CONNECTOME-BASED REINFORCEMENT LEARNING FOR EMBODIED NEURAL CIRCUIT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning policies are typically represented by generic neural networks, whose architectures are largely independent of biological neural organization. In contrast, animal behavior emerges from efficient structured neural circuits through continuous interaction with the body and environment. Whether an experimentally grounded connectome can serve directly as a reinforcement learning policy remains largely unexplored. To address this question, we propose *Connectome-as-Policy*, which formulates a connectome-derived *Caenorhabditis elegans* salt chemotaxis subcircuit as the actor in a closed-loop embodied learning system. The framework combines a fixed circuit topology, continuous neural dynamics, a simulated worm body, and a two-dimensional salt field. Reinforcement learning optimizes synaptic strengths and neuronal parameters while preserving the predefined connectivity. Across six reinforcement learning algorithms, the connectome policy learns stable chemotactic navigation and concentration regulation. Compared with conventional neural backbones, it achieves competitive performance with substantially fewer trainable parameters while enabling circuit-level interpretability. Robustness analyses reveal smooth behavioral transitions and stable performance under varying concentrations and sensory noise. Additional reinforcement learning benchmarks demonstrate comparable performance across standard control tasks. These results suggest that biological circuit priors offer a compact and interpretable foundation for embodied reinforcement learning.

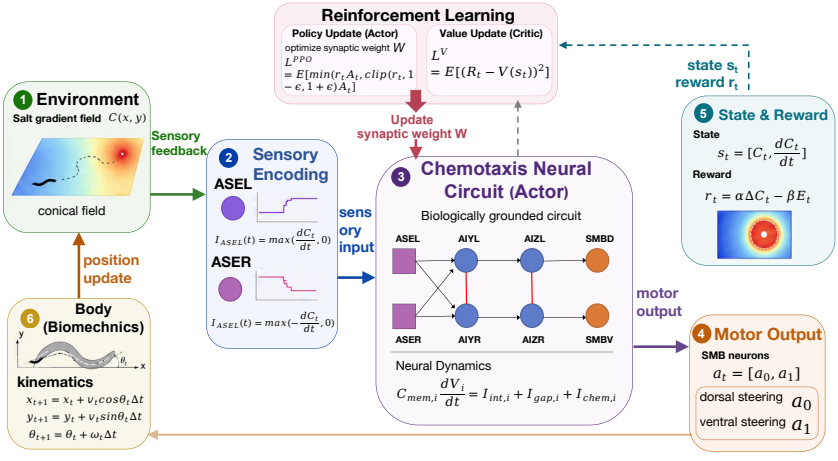


Figure 1: **Closed-loop Connectome-as-Policy framework.** The worm agent interacts with a continuous salt concentration field through a biologically grounded sensorimotor circuit. Sensory neurons encode local salt concentration and temporal changes, the chemotaxis circuit generates continuous motor commands, and body motion updates subsequent observations. The reinforcement learning actor optimizes neuronal and synaptic parameters, while the critic estimates state values to stabilize learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Deep reinforcement learning has achieved remarkable success in sequential decision-making by representing policies with generic neural architectures such as multilayer perceptrons, convolutional networks, and recurrent neural networks Pineau et al. (2009); Mitchell & Petzold (2018); Shlizerman et al. (2014). Despite their strong function approximation capability, these architectures are largely designed as unconstrained universal approximators and rely almost entirely on data to discover effective control strategies. Consequently, they often require extensive training samples, contain a large number of parameters, and provide limited interpretability regarding the internal mechanisms underlying learned behaviors White et al. (1986); Franconville et al. (2018); Kim et al. (2019). In contrast, biological nervous systems are not randomly organized. Through long-term evolution, animals have developed highly structured connectomes with sparse connectivity, recurrent motifs, and specialized functional organization, providing strong inductive priors for efficient sensorimotor control Zuo et al. (2024); Mnih et al. (2015); Itskovits et al. (2018). This naturally raises an important question: *can an experimentally grounded biological connectome serve directly as a reinforcement learning policy?*

