N = 5 \times 10^3$ independent particles, while the box-and-whiskers symbols report respectively the 10th, 25th, 50th, 75th, and 90th percentiles. The blue dashed line represents the average search time of a completely passive particle. The red dashed line provides a one-parameter fit of the learning curve to a stretched exponential behavior. The constants C and D are respectively equal to the average time to reach the target at the 10^3-th episode and to the difference between the values at the 1-st and 10^3-th episode of the same observable. Inset: Time required to reach the target at the 10^3-th episode as a function of the Péclet number for $\ell^* = 1$. The hyperparameters of the PS algorithm for each Péclet number are reported in the Methods section.

maximize its total reward, the agent learns to minimize the target-search time, *i.e.* to optimize the target-search efficiency.

In the following, we fix the length unit as the size of the box L and the time unit as the typical time $\tau := L^2/4D$ required by a passive particle to cover this distance. The model has thus two free dimensionless parameters: The Péclet number $\mathrm{Pe} := v\tau/L$, measuring the magnitude of the activity, and the persistence $\ell^* := v/D_9L$, representing the persistence of directed motion in the ABP phase. The size and the speed of natural and artificial self-propelled microswimmers range over a few orders of magnitude.[50] Furthermore, also the environmental conditions such as the size of the search space and of the targets, as well as the target density and shape can be very variegated. Nevertheless, one can translate our model units to physical units when considering a typical experiment having a box size $L = 10^2$ μm and Janus particles with diffusion coefficient $D = 2\ \mu\mathrm{m}^2\ \mathrm{s}^{-1}$, resulting in a $\tau = 1.25 \times 10^3$ s.

3 Results

Resorting on the RL algorithm projective simulation[54] (PS) described in the Methods section, the learning performances of our agents are evaluated, for each set of free parameters, by checking how the average time to reach the target (also known as the mean first-passage time[55,56]) evolves during subsequent episodes of duration 20τ. The average is performed over N independent agents, all with the same initial policy in which, independently of the current phase duration, the probabilities of switching phase are 10^{-2} and 10^{-3} when being in the passive and in the active phase respectively. Such initial policy is purposely chosen to have particles with a rather poor initial target-search efficiency, with an average searching time at least twice that of a pure passive particle, the latter being about 1.14τ.

We first consider the case in which the learning particle, when in the ABP phase, has a large activity and a persistence length equal to the box size. To do so, we set the persistence to $\ell^* = 1$ and the Péclet number to Pe = 100 which means that the ratio between the typical length traveled because of the self-propulsion and the typical length traveled due to diffusion is 1 at the minimal phase duration, corresponding to the integration time step $\Delta t = 10^{-4}\tau$, and grows up to 100 for a phase duration equal to the time unit τ (see Methods section for details). In such a situation, the learning particle outperforms the target-search performances of a purely passive particle already after two episodes, see Fig. 1. During subsequent episodes, the average time required to find the target keeps decreasing, following a stretched exponential behavior and after 10^3 episodes it is about 4 times smaller than the benchmark value corresponding to the fully passive particle. Furthermore, also the spread of the average search times among the N different agents decreases during the learning process, with the difference between the first quartile and the third quartile reducing from about 1.4τ to about 0.1τ along the 10^3 episodes, see Fig. 1. The observed stretched exponential learning shows that convergence to the optimal policy does not have a constant rate during the learning process but that this rate decreases with increasing number of episodes. We are unable to explain this behavior starting from a closer inspection of the considered environment and/or of the PS algorithm. On the other hand, we also stress that the exact value of the stretching exponent β has no particular importance and that it depends on the parameters of both the model and the PS algorithm.

An important question is how the target-search efficiency depends on the activity of the particle. To address this issue, we investigate how the average time to reach the target during the 10^3-th episode varies when changing the Péclet number. This is

