# Implementation of reinforcement learning in chemical reaction networks: application to phototaxis as curiosity-driven exploration

Ruyi Tang, Grégoire Sergeant-Perthuis, David Colliaux

# Implementation of reinforcement learning in chemical reaction networks: application to phototaxis as curiosity-driven exploration

Ruyi Tang[1], Grégoire Sergeant-Perthuis[1], and David Colliaux[2]

[1]CQSB, Sorbonne University, France
[2]Sony CSL, France
evatang2002@sina.com

## Abstract

Living systems navigate environments using noisy and incomplete sensory signals. In unicellular algae, phototaxis is often modeled as a mechanistic run–tumble process driven by stimulus–response rules. However, such descriptions overlook how organisms actively sample their environment to reduce sensory ambiguity. From a minimal cognition perspective, we reframe this navigation as a subjective, information-driven sensorimotor process. To this end, we propose a framework linking a Partially Observable Markov Decision Process (POMDP) with biochemical reaction dynamics. Environmental variables are hidden, while the cell maintains a minimal internal state updated from the current observation through a memoryless Bayesian reweighting step. These internal dynamics balance orienting toward light with exploratory reorientation and can be implemented through Chemical-Reaction-Network Ordinary Differential Equations (CRN–ODEs), showing how biochemical processes can physically realize the required information-processing mechanisms. Our model includes a biophysical observation process for photoreception and a chemically computable polynomial bound on information gain. Using Inverse Reinforcement Learning (IRL) on 30 experimentally recorded Chlamydomonas trajectories, we infer the behavioral objective consistent with observed phototactic motion and benchmark the resulting dynamics with standard Stochastic Simulation Algorithm (SSA) baselines. Our model reproduces the empirical distribution of alignment with the light source and achieves alignment statistics comparable to objective SSA baselines on this dataset. Within this framework, run–tumble alternation emerges as an information-acquisition strategy: tumbling reorients the cell to sample new sensory configurations and resolve sensor ambiguity, demonstrating how intracellular biochemical networks can support adaptive information-seeking behavior in cellular navigation.

Submission type: **Full Paper**

Data/Code available at: `https://github.com/giveyourselfaTRY/phototaxis-pomdp-crn`



## Introduction

Across biological scales—from insects tracking turbulent plumes ( Heinonen et al. (2025)) to bacteria climbing microscopic gradients ( Auconi et al. (2022))—organisms navigate complex environments using noisy and incomplete sensory cues. In minimal organisms, such navigation emerges without neural systems, raising a central question in artificial life and minimal cognition: how can adaptive behavior arise from simple biochemical dynamics interacting with uncertain environments?

In this paper, we investigate whether tumbling in phototaxis can be interpreted as an information-seeking response to sensory ambiguity, complementing standard stimulus-response descriptions

Recent work suggests that cellular signaling networks can function as distributed information-processing systems ( Colliaux et al. (2017)). At the molecular level, biochemical reaction networks are capable of implementing probabilistic computations, including Bayesian-like inference, through populations of interacting molecules and diffusible messengers ( Wiuf et al. (2023)). These results point toward a possible bridge between statistical models of inference and the physical dynamics of intracellular chemistry. At the behavioral scale, navigation can also be interpreted as an information-driven sensorimotor process, where organisms actively sample their environment to reduce uncertainty—a principle closely related to active sensing ( Gottlieb and Oudeyer (2018)). To mathematically formalize this mechanism, Reinforcement Learning (RL) offers a powerful paradigm. Specifically, the concept of *curiosity* in RL is formally defined as an intrinsic reward signal proportional to the expected information gain (or reduction in uncertainty) about the environment's latent states Pathak et al. (2017). However, how such curiosity-driven, information-seeking dynamics could arise from biochemically realizable mechanisms remains largely unexplored.

Here we study algal phototaxis as an adaptive curiosity-driven decision-making process under partial observability. We formulate the navigation problem as a Partially Observable Markov Decision Process (POMDP) and demonstrate

that its internal inference and decision dynamics can be implemented using modular Chemical-Reaction-Network Ordinary Differential equations (CRN–ODEs). This connects RL prescribed policies and dynamics with potential biochemical processes.

## Contributions

We analyze a novel dataset of 30 experimentally recorded *Chlamydomonas* trajectories and fit a subjective POMDP using maximum-likelihood learning and inverse reinforcement learning. The resulting learned dynamics reproduce empirical alignment–to–light distributions with high fidelity and match the performance of standard mechanistic baselines on this dataset. We then propose the first complete chemical reaction network–ordinary differential equation (CRN–ODE) model that implements a POMDP policy from sensory activation through flagellar motor behavior, balancing exploration and reward. We demonstrate that the steady state of this CRN–ODE closely replicates both the policies and the dynamics learned by the POMDP. We show that tumbling behavior helps the organism actively resolve the front-back ambiguity created by the side-located eyespot, providing the first quantitative explanation of how tumbling supports phototaxis rather than acting as stochastic motor noise.

Together, these results suggest that intracellular biochemical networks can support adaptive information-seeking behavior, linking cellular chemistry, RL information processing, and minimal cognition in microbial navigation.

# From objective to subjective probabilistic model for phototaxis

## An objective model for mechanistic run-and-tumble motion.

To establish a behavioral baseline, we first formulate an idealized "objective" model of phototaxis, implemented via the standard Stochastic Simulation Algorithm (SSA) ( Gillespie (1977)). Modeling run-and-tumble navigation as a stochastic velocity-jump process governed by environment-dependent transition rates is a widely validated in theoretical biology ( Erban and Othmer (2004); Tindall et al. (2008); Polin et al. (2009)). This mechanistic framework adopts an omniscient observer's perspective, assuming the dynamics are driven strictly by exact, global physical quantities.