Among biological model organisms, *Caenorhabditis elegans* offers a unique opportunity to investigate this question Beverly et al. (2011); Gabel et al. (2007); Chen et al. (2006). As the only animal with a complete connectome and extensively characterized sensorimotor circuits, it provides a biologically grounded policy architecture rather than an artificially designed neural network Churchland & Garcia-Ojalvo (2026); Lechner et al. (2017); Bhattasali et al. (2022); Kim & Shlizerman (2020). In particular, the minimal salt chemotaxis circuit has been experimentally identified from sensory neurons (ASEL/ASER) through interneurons (AIY, AIZ) to downstream motor neurons, forming a compact perception–decision–action pathway for continuous navigation Pierce-Shimomura et al. (1999); Queirós et al. (2021); Luo et al. (2014). While the circuit topology has been largely established, the neuronal and synaptic parameters responsible for adaptive chemotactic behavior remain largely unknown Salimans et al. (2017); Bargmann (2006); Bing et al. (2018). Biological learning is widely believed to arise from activity-dependent synaptic plasticity within an existing circuit instead of constructing entirely new neural architectures Demin & Vityaev (2014); Frankle & Carbin (2018); Hasani et al. (2020); Boyle et al. (2012). Therefore, the fundamental problem is no longer discovering the network topology, but learning biologically meaningful neuronal and synaptic parameters under a fixed connectome Heess et al. (2016); Iscen et al. (2018); Izquierdo & Beer (2015).

Reinforcement learning provides a natural computational framework for addressing this problem Lechner et al. (2019); Lillicrap et al. (2015); Luo (2021). Rather than viewing reinforcement learning merely as a control optimization algorithm, we interpret it as a computational model of experience-dependent synaptic plasticity. Under this perspective, evolution determines the connectome topology, whereas reinforcement learning models functional adaptation by optimizing neuronal and synaptic parameters through continuous interaction with the environment Merel et al. (2019; 2018); Keller et al. (2026); Mori et al. (2022). However, existing biologically inspired reinforcement learning methods typically redesign sparse neural architectures or simplify biological circuits for generic control tasks, while most computational studies of *C. elegans* focus on reproducing neural dynamics rather than learning adaptive behavior. Whether an experimentally identified connectome can function directly as a trainable reinforcement learning policy therefore remains largely unexplored Izquierdo & Beer (2013); Yamaguchi et al. (2018); Kim & Shlizerman (2020); Hasani et al. (2020).

To address this challenge, we propose **Connectome-as-Policy**, a biologically grounded reinforcement learning framework that directly formulates the experimentally grounded minimal *C. elegans* klinotaxis subcircuit as the policy network. The biological circuit topology is preserved throughout learning, while reinforcement learning optimizes only neuronal and synaptic parameters within a closed-loop embodied system integrating the neural circuit, worm body, and salt environment (as Figure 1 illustrates). Consequently, adaptive chemotactic behavior emerges through continuous brain–body–environment interaction instead of architectural modification Chen et al. (2023); Szigeti et al. (2014); Gleeson et al. (2018); Zhao et al. (2024); Wang et al. (2025). Furthermore, we demonstrate that the learned connectome provides an interpretable policy representation that can be transferred beyond its native chemotaxis task to standard continuous-control reinforcement learning benchmarks.

The main contributions of this work are summarized as follows:

- We propose **Connectome-as-Policy**, a biologically grounded policy representation that optimizes synaptic and neuronal parameters within a fixed *C. elegans* chemotaxis circuit to generate adaptive behavior through closed-loop interaction.
- The connectome-based policy achieves competitive control performance with substantially fewer trainable parameters than conventional neural backbones, while enabling circuit-level analysis of sensory, interneuron, and motor dynamics.
- Evaluations on standard reinforcement learning benchmarks show that the proposed representation achieves performance comparable to conventional and biologically inspired policy backbones, supporting its cross-task applicability beyond biological behavioral tasks.