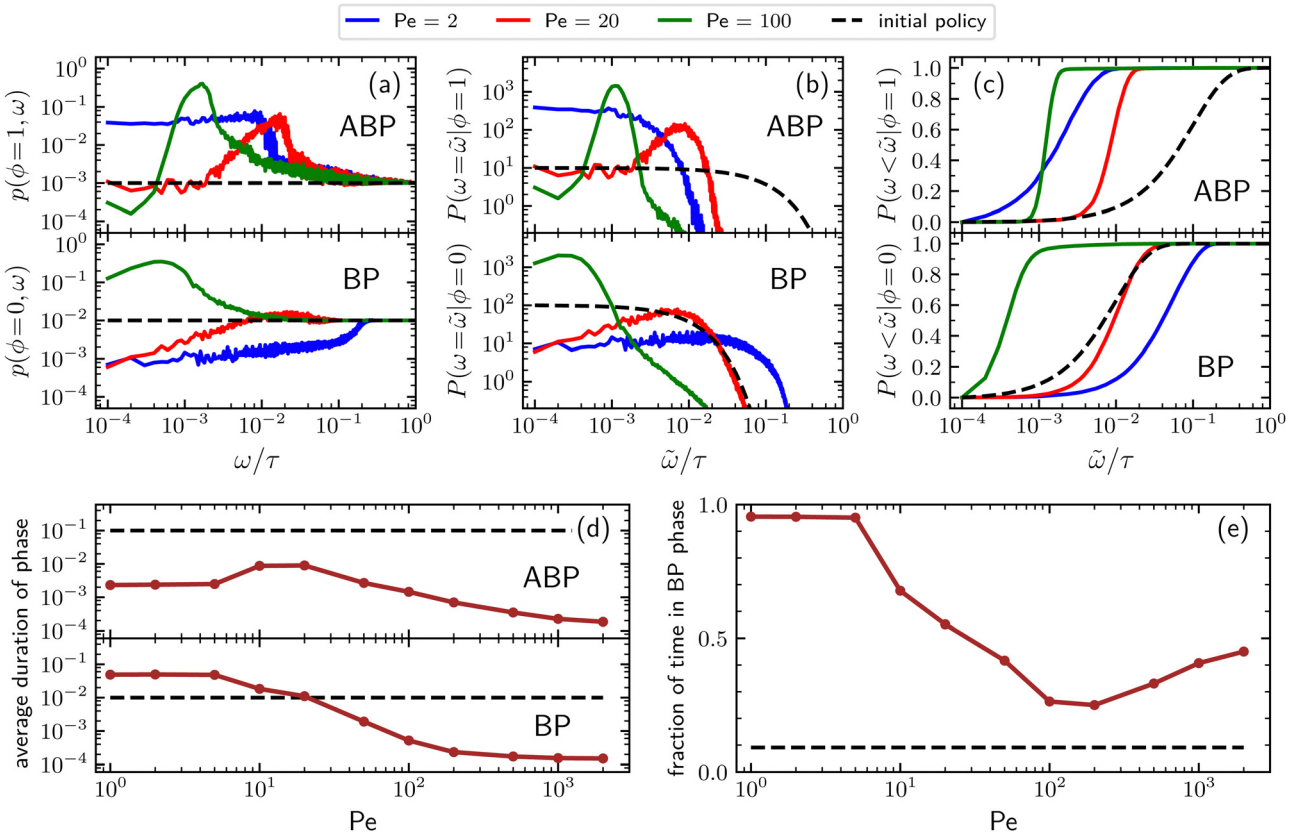

Fig. 2 (a) Probabilities of switching from BP to ABP motion (lower panel) and from ABP to BP motion (upper panel) as a function of the phase duration and for different Péclet numbers. Data are obtained after 10^3 episodes and averaged over $N = 5 \times 10^3$ independent particles; (b) distribution of phase duration for different Péclet numbers (BP phase, lower panel – ABP phase upper panel); (c) cumulative distribution of the phase duration for different Péclet numbers; (d) average duration of BP and ABP phases as a function of the activity; (e) average fraction of time spent in the passive phase as a function of the activity. In all panels, the black dashed line represents the corresponding observable as obtained from the initial policy, which is independent of the Péclet number.

reported in the inset of Fig. 1, which shows that the learning particle has performances comparable to those of a passive particle as long as the Péclet number is smaller than about Pe $\approx$ 10 and then the average time to reach the target decreases with increasing activity until it reaches a plateau for Pe $\lessapprox$ 200. Such a phenomenology is consistent with the results already found in ref. 49 and is intuitively understood as follows: Since the typical distance covered by pure diffusion grows with time as $t^{1/2}$ while the one due to the self-propulsion grows about as t, for small activities, diffusion dominates the relocation process during the short phases. On the other hand, having long active phases is not favorable because of the particle's inability to find the target when in the ABP phase. Thus, at low Péclet numbers, the learning particle tries to maximize the time spent in the passive phase, with the resulting performances equivalent to those of a purely passive particle. In contrast, at large Péclet the self-propulsion velocity is enough to allow relocation at a distance larger than the target size even for very short active phases and the better performances of the learning particle are in accordance with the idea that having an intermittent-search strategy is more efficient than having a simple diffusive process.