As an objective model, we consider the organism to change orientation at random times with a rate increasing with the misalignment of its motion relative to the light source. We suppose the cell swims in a straight line in-between those tumbling events. Specifically, at any time $t$, the agent's spatial state is defined by its position $x_t$ and heading vector $\hat{u}_t$. The objective misalignment angle $\theta$ is derived from the true alignment cosine between the heading and the absolute light source $\mu$: $\cos\theta = \hat{u}_t \cdot (\mu - x_t)/\|\mu - x_t\|$.

Following the classical instantaneous reorientation framework, the time interval $\Delta t$ between two tumbling events is drawn from an exponential distribution $\Delta t \sim \text{Exp}(\lambda(\theta))$. To capture the mechanistic intuition that tumbling probability increases as the agent deviates from the target, we propose the following cosine-based penalty function for the tumbling rate:

$$\lambda(\theta) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \left( \frac{1 - \cos\theta}{2} \right)^{\gamma} \quad (1)$$

During the drawn interval $\Delta t$, the agent executes a *run advance*. Its position is explicitly updated via the kinematic equation $x_{t+\Delta t} = x_t + c_{\text{run}} \hat{u}_t \Delta t$, while its heading $\hat{u}_t$ is subject to marginal rotational noise $\mathcal{N}(0, \sigma^2)$. Once the tumbling event is triggered, the reorientation is assumed to be instantaneous: the agent immediately samples a new heading uniformly from all possible directions, $\hat{u}_{\text{new}} \sim \text{Unif}(\mathbb{S}^1)$, without any translational displacement.

While this objective model successfully reproduces macroscopic gradient-ascent trajectories, it rests on a fundamentally unrealistic premise: it assumes the cell has a built-in global navigator to compute the exact misalignment $\theta$ and evaluate Equation 1. In reality, algae must make navigational decisions relying solely on localized, noisy photon fluxes impinging on its physical eyespot. To understand how the cell actually decides to act under severe partial observability, we must shift our perspective from objective mechanics to subjective inference.

## A subjective model for the curiosity-driven directed run-and-tumble motion.

**Partial observability: biophysical model** To model the subjective perspective of the cell, we first define its sensory limitations. Unlike the objective model which assumes access to global alignment, real algae rely on localized receptors. Although the real alga (*Chlamydomonas reinhardtii*) possesses a single eyespot and swims while spinning in 3D, we adopt a 2D projection model to capture its sensory dynamics. The incident light flux is decomposed into left and right channels relative to the current heading. It is important to note that these "two eyes" serve purely as a modeling surrogate to encode directionality and body occlusion in 2D, rather than an anatomical claim.

As illustrated in Figure 1, we model the agent as a circular body of radius $R > 0$, located at $x_t \in \mathbb{R}^2$ with a unit heading vector $v_t \in \mathbb{S}^1$. Two virtual sensors are mounted on the body rim at symmetric offset angles $\pm\gamma$ relative to the heading. Furthermore, the optical axes of these sensors are directed at angles $\pm\alpha$ relative to $v_t$, and each sensor can only perceive incoming light within a restricted half field-of-view ($FoV$).

This physical embodiment fundamentally acts as a sensory bottleneck. As we will formalize in the subsequent POMDP framework, this restricted sensor configuration induces inherent *geometric ambiguity*. Specifically, due to the

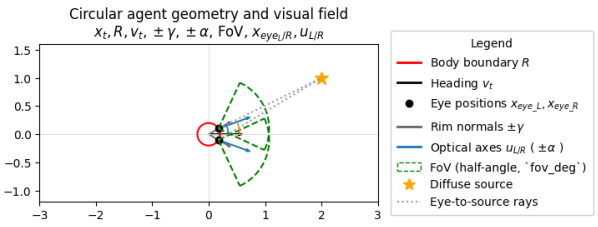



Figure 1: Illustration of the biophysical agent in 2D. The directional shielding and limited field-of-view of the sensors create blind spots and geometric ambiguity regarding the true light source position.

Field-of-View (FoV) limits of each sensor, when the light source lies directly behind the agent or directly in front, the alga receives essentially the same low sensory input on its lateral photoreceptors, resulting in minimal activation that obscures the true light direction.

**POMDP** Given the sensory bottleneck and the resulting partial observability, we cast the phototaxis navigation as a subjective inference process by formulating a Partially Observable Markov Decision Process (POMDP).

**States and actions:** The hidden state $h \in \{0, \dots, n-1\}$ (with $n = 5$ in our experiments) discretizes the relative angle between the agent's heading and the light source. Each $h$ is associated with an angle $\theta(h) = \frac{2\pi h}{n}$. The action space $\mathcal{A} = \{0, 1\}$ dictates movement: $a = 0$ (run) maintains the current heading with forward displacement, while $a = 1$ (tumble) pauses forward motion to randomly sample a new heading from a uniform distribution.

**Observations:** At each time step $t$, the agent receives a discrete observation $o_t = (o_L, o_R) \in \{0, \dots, m-1\}^2$. To explain how this observation is calculated from the spatial geometry, let $\theta$ be the relative angle to the light source. The incident light intensity $I_c$ on each sensor $c \in \{L, R\}$ is proportional to its optical axis alignment $\max(0, \cos(\theta \mp \alpha))$ and is explicitly gated by the Field-of-View (FoV). To model the non-linear photoreceptor physics, this raw signal is passed through a logistic saturation function and quantized into $m$ discrete levels (set to $m = 2$, yielding 4 distinct observation pairs):

$$o_c = \text{bin}\left(\frac{1}{1 + \exp(-\beta_s(I_c - I_{50}))}\right) \tag{2}$$

where $\beta_s$ controls the saturation steepness and $I_{50}$ is the half-response threshold.