2 Related Work

***C. elegans* Chemotaxis Models** Salt chemotaxis in *C. elegans* has become a canonical model for studying sensorimotor transformation. Experimental studies showed that ASEL and ASER exhibit asymmetric responses to increases and decreases in NaCl concentration, while experience-dependent plasticity shapes salt preference through downstream interneurons Pierce-Shimomura et al. (1999); Kunitomo et al. (2013). Computational models range from simplified gradient-sensing networks to connectome-constrained neuromechanical simulations Bargmann & Horvitz (1991); Gordus et al. (2015). Evolutionary optimization was first applied to minimal chemotaxis circuits, and later studies identified the ASE–AIY–AIZ–SMB pathway capable of reproducing continuous steering behaviors Izquierdo & Lockery (2010); Izquierdo & Beer (2013); Soh et al. (2018). These models primarily explain biological mechanisms rather than learning task-oriented control policies.

Biologically Grounded Priors for Reinforcement Learning Prior knowledge can be incorporated into reinforcement learning through trajectory, imitation, and architectural priors, with the latter introducing structural inductive biases directly into the policy network Lechner et al. (2017); Hasani et al. (2020); Lillicrap et al. (2015); Schulman et al. (2017); Kober et al. (2013); Levine et al. (2016); Haarnoja et al. (2018). Biologically inspired controllers provide a particularly structured form of architectural prior Lechner et al. (2020); Hasani et al. (2021); Bhattasali et al. (2022); Dan et al. (2024); Ramdya & Ijspeert (2023). Earlier studies optimized modeled *C. elegans* circuits for robotic control, while Neural Circuit Policies, Liquid Neural Networks, and Neural Circuit Architectural Priors translated sparse neural dynamics or worm locomotor organization into interpretable policy architectures Pineau et al. (2009); Schulman et al. (2017); Lillicrap et al. (2020); Haarnoja et al. (2018); Fujimoto et al. (2018); Lowe et al. (2017); Schulman et al. (2015). More recently, ENOMAD optimized a recurrent whole-connectome model using biologically informed initialization Churchland & Garcia-Ojalvo (2026); Bhattasali et al. (2022); Kim & Shlizerman (2020).

3 Methodology

3.1 Connectome-as-Policy Framework

Figure 1 illustrates the proposed Connectome-as-Policy framework, which formulates a biologically grounded *C. elegans* chemotaxis subcircuit as the actor policy. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the predefined neural graph. The policy is written as

$$\pi_\theta^{\mathcal{G}} : \mathcal{S} \to \mathcal{A}, \qquad \theta = \{g, C, V\}. \tag{1}$$

where g, C, and V denote synaptic weights, membrane capacitance, and membrane potentials, respectively. The graph topology $\mathcal{G}$ remains fixed throughout training, while reinforcement learning optimizes only θ. Environmental observations are encoded by sensory neurons and propagated through the circuit to generate motor commands, forming a closed-loop interaction.

3.2 Biological Neural Dynamics

Figure 1 depict the actor consists of sensory neurons (ASEL/ASER), interneurons (AIYL/AIYR and AIZL/AIZR), and motor neurons (SMBV/SMBD), forming a four-layer sensorimotor circuit. Each neuron is modeled as a continuous-time leaky integrator:

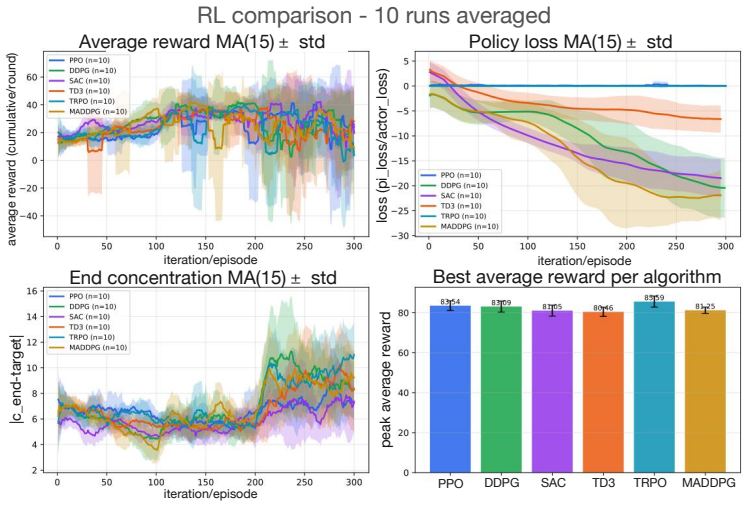


Figure 2: **Comparison of reinforcement learning algorithms for chemotaxis policy optimization.** Curves show mean performance over ten independent runs, with shaded regions denoting one standard deviation. Panels report cumulative reward, actor loss, terminal concentration error, and peak reward. PPO exhibits comparatively stable optimization and competitive navigation performance.