Additional insight into how the PS algorithm encodes learning successful strategies can be gained by directly investigating the policy, *i.e.* the probabilities of switching phase given the state. This is done in Fig. 2 which reports these switching probabilities (panel a) and related observables as learned after 10^3 episodes for Pe = 2, 20, and 100, respectively corresponding to a low, intermediate, and high value of the activity. Among the related observables, we report the probability of having a phase with a certain duration $\tilde{\omega}$ conditioned to being in phase ϕ, $P(\omega = \tilde{\omega}|\phi)$ (Fig. 2b) and the cumulative probability of having a phase duration shorter than $\tilde{\omega}$ conditioned to being in the phase ϕ, $P(\omega < \tilde{\omega}|\phi)$ (Fig. 2c). These quantities can be obtained directly from the switching probabilities $p(s) = p(\phi,\omega)$ related to a given state $s = (\phi,\omega)$ as

$$P(\omega = \tilde{\omega}|\phi) = \frac{1}{\Delta t}\prod_{n'=1}^{\tilde{n}-1}[1 - p(\phi, n'\Delta t)]p(\phi, \tilde{n}\Delta t), \qquad (4)$$

where we discretized the time introducing the integer variable $n = \omega/\Delta t$ and the factor $1/\Delta t$ in front of the right-hand side accounts for the correct normalization, and

$$P(\omega < \tilde{\omega}|\phi) = \sum_{n'=1}^{\tilde{n}-1} P(\omega' = n'\Delta t|\phi)\Delta t\ . \qquad (5)$$

Further observables reported in Fig. 2 are the average duration of a phase (panel d) and the fraction of time spent

in the passive phase (panel e) as a function of the Péclet number. However, before discussing the details of the learned policies, it is important to clarify that, for large enough ω's, the value of the switching probability $p(s)$ always drops to the corresponding value in the initial (arbitrary) policy. In fact, the longer the phase duration ω of a given state $s = (\phi,\omega)$, the more rarely this state is visited during the learning process, with the frequency of these visits depending on the switching probabilities associated with the states $s = (\phi,\omega')$ having the same phase ϕ and lower phase duration $\omega' < \omega$. This results in practical limitation in sampling states with a large phase duration. In spite of this issue, the target-search abilities of the trained agent are not affected since the rarer it is to visit a given state, the smaller the contribution of the action following that state to the overall performances of the particle.

For low activity (Pe = 2), the probability of switching from the passive to the active phase decreases from the value 10^{-2} corresponding to the initial policy to a value of about 10^{-3} for very short phase duration. Such a probability increases with a power-law behavior for about three decades, until it quickly converges to the initial policy value for a duration of the phase larger than about $10^{-1}\tau$ (see Fig. 2a, lower panel). On the other hand, the probability of switching from the active to the passive phase (Fig. 2a, upper panel) increases from the initial policy's value 10^{-3} to about 4×10^{-2} and drops to the initial policy for a duration of the phase larger than about $10^{-2}\tau$. These results, together with the corresponding ones in panels b and c, indicate that, for low activity, the trained particle prefers to alternate relatively long passive phases with short active ones, confirming the previously mentioned expectations. For Pe = 20, the probability of having a phase with a certain duration ω of a given phase shows a peak at around $\omega = 10^{-2}\tau$ both when conditioned to be in the ABP phase ($\phi = 1$) and in the BP one ($\phi = 0$), see Fig. 2b. Consequently, for this value of the activity, the best strategy consists of alternating between active and passive phases both having a typical duration of about $10^{-2}\tau$ (see also Fig. 2d), with the duration of the passive phase having a larger variance as indicated by the fact that the peak of the distribution conditioned to being in the active phase is narrower than the one of the distribution in the passive phase. We stress that, because of its self-propulsion, an ABP with Pe = 20 and $\ell^* = 1$ in a time interval of $10^{-2}\tau$ covers a typical distance of about $0.2L$ which is twice the target diameter. Finally, for Pe = 100, Fig. 2 shows that the distribution of phase durations displays a sharp peak at around $10^{-3}\tau$ for the ABP phase and a rather broad peak at a few integration time steps for the BP phase. Concomitantly, the learned strategy alternates between very short active phases with an average duration of about $1.4 \times 10^{-3}\tau$ and even shorter passive phases lasting about $0.5 \times 10^{-3}\tau$ on average. In this case, the typical distance traveled during the active phase because of the particle's self-propulsion is about $0.14L$ which is of the same order as the one registered in the case of Pe = 20 even though the activity is now 5 times larger.