**Transition and emission:** The transition matrix $T_a(h'|h)$ encodes the rotational dynamics. A *run* preserves the relative angle with marginal diffusion leakage $\epsilon$, yielding a diagonally dominant matrix: $T_{\text{run}}(h'|h) = \frac{\delta_{h',h} + \epsilon}{1 + \epsilon n}$. A *tumble* completely randomizes the heading, yielding a uniform transition: $T_{\text{tumble}}(h'|h) = 1/n$.

The emission kernel $Z(o|h) = P(o|h)$ mathematically bridges the abstract hidden states and the physical observations. To construct this matrix, each hidden state $h$ is assigned a prototype angle $\theta_h = \frac{2\pi h}{n}$. We pass $\theta_h$ through the biophysical observation model (Eq. 2) to determine its deterministic target observation $o^*$. We then assign a high probability spike to $o^*$ and a uniform noise floor $\epsilon_{\text{obs}}$ to all other bins, followed by column normalization. This yields a robust, physically-grounded likelihood matrix.

**Memoryless belief update:** The agent maintains a subjective belief $b_t$, a probability distribution over the hidden states $h$. We formulate a memoryless Bayesian update so that it can be implemented by executable biochemical dynamics. At each step $t$, the posterior belief $b_t^+$ is computed via a single-step likelihood reweighting over a fixed uniform prior $b_0$:

$$b_t^+(h) = \frac{Z(o_t|h)b_0(h)}{\sum_{j=0}^{n-1} Z(o_t|j)b_0(j)} \tag{3}$$

**Curiosity-driven one-step-look-ahead policy** To select action based on subjective memoryless belief, we adopt a one-step look-ahead policy, scoring each candidate action by balancing task-oriented exploitation and information-gathering exploration. After contemplating an action $a$, the anticipated next-state belief is $b'_a = T_a b_t^+$. The total value of an action is defined as:

$$V(a) = V^{(1)}(a) + \lambda V^{(2)}(a) \tag{4}$$

where $\lambda \geq 0$ is a tunable hyperparameter balancing the exploration-exploitation trade-off. The exploitation term $V^{(1)}$ represents the expected immediate task reward:

$$V^{(1)}(a) = \mathbb{E}_{h' \sim b'_a}[r] = r^\top(b'_a) \tag{5}$$

where the reward vector $r \in \mathbb{R}^n$ assigns higher values to hidden states that are closely aligned with the light source, which is why we take the global alignment into account during the next-stage learning. The exploration (curiosity) term $V^{(2)}$ evaluates the expected information gain about the hidden state $H$ from the anticipated future observation $O$. We quantify this using Mutual Information (MI):

$$V^{(2)}(a) = I(H; O \mid b'_a) = \sum_{o,h} b'_a(h) Z(o|h) \log \frac{Z(o|h)}{p_O(o)} \tag{6}$$

where $p_O(o) = \sum_h Z(o|h)b'_a(h)$. Finally, actions are selected either via a Boltzmann (softmax) policy, $\pi(a) \propto \exp(V(a)/\tau)$, where the temperature $\tau$ modulates the

stochasticity of the behavior; or deterministically via $\arg\max_a V(a)$.

By formulating this subjective policy, we hypothesize that the curiosity term $V^{(2)}$ might drive the agent to actively resolve the geometric ambiguity highlighted in the previous section. Whether this theoretically translates to functional tumbling behavior will be empirically examined in our simulation results.

## Implementation in chemical reaction networks

### Polynomial upper bound of curiosity.

Recent theoretical frameworks demonstrate that biochemical cascades can function as "probabilistic computers," where mass-action kinetics at steady state can compute any Rational Function with Non-negative Coefficients (RFNC) ( Colliaux et al. (2017)).

While the exploitation term $V^{(1)}$ is intrinsically linear and thus directly computable by Chemical Reaction Networks (CRNs), the exploration term $V^{(2)}$ requires computing logarithms. Since the logarithm is not an RFNC, it cannot be directly evaluated by mass-action polynomials.

To obtain a chemically implementable surrogate, we upper-bound the mutual-information term by applying the algebraic inequality $\log x \leq x - 1$ to the logarithmic factor and then collecting the action-independent constant terms. This yields the following RFNC-compatible polynomial form:

$$V^{(2)}(a) \leq \sum_{o,h} b'_a(h) Z(o|h)^2 + m^2 - 2 \tag{7}$$

Since $m^2 - 2$ is constant across all actions, it does not affect the action-selection policy. Thus, we substitute $V^{(2)}(a)$ with a fully chemically computable polynomial curiosity score:

$$\tilde{V}^{(2)}(a) = \sum_h b'_a(h) \sum_o Z(o|h)^2 \tag{8}$$

This pure RFNC formulation allows us to seamlessly construct the corresponding CRN-ODE modules.

### CRN implementation.

Systematic CRN compilers have been developed to program diverse input functions. Classic strands include deterministic computation of semilinear functions (Chen et al. (2014)) and the synthesis of CRNs for finite-support discrete probability distributions (Cardelli et al. (2018)). More recently, Wiuf et al. (2023) realized dynamic inference by translating Hidden Markov Models (HMMs) into CRNs that simulate the Baum-Welch algorithm via mass-action kinetics.