$$C_{\mathrm{mem},j}\frac{dV_j}{dt} = I_{\mathrm{int},j} + \sum_{k\in\mathcal{C}_j} g_{kj}^{\mathrm{chem}} s_k (E_{kj}^{\mathrm{rev}} - V_j) + \sum_{k\in\mathcal{G}_j} g_{kj}^{\mathrm{gap}}(V_k - V_j), \tag{2}$$

$$I_{\mathrm{int},j} = G_{\mathrm{mem}}(V_{\mathrm{rest}} - V_j) + \frac{G_{\mathrm{act}}}{1+\exp[-k_{\mathrm{act}}(V_j - V_{\mathrm{act}})]} + I_{\mathrm{SR}}, \tag{3}$$

where $\mathcal{C}_j$ and $\mathcal{G}_j$ denote the sets of chemical synapses and electrical synapses connected to neuron j, respectively. The parameter g_{kj}^{chem} represents the maximal conductance of the chemical synapse from neuron k to neuron j, while s_k denotes the synaptic release variable, which can be modeled as the sigmoid activation of the presynaptic neuron output. E_{kj}^{rev} is the reversal potential of the chemical synapse. The parameter g_{kj}^{gap} denotes the electrical coupling conductance between neurons k and j, and the term $(V_k - V_j)$ represents the membrane potential difference driving the gap-junction current.

3.3 Embodied Interaction

As Figure 1 illustrates, the worm navigates in a two-dimensional truncated conical salt field:

$$C(x,y) = \max\Big(0,\ C_0 - \alpha\sqrt{(x-x_c)^2 + (y-y_c)^2}\Big), \tag{4}$$

where C_0, (x_c, y_c), and α denote the peak concentration, source location, and radial slope, respectively. At time step t, the policy receives

$$s_t = [C_t, \Delta C_t] \tag{5}$$

and outputs the continuous action

$$a_t = [a_0, a_1] = tanh(SMBV, SMBD), \tag{6}$$

where a_0 and a_1 correspond to the intensity of dorsal and ventral steering.

The worm body is modeled as a neuromechanical system with 50 coupled segments and 24 muscle blocks (6 head and 18 body muscles). Each segment contains rigid rods and elastic springs, where

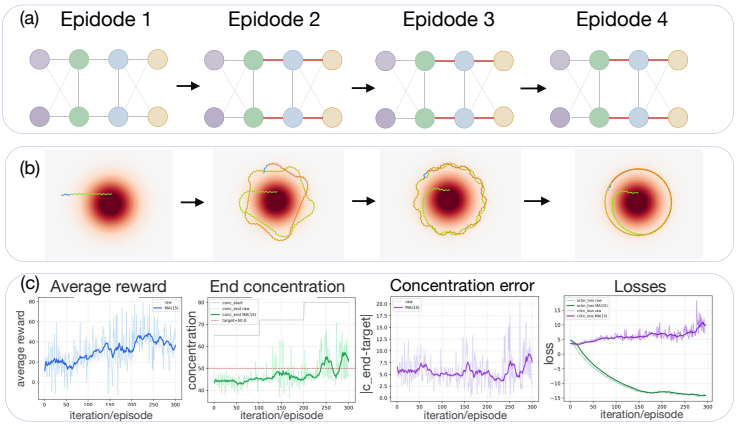


Figure 3: **Training process of chemotaxis policy.** Synaptic parameters are optimized while the circuit topology remains fixed. As training progresses, worm trajectories become more stable around the preferred salt region. The bottom panels report reward, terminal concentration, concentration error, and actor-critic losses during the training.