It is interesting to note that, as reported in Fig. 2d, while the average passive phase duration monotonically decreases with the Péclet number, its counterpart for the active phase has a non-monotonic behavior that can be rationalized as follows: Both at large and low values of the activity the ABP phases are very short but for two different reasons. At low activity, these are short because active relocation to a distance greater than the target size would require too much time and the agent responds by minimizing the time spent in this phase. In contrast, for large activity, very short active phases are already sufficient to allow the particle to relocate elsewhere in the simulation box and improve the target-search performances of the smart particle. For intermediate Péclet numbers, the agent instead finds an optimal duration of the active phase reflecting the compromise between the utility of active phases for quick relocation and the fact that during these phases the target cannot be detected. The effect of the monotonic and non-monotonic behaviors of the average duration of respectively the

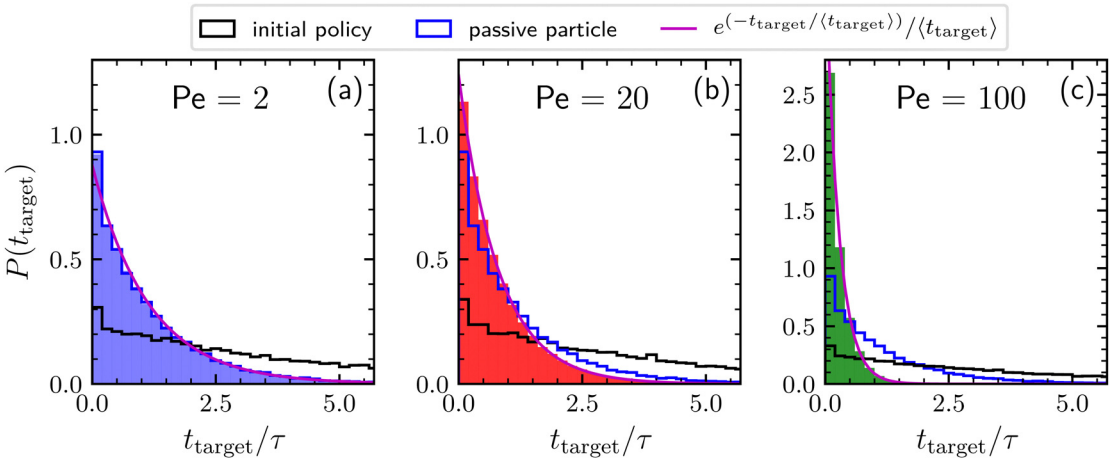

Fig. 3 Distribution of times needed to find the target collected during a time interval of length $10^6\tau$, for Pe = 2, 20, and 100 (panels a, b, and c respectively). We consider here agents behaving according to the policies learned after 10^3 episodes as reported in Fig. 2 (solid bars), to the initial policy prior to learning (black line), and completely passive particles (blue line). The magenta line is the exponential distribution having decay time given by the average time to find the target.

passive and the active phase, also results in a non-monotonic behavior of the fraction of time spent in the passive phase. In fact, this quantity is close to 1 for small Péclet numbers, decreases to about 0.25 for Pe $\approx$ 100, and then increases again to 0.5 which is the value expected for extremely large levels of activity, see Fig. 2e.

Finally, Fig. 3 shows the distributions of times needed to find the target by agents adopting the learned policies previously discussed for Pe = 2, 20, and 100. These distributions are exponential, meaning that the kinetics is completely characterized by the mean first-passage time. For a comparison, the distributions obtained by a searching particle adopting the initial policy and by a completely passive particle are also reported. Concerning the results obtained by adopting the initial policy, note that, even if the policy remains the same, in principle the resulting distribution depends also on the value of the activity. However, this dependence appears to be very weak, as revealed by the similar behavior of the distribution corresponding to the three different Péclet numbers. For low activity (Pe = 2) the distribution of the searching times related to the learned policy is very similar to that of a simple BP, confirming the passive-like behavior of the agents in this Péclet regime. As expected, increasing the activity, the distribution for the optimized particle becomes more and more narrow. In particular, for Pe = 100, the large majority of targets is found within the unit time τ.

4 Conclusions

In summary, we have introduced a RL-based method to probe the potential of smart microswimmers in target-search problems involving targets of unknown positions in a homogeneous environment. Specifically, we applied a Projective Simulation based approach[48] to micro-particles able to perform either passive or active Brownian motion and to switch from one to the other on the basis of a probabilistic policy, the latter being necessary to reproduce a stochastic strategy. Our findings demonstrate that, during repeated learning episodes, the agent optimizes its target-search performance and that the optimal policy strongly depends on the magnitude of the self-propulsion during the active phase. For low activity, the behavior of the smart particle is similar to that of a completely passive particle while, for large activity, the agent takes advantage of the active phases to quickly cover more ground and increase the target-finding odds. More in detail, the duration of the passive phases decreases with increasing Péclet number while the duration of the active phases displays a non-monotonic behavior with a maximum at intermediate Péclet numbers. The proposed model is inspired by the intermittent search strategies developed by Bénichou and coworkers.[1,24,25] In this framework, an exponential distribution of phases is assumed to allow analytical tractability and in agreement with some experimental observations.[1] However, in our case, the duration of a phase is part of the state sensed by the agent, meaning that the agent is endowed with some sort of temporal memory. Consequently, distributions of phases different from the exponential one may arise, which is indeed what is observed in the learned policies. Our results complement and extend those of a previous study based on a genetic algorithm[49] and demonstrate that also reinforcement learning is a powerful tool to investigate target-search problems for agents undergoing a stochastic dynamics.