Finding a middle ground between static CRN compilers and complex dynamic HMM networks, our CRN architecture integrates the monotone-accumulation principles of static computation with a forward-style inference scaffold. We construct parallel sub-networks for each action $a \in \{0(\text{run}), 1(\text{tumble})\}$, which share the same structural motifs but differ in reaction rates parameterized by $T_a$. The network operates through mass-action Ordinary Differential Equations (ODEs) across three modular blocks:

**1. Memoryless one-step belief update:** Let $A_h^{(1)}$ denote the species storing the prior belief $b_0(h)$, and $E_v$ be a boolean catalyst activated by receiving observation index $v$. The unnormalized likelihood $U_h$ is accumulated catalytically:

$$A_h^{(1)} + E_v \xrightarrow{k_{v,h}} A_h^{(1)} + U_h + E_v \quad (k_{v,h} = Z[v,h]) \tag{9}$$

To execute Bayesian normalization without complex division circuits, we introduce a shared resource pool $M$ (initialized to 1.0). The posterior belief species $A_h^{(2)}$ is produced by consuming $M$ proportionally to $U_h$:

$$M + U_h \xrightarrow{1} A_h^{(2)} + U_h. \tag{10}$$

This reaction must be considered jointly with the depletion of the shared pool,

$$\frac{d[M]}{dt} = -k[M] \sum_j [U_j], \qquad \frac{d[A_h^{(2)}]}{dt} = k[M][U_h]. \tag{11}$$

Since the $U_h$ are fixed during one observation step, all species $A_h^{(2)}$ compete for the same normalization resource $M$. Hence at steady state as $t \to \infty$ and $M \to 0$, the concentration $[A_h^{(2)}]$ cleanly converges to the normalized posterior $b_t^+(h)$.

$$[A_h^{(2)}]_\infty = \frac{[U_h]}{\sum_j [U_j]} = \frac{Z[o_t, h]\, b_0(h)}{\sum_j Z[o_t, j]\, b_0(j)} = b_t^+(h). \tag{12}$$

For a POMDP with $n$ hidden states and $m^2$ possible observation pairs, this modular design scales efficiently, requiring only $\mathcal{O}(n + m^2)$ chemical species and $\mathcal{O}(n \cdot m^2)$ reactions.

**2. Exploitation (immediate expected reward):** To compute $V^{(1)}(a) = r^\top T_a b_t^+$, we employ a "leaky reservoir" motif. A reward species $R$ is catalytically produced by the posterior $A_h^{(2)}$ and simultaneously degrades:

$$A_h^{(2)} \xrightarrow{w_a(h)} A_h^{(2)} + R, \quad R \xrightarrow{1} \emptyset \tag{13}$$

where the production rate is $w_a(h) = (r^\top T_a)_h$. By solving the mass-action ODE $\dot{R} = 0$, the steady-state concentration $[R]_\infty$ exactly yields the exploitation score $V^{(1)}(a)$.

**3. Exploration (curiosity polynomial upper bound):** Similarly, to compute the polynomial bound $\tilde{V}^{(2)}(a)$ (Eq. 8), we introduce a curiosity species $C_h$ for each state, using the transitioned likelihood squared as the production rate:

$$A_h^{(2)} \xrightarrow{c_h} A_h^{(2)} + C_h, \quad C_h \xrightarrow{1} \emptyset \tag{14}$$

where $c_h = (T_a)_{h,:} \sum_v Z[v,\cdot]^2$.

Ultimately, the total CRN-ODE action value is read out at steady state as $V_{\text{CRN}}(a) = [R]_\infty + \lambda \sum_h [C_h]_\infty$. By selecting the action $a^*$ that maximizes this steady-state concentration, the biochemical network physically executes the curiosity-driven POMDP policy.

## Simulations & comparison to data

To rigorously evaluate our framework, we divide our experiments into two stages. First, we provide simulations for phototactic behavior under a single light source to verify the mathematical and biochemical equivalence of our CRN design. Second, we transition to data-driven learning, extracting latent policies from real biological trajectories and benchmarking them against objective mechanistic models.

### Numerical & CRN POMDP simulations.

We first validate the computational integrity of our subjective POMDP and its CRN-ODE implementation in an idealized single-source environment. The agent is placed in a simulated 2D field with a distant light source and operates under a fixed, hand-specified reward vector designed to favor light alignment.

Our goal here is to verify that the mass-action CRN-ODE faithfully reproduces the numerical policy. We simulated trajectories using both exact numerical computations and numerical integration of the CRN-ODEs (running until steady-state convergence at each decision step).

| Simulation | Curiosity Weight ($\lambda$) | Tumble Ratio |
|---|---|---|
| numerical, exact MI | 0.3 | $9/4000 \approx 2.3 \times 10^{-3}$ |
| numerical, exact MI | 6.83 (coarsely-tuned) | $29/4000 \approx 7.3 \times 10^{-3}$ |
| numerical, MI upper bound | 6.83 | $7/2000 = 3.5 \times 10^{-3}$ |
| CRN–ODE | 6.83 | $4/2000 = 2.0 \times 10^{-3}$ |
| numerical, MI upper bound | 2(grid searched) | $2/2000 = 1.0 \times 10^{-3}$ |
| numerical, soft policy | 0 | $7/2000 = 3.5 \times 10^{-3}$ |
| CRN–ODE, soft policy | 0 | $10/2000 = 5.0 \times 10^{-3}$ |

Table 1: Tumble ratio for different policy simulations and curiosity weights with single-light

As summarized in Table 1, the tumble ratios (calculated as the fraction of tumble steps over the entire trajectory) are consistently maintained at a low magnitude ($\sim 10^{-3}$) across all models. This indicates that all configurations successfully guide the agent toward the light source with high locomotive efficiency. Crucially, replacing the exact MI with the polynomial upper bound only marginally alters the tumble ratio.