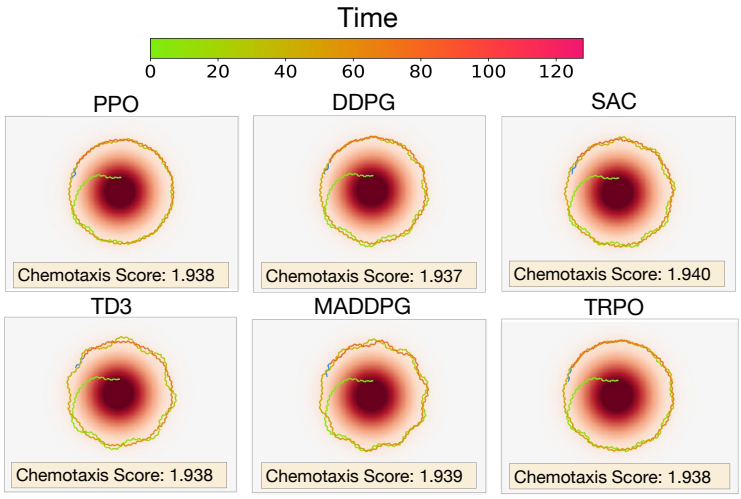


Figure 4: **Chemotactic behaviors learned by different reinforcement learning algorithms.** Trajectories show post-training navigation in the continuous salt field, with color indicating time. The chemotaxis score measures target approach and maintenance. SAC produces smoother contour following, whereas other methods exhibit larger oscillations or reduced stability.

dorsal–ventral muscle activation drives deformation and locomotion through environmental interaction. Stretch receptors convert body strain into proprioceptive currents for muscle control. The updated position determines the subsequent sensory input, forming the embodied feedback loop.

3.4 Reward Formulation

Chemotaxis is formulated as continuous regulation around the preferred concentration $C^* = 50\,\text{mM}$. Let $e_t = |C_t - C^*|$ denote the concentration error. The step-wise reward is

$$r_t = (e_{t-1} - e_t) + \alpha \mathbf{1}(e_t < 10) - \beta e_t \mathbf{1}(e_t < 4) - \lambda \|a_t\|_2^2 \tag{7}$$

where the first term rewards reductions in concentration error, the indicator terms encourage target detection and maintenance, and the action penalty discourages excessive steering. This dense concentration-error reward provides continuous guidance for gradient following and stable contour regulation.

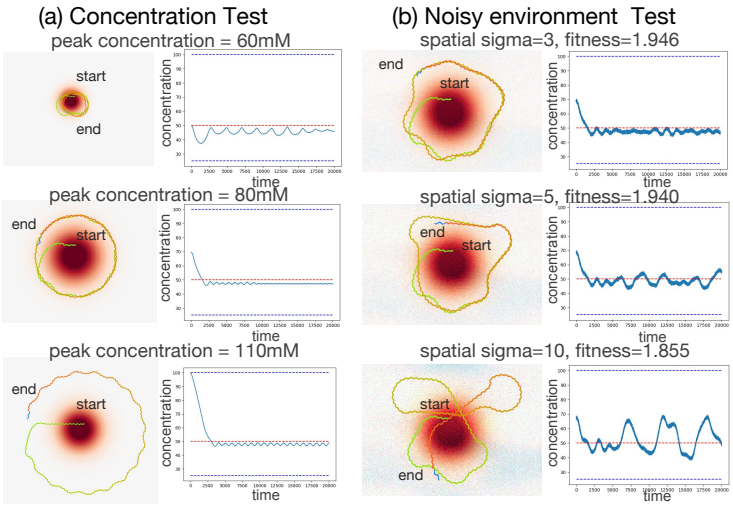


Figure 5: **Robustness test on concentration variation and noisy environment.** (a)Each pair shows the spatial trajectory and corresponding concentration profile over time. The red dashed line denotes the 50 mM target, and blue dashed lines indicate concentration bounds. (b) The learned policy is tested on noisy concentration.