With respect to previous literature on stochastic target search,[1,21–26,48] which mainly applies to generic scenarios, our investigation is more focused on the microscopic world, namely we are interested in natural or artificial microswimmers. This is the main reason to resort to the active Brownian particle model. In fact, this model, besides being the paradigmatic model in the framework of non-equilibrium dynamics,[57–60] also provides a faithful representation of the behavior of artificial microswimmers such as the Janus particles.[50] Remarkably, nowadays it is already possible to perform experiments in which the activity of artificial microswimmers is controlled by an external illuminating system.[38] Thus, the target-search strategies developed in the present manuscript can potentially be tested in a laboratory. Furthermore, the *intermittent* active Brownian dynamics that we introduce in the Model section can be also considered, in the case of relatively large activity and persistence, as a first proxy for the run-and-tumble dynamics which is the typical theoretical model describing the motion of bacteria.[10,50,61]

The proposed framework offers new insight into target-search problems in homogeneous enviroments and paves the way to further research. In particular, it can be leveraged to explore more complex scenarios such as, for instance, target search with resetting events,[62–64] multiple and/or motile targets problems,[1] or searchers with multiple migration modes, the latter being relevant to dendritic cells searching for infections.[65] Moreover, other possible developments, particularly relevant for the envisioned medical and environmental application of smart active particles, entail heterogeneous environments involving the presence of obstacles, boundaries, and energy barriers.[29,66–68] Finally, endowing the agent with a limited memory of the recently visited locations[69] or with the ability to sense directional cues coming from the target itself, may also be an extension going in the direction of better modelling biological microswimmers.

5 Methods

To identify effective target-search strategies, we used the RL algorithm projective simulation (PS), which was originally created as a platform for the design of autonomous quantum learning agents[54] and was shown to have competitive performance also in classical RL problems.[70,71]

The core idea of this algorithm is to use the notion of a particular kind of memory, called *episodic and compositional memory* (ECM) which is mathematically described by a graph connecting units called *clips*. Clips can be either percept or decision units, corresponding to states and actions respectively,

or a combination of those. We design our target-search problem as a Markov decision process,[35] *i.e.* at each learning step, the agent is in some state s, takes an action a according to a policy defined by the conditional probabilities $\pi(a|s)$, and receives a reward $\mathscr{R}$ as a consequence of this action. In such a case, the ECM structure consists of a layer of states fully connected with a layer of actions. Each edge of the graph, *i.e.* each state-action pair (s,a), is assigned with a real-value weight $h(s,a)$, called the h-value, which determines the policy according to

$$\pi(a|s) = \frac{h(s,a)}{\sum_{a' \in \mathscr{A}} h(s,a')}, \tag{6}$$

where $\mathscr{A}$ represents the set of all possible actions. Furthermore, a non-negative glow value $g(s,a)$ stores the information on which and, implicitly, how frequently state-action pairs have been visited during the learning process. Such information is then exploited when updating the policy with the goal of maximizing the total expected reward.

This last feature of the PS algorithm makes it particularly apt to solve our target-search problem. The reason is that, on average, the equations of motion (1–3) have to be iterated a large number of times before a target is found and the agent obtains its reward. Consequently, the reward signal is very sparse and has only a very low correlation with the particular state-action pair encountered when the target is found. Approaches taking into account long sequences of visited state-action pairs, as the PS algorithm, should then be preferred with respect to typical action-value methods such as one-step Q-learning or SARSA.[35] Indeed, we failed to obtain successful target-search policies when applying standard Q-learning to our setup.