Furthermore, we explicitly distinguish between the discrete *numerical* POMDP (which computes exact matrix multiplications instantaneously) and the *CRN-ODE* implementation (which relies on continuous-time integration of mass-action kinetics until steady-state). The CRN-ODE yields tumbling statistics ($2.0 \times 10^{-3}$) that closely track its numerical counterpart ($3.5 \times 10^{-3}$). The slight numerical divergence arises naturally from the finite integration time in ODE solvers compared to instantaneous discrete math. Nonetheless, this confirms that the complex Bayesian belief updating and curiosity evaluation can be robustly executed by physical biochemical reaction networks.

**Key insight: tumbling as active exploration under geometric ambiguity.** Recall the *geometric ambiguity* introduced by the sensory bottleneck: The eyespot is positioned laterally and slightly behind the cell, light sources located at specific off-axis angles in front of or behind the cell can produce indistinguishable bilateral readings. In these ambiguous regimes, the posterior belief over the hidden angle flattens, causing the agent's spatial uncertainty to spike. When this uncertainty is high, the curiosity term assigns disproportionately greater value to actions that promise information gain. Consequently, *tumbling* naturally emerges as an active exploration step: by randomizing its heading, the cell drastically alters its illumination perspective, thereby gathering the critical sensory evidence needed to resolve the front–back ambiguity. In this light, run–tumble alternation is justified not merely as motor noise, but as a purposeful, active sensing strategy to conquer partial observability.

### Data-driven policy learning.

The above simulations establish the computational feasibility of our CRN-POMDP framework for phototaxis scenarios. However, establishing true biological validity requires learning the policy from real organism trajectories. Our goal in this section is to train the policy on experimental data and combine it with Inverse Reinforcement Learning (IRL) to infer the true exploitation rewards, ultimately obtaining an optimal policy that matches empirical real-data statistics.

**Data preprocessing** We utilize 30 real-world 2D trajectories of *Chlamydomonas* swimming under optical fiber illumination. To translate this raw tracking data into a discrete-time POMDP format, we first reconstructed the absolute time axis from camera timestamps to compute instantaneous translational speeds $|v|$ and wrapped angular velocities $\omega$. We then applied a labeling rule: intervals were classified as a *tumble* if their speed fell below a robust low-speed threshold (the 0.05 $\times$ global speeds) and their angular velocity exceeded a high-turn threshold (derived via K-Means clustering on $|\omega|$). All other intervals were labeled as *runs*.

As the original camera frame intervals are irregular, we projected the trajectories onto a uniform time grid ($\Delta t$=0.125,s) via linear interpolation. The discrete action labels were mapped to this uniform grid using a majority vote based on temporal overlap. This pipeline effectively mitigates interpolation-induced drift and yields a standardized sequence of state-action-observation tuples for downstream inference.

To further evaluate the generalization of our learned models, we performed a strict train-test split on the dataset. Out of the 30 processed trajectories, we randomly reserved 4 trajectories as a held-out test set strictly for downstream benchmarking. The remaining 26 trajectories were exclusively utilized for training the POMDP policy and performing IRL.

**Policy learning & benchmarking** To recover the organism's navigational strategy, we decompose the learning process into 2 phases: (1) joint learning of the observation policy, kinematics, and POMDP emission kernels, and (2) IRL to extract the underlying exploitation reward and benchmark against objective models.

**Phase 1: joint policy learning** In our previous idealized simulations, we assumed tumbling agents reorient instantaneously in place ($c_{\text{tum}} = 0$). However, real cells exhibit some continuous translational displacement even while reorienting. To rigorously reflect this, we generalize our kinematics for the data-driven pipeline by introducing learnable speed scales for both actions: $c_{\text{run}}$ and $c_{\text{tum}}$ (where typically $0 < c_{\text{tum}} \leq c_{\text{run}}$).

We train a 2-layer MLP policy $\pi_\theta(a|o)$ mapping one-hot observations to action probabilities, jointly with the kinematic scales and the POMDP emission kernel $Z(o|h)$. The training minimizes a combined objective:

$$\mathcal{L}_{\text{Phase1}} = L_{\text{kin}} + \beta_{\text{obs}} L_{\text{obs}} + \alpha_Z \text{KL}(Z \| Z_0) \tag{15}$$

where $L_{\text{obs}}$ maximizes the transition likelihood, $\text{KL}(Z\|Z_0)$ regularizes the emission kernel against the biophysical prior $Z_0$, and $L_{\text{kin}}$ is the kinematic reconstruction loss:

$$L_{\text{kin}} = \mathbb{E}_t \left[ \Delta t \left\{ (1 - p_{\text{tum}}) \ell_{\text{run}} + p_{\text{tum}} \ell_{\text{tum}} \right\} \right] \tag{16}$$

Here, $\ell_{\text{run/tum}}$ represents the mean squared error between the predicted step displacement under action and the actual trajectory step. After training, the model successfully extracted biological speeds ($c_{\text{run}} \approx 18.8, px/s \geq c_{\text{tum}} \approx 0.03, px/s > 0$). Since this supervised policy primarily fits physical displacements rather than exploratory motives (producing almost zero tumbles when embedded in the POMDP), we freeze these parameters and proceed to Phase 2.

**Phase 2: IRL & baseline benchmarking** However, the purely supervised policy in phase 1 failed to capture the exploratory run-and-tumble alternation, producing almost zero tumbles. To uncover the latent functional objective driving this behavior, we apply Inverse Reinforcement Learning (IRL) that recovers the reward vector $r$ to reproduce real-data behavioral patterns with fixed learned parameters ( Choi and Kim (2011); Djeumou et al. (2022)).