3.5 Chemotaxis Fitness Score

In addition to the training reward, episode-level performance is measured using a chemotaxis fitness score that combines arrival efficiency and target maintenance:

$$F_{\text{chemotaxis}} = f_{\text{arrival}} + f_{\text{maintenance}}, \tag{8}$$

$$f_{\text{arrival}} = 1 - \frac{t_{\text{arrival}}}{T_{\text{sim}}}, \tag{9}$$

$$f_{\text{maintenance}} = 1 - \frac{1}{T_{\text{sim}} - t_{\text{arrival}}} \times \sum_{t=t_{\text{arrival}}}^{T_{\text{sim}}} \frac{|C(t) - C_{\text{pref}}|}{C_{\max} - C_{\min}}. \tag{10}$$

where t_{arrival} is the first time at which the worm reaches the preferred 50mM concentration and T_{sim} is the episode length.

3.6 Reinforcement Learning Optimization

The connectome-based actor is optimized using six continuous-control algorithms: Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), Twin Delayed Deep Deterministic Policy Gradient (TD3), Trust Region Policy Optimization (TRPO), and Multi-Agent Deep Deterministic Policy Gradient (MADDPG). All methods use the same connectome topology, observation and action spaces, reward function, and environmental settings; only their sampling and policy-update mechanisms differ. Each episode contains a fixed number of control steps, and a curriculum progressively increases the initial concentration to improve training stability. Algorithm-specific hyperparameters are provided in the supplementary material.

For on-policy optimization, PPO models the actor as a stochastic policy,

$$a_t \sim \pi_\theta(a_t|s_t), \tag{11}$$

and updates the policy using the clipped surrogate objective,

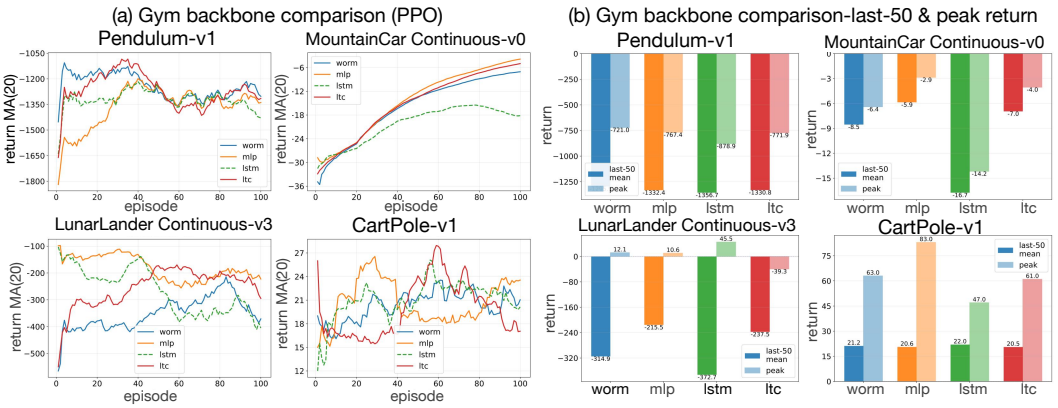


Figure 6: **Generalization of Connectome-as-Policy across standard reinforcement learning control benchmarks.** The proposed biologically grounded connectome policy (Worm) is compared with conventional policy backbones, including MLP, LSTM, and Liquid Time-Constant (LTC) networks, using PPO on four OpenAI Gym environments. Left panels show moving-average returns, while right panels report Last-50 and peak returns.

$$L_{\text{PPO}} = \mathbb{E}\left[\min\left(r_t A_t,\ \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t\right)\right], \tag{12}$$

where A_t is estimated using generalized advantage estimation.

For off-policy optimization, DDPG learns a deterministic actor

$$a_t = \mu_\theta(s_t), \tag{13}$$

together with an action-value critic

$$Q(s_t, a_t), \tag{14}$$

using experience replay and slowly updated target networks. Consequently, learning reshapes neural dynamics while preserving the structural organization of the connectome.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed Connectome-as-Policy framework on a continuous salt chemotaxis task in a two-dimensional field with a target concentration of 50 mM. The actor adopts the fixed biologically grounded chemotaxis circuit described in Section 3, while reinforcement learning optimizes only neuronal and synaptic parameters. We compare six RL algorithms under identical environments, observations, and rewards using episodic reward, terminal concentration error, and navigation trajectories. Each method is evaluated over ten independent runs with different random seeds.