Applying the PS framework to the model illustrated in the dedicated section and taking into consideration that in our case the action a can be described as a binary variable, with $a = 1$ corresponding to a switch of the phase (passive or directed motion) and $a = 0$ to maintaining the current phase, a single learning step consists of the following operation:

- Given the current state $s_t = (\phi_t,\omega_t)$, the probability of switching phase p_t is determined as

$$p_t = \pi(a_t = 1|s_t) = h(s_t,1)/[h(s_t,0) + h(s_t,1)]$$

and the next phase ϕ_{t+1} is selected accordingly;
- The glow matrix is damped following the update rule $G \leftarrow (1-\eta)G$, where η is called the glow parameter and determines how much a delayed reward should be discounted;
- The glow matrix is updated by adding a unit to the visited state-action pair, $g(s_t,a_t) \leftarrow g(s_t,a_t) + 1$;
- The position and the direction of the particle evolve according to eqn (2) and (3);
- The whole matrix of h-values is updated according to the learning rule of the PS model, $H \leftarrow (1-\gamma)H + \gamma H_0 + \mathscr{R}G$, where $\mathscr{R}$ is the reward being zero if no target is found by the particle located at the updated position and 1 otherwise. Here, γ is called the damping parameter and specifies how quickly the H matrix returns to an initial matrix H_0.

Table 1 Hyperparameters used to obtain the results presented in the present work

Pe	≤ 5	10	20	50	≥ 100
γ	10^{-7}	10^{-6}	10^{-6}	10^{-6}	10^{-5}
η	10^{-2}	10^{-3}	10^{-3}	10^{-2}	10^{-2}

Note that the reward $\mathscr{R}$ is different from zero only when the target is found, it is thus possible to optimize the computational costs by iterating the single-step update rule of the H matrix and updating this matrix as a whole only when the target is found. More in details, if a target is found at time step t_2 and the previous target was found at time step t_1, then the update rule of the whole H matrix at time t_2 (given that the last time it was updated was at the end of step t_1) is $H \leftarrow (1-\gamma)^{t_2-t_1}H + \gamma\left[\sum_{t'=0}^{t_2-t_1-1}(1-\gamma)^{t'}\right]H_0 + \mathscr{R}G$. In doing so, one has to pay attention that, during the learning steps t with $t_1 < t \leq t_2$, the switching probability p_t has to be determined according to $p_t = \tilde{h}(s_t,1)/[\tilde{h}(s_t,0) + \tilde{h}(s_t,1)]$, where $\tilde{h}(s_t,a_t)$ is a temporarily updated h-value obtained as $\tilde{h}(s_t,a_t) = (1-\gamma)^{t-t_1-1}h(s_t,a_t) + \gamma\left[\sum_{t'=0}^{t-t_1-2}(1-\gamma)^{t'}\right]h_0(s_t,a_t)$, with $h_0(s_t,a_t)$ the element of the initial matrix H_0 corresponding to state s_t and action a_t.

The initial policy is such that the probabilities of switching phase are 10^{-2} and 10^{-3} when being in the passive and in the active phase respectively. This is obtained by setting, for each t, $h_0(s_t,a_t = 1) = 10^{-2}$ and $h_0(s_t,a_t = 0) = (1 - 10^{-2})$ if the state is in a passive phase, and $h_0(s_t,a_t = 1) = 10^{-3}$ and $h_0(s_t,a_t = 0) = (1 - 10^{-3})$ if the state is in an active phase. All the terms of the G matrix are initialized to zero at the beginning of each episode.

We set the integration time step to $\Delta t = 10^{-4}\tau$ and, to have a finite set of states, we limit the duration of a given phase ω to be not longer than τ. This results in a total of 2×10^4 states $s_t = (\phi_t,\omega_t)$, being $\phi_t = 0,1$ (see Model section) and $\omega_t = 1,\ldots,10^4$. The glow and the damping parameters are considered hyperparameters of the model and, for each value of the activity Pe and of the persistence ℓ^*, are adjusted to obtain the best learning performances. Their values are reported in Table 1. Finally, to investigate how the learning process evolves, we split the whole process into several episodes, each lasting 20τ. At the beginning of each episode, each element of the glow matrix is initialized to zero.

Author contributions

M. C. developed the software and analyzed the results. All authors conceived the research and wrote and reviewed the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

H. K. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 847476; M. C. is supported by FWF: P 35872-N; T. F. acknowledges funding by FWF: P 35580-N; A. L. and H. J. B. acknowledge support by the Volkswagen Foundation (Az: 97721); H. J. B. acknowledges funding from FWF through SFB BeyondC F7102, and the European Research Council (ERC, Quant AI, Project No. 10105529). G. M.-G also acknowledges funding from the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Council or the European Research Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them. This research was funded in whole or in part by the Austrian Science Fund (FWF) [P 35872-N; P 35580-N; SFB BeyondC F7102]. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