Crucially, to avoid the identifiability problem in reward learning —where intrinsic exploration bonuses and extrinsic task rewards can confound one another —we intentionally set the curiosity weight to zero ($\lambda = 0$) during the IRL phase. This isolates the extrinsic motivation, ensuring that the recovered reward vector $r \in \mathbb{R}^n$ purely reflects the organism's physical preference for spatial alignment. We freeze the learned kernels and isolate this exploitation value function:

$$Q_r(o, a) = r^\top (T_a b_t^+) \tag{17}$$

With the extrinsic goal $r$ strictly anchored, the organism's inherent exploratory drive is captured macroscopically via a soft Boltzmann policy, governed by a temperature hyperparameter $\tau$:

$$\pi_r(a \mid o) = \frac{\exp(Q_r(o, a)/\tau)}{\sum_{a'} \exp(Q_r(o, a')/\tau)} \tag{18}$$

By computing an empirical expert distribution $P_{\text{exp}}(a \mid o)$ from the real data via Laplacian smoothing, we recover the optimal reward $r^*$ by minimizing the cross-entropy loss with $L_2$ regularization:

$$\min_r \mathbb{E}_o \left[ -\sum_a P_{\text{exp}}(a \mid o) \log \pi_r(a \mid o) \right] + \lambda_r \|r\|_2^2 \tag{19}$$

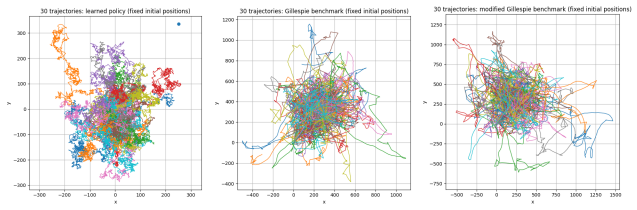
The recovered reward $r^*$ successfully assigns higher values to hidden states directly facing the light, confirming our theoretical exploitation design ( Ziebart et al. (2008)).

**Benchmarking: objective vs. subjective models** To fairly evaluate our data-driven subjective POMDP, we benchmark it against the objective models. The *Standard SSA* treats tumbling as an instantaneous event without displacement ($c_{\text{tum}} = 0$), serving as our classical theoretical baseline. However, since our Phase 1 learning revealed a non-zero tumbling speed ($c_{\text{tum}} > 0$), we explicitly introduce a *Modified CTMC* objective baseline alongside the *Standard* objective model. In the Modified CTMC version, the agent does not merely spin in place; instead, during a tumble, it experiences continuous translational displacement: $\Delta x = \Delta t \cdot c_{\text{tum}} \hat{u}_{\text{tum}}$, where $\hat{u}_{\text{tum}} \sim \text{Unif}(\mathbb{S}^1)$.

To quantitatively compare performance, we evaluate the alignment cosine between the agent's heading $v_t$ and the true light source $\mu$:

$$\cos \theta_{\text{align}} = \frac{v_t \cdot (\mu - x_t)}{\|v_t\| \|\mu - x_t\|} \tag{20}$$

Figure 2 provides a qualitative comparison of 30 simulated trajectories under the learned subjective POMDP, the standard objective model, and the modified CTMC objective model. Visual inspection suggests that the learned POMDP produces a broader spatial dispersion than the two objective baselines; we therefore quantify model behavior below

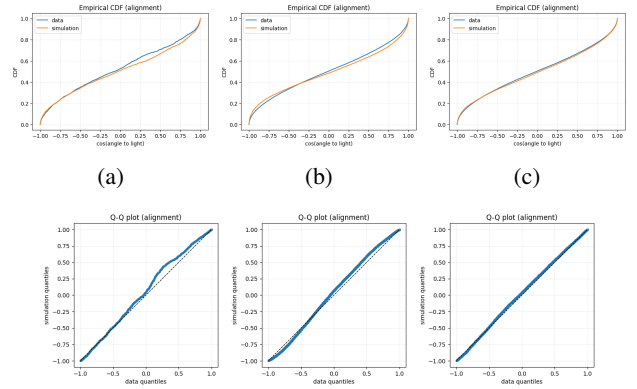



(a) (b) (c)

Figure 2: 30 trajectories simulated from: a) learned subjective POMDP model; b) standard objective model; c) modified CTMC objective model. All with the same fixed initial position and light source center (blue spot in the figure (a) ).

using alignment-distribution metrics rather than relying on trajectory plots alone.

We continue to analyze the alignment distribution, with the Empirical Cumulative Distribution Functions (CDFs) of $\cos\theta_{\text{align}}$ and quantile-quantile (Q-Q) plots, first comparing the learned POMDP policy with modified SSA and then with standard SSA in the following Figure 3:




(a) (b) (c)

(d) (e) (f)

Figure 3: **Alignment distribution analysis: Subjective POMDP vs. Real Data and Objective Baselines.** The learned subjective POMDP policy is evaluated against three targets across the columns: the real biological test data **(a, d)**, the idealized *Standard SSA* objective model ($c_{\text{tum}} = 0$) **(b, e)**, and the *Modified CTMC* objective model ($c_{\text{tum}} > 0$) **(c, f)**. **(Top row)** Empirical Cumulative Distribution Functions (CDFs) of the alignment cosine ($\cos\theta_{\text{align}}$). **(Bottom row)** Quantile-Quantile (Q-Q) plots evaluating the structural congruence of the distributions.