4.2 Reinforcement Learning Optimizes the Connectome-Based Policy

We compare six representative reinforcement learning algorithms, and in all experiments, the biological connectome remains fixed, while only neuronal and synaptic parameters are optimized.

Learning Dynamics and Policy Evolution. Figure 2 shows the learning process of different reinforcement learning algorithms. Across all methods, the episodic reward increases steadily during training, while the concentration error decreases, demonstrating that reinforcement learning can successfully optimize the neuronal dynamics of the biological connectome. Among the evaluated

methods, SAC exhibits the fastest convergence and achieves the highest average reward, whereas PPO provides more stable optimization with lower variance throughout training.

Figure 3 illustrates how the connectome-based policy and its chemotactic behavior evolve during reinforcement learning. The circuit snapshots show that learning modifies synaptic parameters while preserving the predefined biological topology. Correspondingly, the navigation trajectories gradually transition from broad and irregular exploration to more compact contour-following motion around the preferred concentration region. The quantitative curves show an overall increase in average reward, while the terminal concentration approaches the target value of 50 mM and the concentration error decreases over most of training, despite fluctuations caused by stochastic exploration and curriculum changes. Additional comparison experiments and parameter ablation tests reported in the Appendix show that reinforcement learning achieves higher concentration fitness than evolutionary optimization, indicating more accurate chemotactic regulation.

Table 1: Performance comparison of reinforcement learning algorithms. Here, Avg. R denotes the average reward over the entire training process, Err._{50} denotes the average concentration error over the last 50 steps, and Conc._{50} denotes the average concentration over the last 50 steps.

Alg.	**Avg. R**	**Err.$_{50}$**	**Conc.$_{50}$**
PPO	23.86 ± 4.15	7.71 ± 0.71	51.49 ± 1.75
DDPG	26.90 ± 3.96	9.02 ± 2.00	54.50 ± 3.70
SAC	28.54 ± 1.90	6.95 ± 2.07	50.50 ± 3.03
TD3	24.14 ± 5.52	8.28 ± 2.19	52.69 ± 3.51
TRPO	23.38 ± 3.37	10.20 ± 2.66	55.38 ± 4.38
MADDPG	24.47 ± 5.41	8.48 ± 2.18	53.31 ± 4.03

Learned Navigation Behaviors. The learned navigation behaviors are further illustrated in Figure 4. All algorithms acquire gradient-following behavior and successfully guide the simulated worm toward the preferred salt concentration region. However, clear differences are observed in contour-following stability. SAC and TRPO produce smoother trajectories with fewer oscillations around the target contour, whereas PPO and DDPG generate slightly larger steering fluctuations. TD3 and MADDPG exhibit less consistent contour tracking, resulting in larger trajectory deviations.

Quantitative and Statistical Comparison. Table 1 and Appendix Table 5 quantitatively compare the optimization performance of all algorithms. SAC achieves the highest average reward during both the entire training process and the final concentration closest to the target value of 50 mM, indicating superior learning efficiency. In contrast, TD3 produces the best final 50-episode reward and smallest final concentration error, demonstrating more accurate long-term chemotactic regulation. One-way ANOVA further indicates that significant differences are observed only for the final concentration error ($F(5, 54) = 2.64$, $p = 0.034$), whereas the remaining metrics do not show statistically significant differences. Overall, these results demonstrate that reinforcement learning can effectively optimize neuronal and synaptic parameters within a fixed biological connectome. The optimized circuit acquires stable chemotactic behavior through continuous interaction with the environment.