Notes and references

1 O. Bénichou, C. Loverdo, M. Moreau and R. Voituriez, *Rev. Mod. Phys.*, 2011, **83**, 81.
2 E. L. Charnov, *Theor. Popul. Biol.*, 1976, **9**, 129.
3 W. J. O'brien, H. I. Browman and B. I. Evans, *Am. Sci.*, 1990, **78**, 152.
4 D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell and D. Morritt, *Nature*, 2008, **451**, 1098.
5 D. L. Kramer and R. L. McLaughlin, *Am. Zool.*, 2015, **41**, 137.
6 J. R. Frost and L. D. Stone, *Review of Search Theory: Advances and Applications to Search and Rescue Decision Support*, U.S. Coast Guard Research and Development Center, 2001.
7 O. G. Berg, R. B. Winter and P. H. Von Hippel, *Biochem.*, 1981, **20**, 6929.
8 J. Gorman and E. C. Greene, *Nat. Struct. Mol. Biol.*, 2008, **15**, 768.
9 J. Elgeti and R. Winkler, *Rep. Prog. Phys.*, 2015, **78**, 056601.
10 H. Berg, *E. coli in Motion*, Springer-Verlag, Heidelberg, 2004.
11 P. N. Devreotes and S. H. Zigmond, *Annu. Rev. Cell Biol.*, 1988, **4**, 649.
12 S. de Oliveira, E. E. Rosowski and A. Huttenlocher, *Nat. Rev. Immunol.*, 2016, **16**, 378.
13 M. Eisenbach and L. C. Giojalas, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 276.
14 M. J. Smanski, H. Zhou, J. Claesen, B. Shen, M. A. Fischbach and C. A. Voigt, *Nat. Rev. Microbiol.*, 2016, **14**, 135.
15 M. You, C. Chen, L. Xu, F. Mou and J. Guan, *Acc. Chem. Res.*, 2018, **51**, 3006.
16 S. Klumpp, C. T. Lefévre, M. Bennet and D. Faivre, *Phys. Rep.*, 2019, **789**, 1.
17 M. Medina-Sánchez, L. Schwarz, A. K. Meyer, F. Hebenstreit and O. G. Schmidt, *Nano Lett.*, 2016, **16**, 555.
18 D. Patra, S. Sengupta, W. Duan, H. Zhang, R. Pavlick and A. Sen, *Nanoscale*, 2013, **5**, 1273.
19 S. Naahidi, M. Jafari, F. Edalat, K. Raymond, A. Khademhosseini and P. Chen, *J. Control. Release*, 2013, **166**, 182.
20 W. Gao and J. Wang, *ACS Nano*, 2014, **8**, 3170.
21 G. M. Viswanathan, M. G. Da Luz, E. Raposo and H. Stanley, *The Physics of foraging: An introduction to random searches and biological encounters*, Cambridge University Press, 2011.
22 G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. E. da Luz, E. P. Raposo and H. E. Stanley, *Nature*, 1999, **401**, 911.
23 G. Viswanathan, E. Raposo and M. da Luz, *Phys. Life Rev.*, 2008, **5**, 133.
24 O. Bénichou, M. Coppey, M. Moreau, P.-H. Suet and R. Voituriez, *Phys. Rev. Lett.*, 2005, **94**, 198101.
25 O. Bénichou, C. Loverdo, M. Moreau and R. Voituriez, *Phys. Rev. E*, 2006, **74**, 020102.
26 C. Loverdo, O. Bénichou, M. Moreau and R. Voituriez, *Phys. Rev. E*, 2009, **80**, 031146.
27 M. C. Santos, E. P. Raposo, G. M. Viswanathan and M. G. E. da Luz, *EPL*, 2004, **67**, 734.
28 F. Bartumeus, J. Catalan, U. L. Fulco, M. L. Lyra and G. M. Viswanathan, *Phys. Rev. Lett.*, 2002, **88**, 097901.
29 G. Volpe and G. Volpe, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 11350.
30 O. Bénichou, C. Loverdo, M. Moreau and R. Voituriez, *J. Phys.: Condens. Matter*, 2007, **19**, 065141.
31 M. A. Lomholt, K. Tal, R. Metzler and K. Joseph, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 11055.
32 S. Benhamou, *J. Theor. Biol.*, 1992, **159**, 67.
33 M. Moreau, O. Bénichou, C. Loverdo and R. Voituriez, *J. Stat. Mech. Theory Exp.*, 2009, P12006.
34 F. Cichos, K. Gustavsson, B. Mehlig and G. Volpe, *Nat. Mach. Intell.*, 2020, **2**, 94.
35 R. S. Sutton and A. G. Barto, *Reinforcement Learning*, The MIT Press, 2nd edn, 2018.
36 M. Mitchell, *An introduction to Genetic Algorithms*, The MIT Press, 1998.
37 E. Schneider and H. Stark, *EPL*, 2019, **127**, 64003.
38 S. Muiños-Landin, A. Fischer, V. Holubec and F. Cichos, *Sci. Robot.*, 2021, **6**, eabd9285.
39 A. C. H. Tsang, P. W. Tong, S. Nallan and O. S. Pak, *Phys. Rev. Fluids*, 2020, **5**, 074101.
40 B. Hartl, M. Hübl, G. Kahl and A. Zöttl, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2019683118.
41 P. A. Monderkamp, F. J. Schwarzendahl, M. A. Klatt and H. Löwen, *Mach. learn.: sci. technol.*, 2022, **3**, 045024.
42 S. Colabrese, K. Gustavsson, A. Celani and L. Biferale, *Phys. Rev. Lett.*, 2017, **118**, 158004.
43 K. Gustavsson, L. Biferale, A. Celani and S. Colabrese, *Eur. Phys. J. E*, 2017, **40**, 110.
44 S. Colabrese, K. Gustavsson, A. Celani and L. Biferale, *Phys. Rev. Fluids*, 2018, **3**, 084301.
45 L. Biferale, F. Bonaccorso, M. Buzzicotti, P. Clark Di Leoni and K. Gustavsson, *Chaos*, 2019, **29**, 103138.