The CDFs (Fig. 3a-c) demonstrate that the subjective POMDP policy tightly tracks the cumulative alignment error of the real biological data as well as both omniscient baselines. The Q-Q plots (Fig. 3d-f) further validate this structural congruence: the quantiles of the learned subjective policy form a near-perfect diagonal with all three targets. The alignment against the Modified CTMC (Fig. 3c, f) further confirm this structural congruence: the quantiles of the learned subjective policy form a near-perfect diagonal with the objective baselines. These results indicate that the learned subjective policy achieves alignment distributions close to those of the objective baselines, with the strongest agreement observed for the modified CTMC baseline.

To verify this congruence, we computed the Wasserstein-1 distance ($W_1$) and the Kolmogorov-Smirnov statistic ($D$), summarized in Table 2.

| **Comparison Pair** | **Wasserstein-1** ($W_1$) | **KS Statistic** ($D$) |
|---|---|---|
| Policy vs. Modified CTMC | **0.0187** | **0.0156** |
| Policy vs. Standard SSA | 0.0441 | 0.0387 |
| Policy vs. Real Data | 0.0458 | 0.0536 |

Table 2: **Statistical distances of alignment distributions.** The metrics evaluate the congruence between the empirical real data (held-out test set), the objective baselines, and the learned subjective POMDP policy. Lower Wasserstein-1 ($W_1$) and Kolmogorov-Smirnov ($D$) values indicate higher distributional similarity.

As detailed in Table 2, the metrics reveal a critical insight: the learned subjective policy structurally aligns best with the *Modified CTMC* ($W_1 \approx 0.0187, D \approx 0.0156$). It proves that our memoryless POMDP successfully mastered the gradient-ascent strategy constrained by realistic fluid kinematics ($c_{\text{tum}} > 0$). It also maintains high fidelity against the held-out Real Data ($W_1 \approx 0.0458, D \approx 0.0536$).

This quantatively proves that a memoryless agent driven by compressed local observations can recover much of the macro-level alignment behavior captured by the objective models. This supports, rather than proves, the hypothesis that the IRL-recovered reward captures biologically relevant alignment preferences. This suggests that the latent reward structure recovered via IRL captures subtle biological preferences that strict mathematical misalignment penalties may overlook. However, despite this near-perfect spatial orientation, a dramatic divergence exists in their action frequencies, which sets the stage for analyzing the biological exploration-exploitation trade-off.

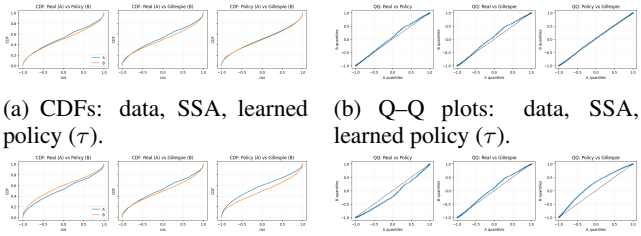
**Exploration-exploitation trade-off** Despite the excellent alignment distribution, a crucial discrepancy emerged in action frequencies. As shown in Table 3, integrating realistic kinematics ($c_{\text{tum}} > 0$) into the Modified CTMC inherently raises the tumble ratio to $0.130$ compared to the idealized Standard SSA ($0.038$). Because reorienting while moving dilutes turning efficiency, the agent must tumble more frequently to achieve the same alignment.

Our data-driven POMDP captures this kinematic reality, which explains why the initial IRL soft policy ($\tau = 1.0$) overestimates the tumbling frequency ($0.503$). To bridge

this gap toward the real experimental data (0.027), we calibrated the temperature hyperparameter to $\tau^* \approx 0.007$. This sharpens the policy, successfully reducing the simulated tumble ratio to 0.0121, bringing it closer to the empirical value while still leaving a measurable gap.

| Model / Data | Setting | Tumble ratio |
|---|---|---|
| Real data | — | 0.027 |
| objective(standard) | instantaneous reorient. | 0.038 |
| objective(modified) | run-while-tumble CTMC | 0.13 |
| Learned subjective (before IRL) | - | 0 |
| Learned subjective (after IRL) | $\tau = 1.0$ | 0.503 |
| Learned subjective (calibrated) | $\tau^\star = 0.007$ | 0.0121 |

Table 3: Tumble ratios for real data, objective baselines and learned subjective models.



(a) CDFs: data, SSA, learned policy ($\tau$).

(b) Q–Q plots: data, SSA, learned policy ($\tau$).

(c) CDFs: data, SSA, learned policy ($\tau^\star$).

(d) Q–Q plots: data, SSA, learned policy ($\tau^\star$).

Figure 4: Empirical CDFs and Q–Q plots of the alignment cosine $\cos\theta_{\text{align}}$ (heading vs. light direction) for real data, SSA baselines and learned POMDP policies, with the original temperature $\tau$ (top) and the calibrated temperature $\tau^\star$ (bottom).

As visualized in Fig. 4, calibrating the temperature hyperparameter $\tau$ suppresses excess tumbling but slightly broadens the alignment distribution (Fig. 4c, d). However, suppressing the tumble frequency naturally comes at a cost: without frequent reorientations to correct drifting errors, the alignment distribution broadens. Quantitatively, the alignment error against the Real Data measurably increased (with $W_1$ rising from 0.0458 to $\mathbf{0.1006}$, and the KS statistic $D$ increasing from 0.0536 to $\mathbf{0.0797}$).