46 J. K. Alageshan, A. K. Verma, J. Bec and R. Pandit, *Phys. Rev. E*, 2020, **101**, 043110.
47 A. C. H. Tsang, E. Demir, Y. Ding and O. S. Pak, *Adv. Intell. Syst.*, 2020, **2**, 1900137.
48 G. Muñoz-Gil, A. López-Incera, L. J. Fiderer and H. J. Briegel, *New J. Phys.*, 2024, **26**, 013010.
49 H. Kaur, T. Franosch and M. Caraglio, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 035008.
50 C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe and G. Volpe, *Rev. Mod. Phys.*, 2016, **88**, 045006.
51 K. O. Stanley and R. Miikkulainen, *Evol. Comput.*, 2002, **10**, 99.
52 J. R. Howse, R. A. L. Jones, A. J. Ryan, T. Gough, R. Vafabakhsh and R. Golestanian, *Phys. Rev. Lett.*, 2007, **99**, 048102.
53 H.-R. Jiang, N. Yoshinaga and M. Sano, *Phys. Rev. Lett.*, 2010, **105**, 268302.
54 H. J. Briegel and G. De Las Cuevas, *Sci. Rep.*, 2012, **2**, 400.
55 S. Redner, *A Guide to First-Passage Processes*, Cambridge University Press, 2001.
56 *First-Passage Phenomena and Their Applications*, ed. R. Metzler, G. Oshanin and S. Redner, World Scientific, Singapore, 2013.
57 M. E. Cates, *Rep. Prog. Phys.*, 2012, **75**, 042601.
58 É. Fodor, C. Nardini, M. E. Cates, J. Tailleur, P. Visco and F. van Wijland, *Phys. Rev. Lett.*, 2016, **117**, 038103.
59 É. Fodor and M. C. Marchetti, *Physica A Stat. Mech. Appl.*, 2018, **504**, 106.
60 M. Caraglio and T. Franosch, *Phys. Rev. Lett.*, 2022, **129**, 158001.
61 I. Santra, U. Basu and S. Sabhapandit, *Phys. Rev. E*, 2020, **101**, 062120.
62 M. R. Evans and S. N. Majumdar, *Phys. Rev. Lett.*, 2011, **106**, 160601.
63 L. Kusmierz, S. N. Majumdar, S. Sabhapandit and G. Schehr, *Phys. Rev. Lett.*, 2014, **113**, 220602.
64 V. Kumar, O. Sadekar and U. Basu, *Phys. Rev. E*, 2020, **102**, 052129.
65 T. Song, Y. Choi, J.-H. Jeon and Y.-K. Cho, *Front. Immunol.*, 2023, **14**, 1129600.
66 L. Zanovello, M. Caraglio, T. Franosch and P. Faccioli, *Phys. Rev. Lett.*, 2021, **126**, 018001.
67 L. Zanovello, P. Faccioli, T. Franosch and M. Caraglio, *J. Chem. Phys.*, 2021, **155**, 084901.
68 L. Zanovello, R. J. G. Löffler, M. Caraglio, T. Franosch, M. M. Hanczyc and P. Faccioli, *Sci. Rep.*, 2023, **13**, 5616.
69 H. Meyer and H. Rieger, *Phys. Rev. Lett.*, 2021, **127**, 070601.
70 J. Mautner, A. Makmal, D. Manzano, M. Tiersch and H. J. Briegel, *New Gener. Comput.*, 2015, **33**, 69.
71 W. L. Boyajian, J. Clausen, L. M. Trenkwalder, V. Dunjko and H. J. Briegel, *Quantum Mach. Intell.*, 2020, **2**, 13.