This statistical shift uncovers a fundamental, physically meaningful biological trade-off. In a viscous environment with residual translational momentum ($c_{\text{tum}} > 0$), reorientation is mechanically inefficient. The temperature $\tau$ therefore modulates not merely mathematical stochasticity, but the organism's vital balance between *locomotive efficiency* (fewer energy-consuming tumbles) and *sensory acquisition* (more frequent exploratory tumbles to correct heading errors and resolve geometric ambiguity). The biological agent naturally navigates this tension, demonstrating that a purely subjective, chemically computable belief loop is sufficient to generate sophisticated, adaptive phototaxis.

## Discussion

By bridging abstract algorithmic reinforcement learning with physically realizable mass-action kinetics, our data-driven POMDP framework demonstrates that macroscopic phototactic behavior can be accurately replicated by subjective, memoryless inference coupled to mass-action-compatible dynamics. Our results suggest a reinterpretation of run-and-tumble behavior as an information-seeking strategy that helps resolve perceptual ambiguity during navigation. In this view, tumbling is not merely stochastic noise but a functional, curiosity-driven mechanism for actively sampling the environment when directional cues are uncertain. By linking this behavioral dynamics to biochemical reaction networks, we show how such exploratory processes can be physically realized through intracellular molecular interactions. This provides a bridge between abstract descriptions of adaptive behavior and the underlying biochemical mechanisms that can implement them.

This work situates itself among recent efforts exploring microbial navigation through advanced computational paradigms. Related work has explored run-and-tumble navigation in terms of reinforcement learning ( Pramanik et al. (2025)) and active inference frameworks emphasizing curiosity-driven exploration Tschantz et al. (2020), particularly in the context of chemotaxis. Other studies have shown that the structure of biochemical pathways involved in chemotaxis or phototaxis may be inferred from the tumbling dynamics Lei et al. (2025). Our approach complements these perspectives by explicitly demonstrating how information-seeking dynamics can be implemented through modular biochemical reaction networks compatible with mass-action kinetics.

From an artificial life perspective, these results highlight how adaptive exploratory behavior can emerge from relatively simple biochemical circuits interacting with uncertain environments. Such mechanisms illustrate how minimal biological systems may realize sophisticated information-processing strategies without requiring complex neural architectures. Behavioral variability, historically dismissed as noise, may therefore play a constructive role in shaping effective exploration.

Looking forward, this framework provides a foundation for several possible avenues. Future work should extend this framework to richer environments, such as multiple or dynamically varying light sources, where more complex collective exploratory strategies may emerge. Further investigation of the biochemical modules may also support the design of synthetic control mechanisms for phototactic microorganisms. Finally, these principles could inspire embodied implementations in minimal robotic systems, where simple sensorimotor loops reproduce biologically inspired exploration and active sensing strategies in the physical world.

## Acknowledgements

We thank Pierre Bessière and Jacques Droulez for advice on the observation design and evaluation protocol, and for discussions that improved our evaluation setup and presentation. We also thank Sorbonne University and Sony CSL for supporting this research.

This internship was funded by AMIES – PEPS project "*Generative models of chemotaxis and phototaxis*".

## References

Auconi, A., Novak, M., and Friedrich, B. M. (2022). Gradient sensing in bayesian chemotaxis. *Europhysics letters*, 138(1):12001.

Cardelli, L., Kwiatkowska, M., and Laurenti, L. (2018). Programming discrete distributions with chemical reaction networks. *Natural computing*, 17(1):131–145.

Chen, H.-L., Doty, D., and Soloveichik, D. (2014). Deterministic function computation with chemical reaction networks. *Natural computing*, 13(4):517–534.

Choi, J. and Kim, K.-E. (2011). Inverse reinforcement learning in partially observable environments. *J. Mach. Learn. Res.*, 12(null):691–730.

Colliaux, D., Bessière, P., and Droulez, J. (2017). Cell signaling as a probabilistic computer. *International Journal of Approximate Reasoning*, 83:385–399.

Djeumou, F., Cubuktepe, M., Lennon, C., and Topcu, U. (2022). Task-guided inverse reinforcement learning under partial information. In *Proceedings of the international conference on automated planning and scheduling*, volume 32, pages 53–61.

Erban, R. and Othmer, H. G. (2004). From individual to collective behavior in bacterial chemotaxis. *SIAM Journal on Applied Mathematics*, 65(2):361–391.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gottlieb, J. and Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature reviews. Neuroscience*, 19(12):758–770.

Heinonen, R. A., Biferale, L., Celani, A., and Vergassola, M. (2025). Optimal trajectories for bayesian olfactory search in turbulent flows: The low information limit and beyond. *Physical review fluids*, 10(4):044601.

Lei, S., Li, Y., Ma, Z., Zhang, H., and Tang, M. (2025). Identification of the governing equation of stimulus-response data for run-and-tumble dynamics. *PLOS Computational Biology*, 21(8):e1013287.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction.

Polin, M., Tuval, I., Drescher, K., Gollub, J. P., and Goldstein, R. E. (2009). Chlamydomonas swims with two "gears" in a eukaryotic version of run-and-tumble locomotion. *Science*, 325(5939):487–490.

Pramanik, R., Mishra, S., and Chatterjee, S. (2025). Run-and-tumble chemotaxis using reinforcement learning. *Physical Review E*, 111(1):014106.

Tindall, M., Porter, S., Maini, P., Gaglia, G., and Armitage, J. (2008). Overview of mathematical approaches used to model bacterial chemotaxis i: The single cell. *Bulletin of mathematical biology*, 70:1525–69.

Tschantz, A., Seth, A. K., and Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS computational biology*, 16(4):e1007805.

Wiuf, C., Behera, A., Singh, A., and Gopalkrishnan, M. (2023). A reaction network scheme for hidden markov model parameter learning. *Journal of The Royal Society Interface*, 20(203):20220877.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438.