Robustness Analysis. Figure 5 evaluates the learned policy under different initial concentrations and environmental noise levels. For initial concentrations of 60, 80, and 110 mM, the worm approaches the 50 mM target contour within approximately $2.5\text{-}3.0 \times 10^3$ simulation steps and subsequently maintains concentrations near 46-48 mM, corresponding to a steady-state error of roughly 2-4 mM. Under low noise, the concentration remains largely within 46-50 mM, whereas moderate and strong noise increase the fluctuation ranges to approximately 44-56 mM and 39-69 mM, respectively. Although stronger perturbations produce increasingly distorted trajectories and larger concentration oscillations, the agent repeatedly returns to or crosses the target contour without diverging from the bounded region. These results indicate that the connectome-based policy is robust to variations in initial conditions and exhibits graceful performance degradation under increasing environmental uncertainty. Supplementary results further demonstrate that dense concentration-difference shaping rewards are better suited for contour-following chemotaxis than discrete goal-based rewards.

4.3 Cross-Task Applicability of Connectome-Based Policies

To investigate whether the connectome-based policy representation can be applied beyond the original chemotaxis task, we evaluate the proposed connectome architecture on four standard continuous control benchmarks, including Pendulum, MountainCarContinuous, LunarLanderContinuous, and CartPole. The WORM policy is compared with conventional neural architectures, including MLP, LSTM, and Liquid Time-Constant (LTC) networks under identical PPO training settings.

Table 2: Comparison of policy backbones on four standard control benchmarks. Peak denotes the highest return achieved during training. Higher returns are better. Parameter counts vary slightly across tasks.

Environment	Backbone	Peak	# Parameters
4*Pendulum	Worm	$\mathbf{-721.1}$	∼350–430
	MLP	$\underline{-767.4}$	∼4.5k–4.9k
	LSTM	-878.9	∼55k–56k
	LTC	-771.9	∼29k–30k
4*MountainCar	Worm	-6.40	∼350–430
	MLP	$\mathbf{-2.87}$	∼4.5k–4.9k
	LSTM	-14.18	∼55k–56k
	LTC	$\underline{-4.05}$	∼29k–30k
4*LunarLander	Worm	$\underline{12.1}$	∼350–430
	MLP	10.6	∼4.5k–4.9k
	LSTM	$\mathbf{45.5}$	∼55k–56k
	LTC	-39.3	∼29k–30k
4*CartPole	Worm	$\underline{63}$	∼350–430
	MLP	$\mathbf{83}$	∼4.5k–4.9k
	LSTM	47	∼55k–56k
	LTC	61	∼29k–30k

Figure 6 and Table 2 summarize the learning curves and final performance across different environments. The connectome-based policy achieves competitive performance with significantly fewer parameters compared with generic recurrent and continuous-time neural architectures. In Pendulum, WORM obtains comparable control performance to MLP and LTC, while maintaining more stable optimization dynamics. In MountainCarContinuous, all architectures gradually improve during training, with MLP and LTC achieving slightly higher final returns. For LunarLanderContinuous, the performance gap between different policies becomes larger, indicating that tasks requiring high-dimensional state integration remain challenging for the compact connectome representation. In CartPole, WORM achieves comparable performance to other baselines despite its substantially smaller parameter space. Additional comparisons between WORM policy and RNN/GRU on both chemotaxis and Gym tasks are in the Appendix.

These results demonstrate that the connectome-based architecture is not limited to chemotaxis control but can serve as a compact structured policy representation for diverse reinforcement learning tasks. It also provides advantages of low-dimensional structured dynamics, parameter efficiency, and biological interpretability.

5 Conclusion

We presented *Connectome-as-Policy*, a biologically grounded reinforcement learning framework that represents an experimentally derived *C. elegans* chemotaxis circuit as a structured policy. By preserving the circuit topology and optimizing neuronal and synaptic parameters, the proposed framework learns adaptive chemotactic behavior through closed-loop interaction. Experiments demonstrate that the connectome-based policy achieves competitive control with substantially fewer trainable parameters than conventional neural backbones while providing circuit-level interpretability. Furthermore, continuous concentration-based rewards improve learning stability for chemotactic regulation. These results suggest that biological connectomes can serve as effective policy priors for reinforcement learning and provide a promising direction toward larger connectome-based controllers and embodied intelligent systems.

AI use statement

In this work, we used generative AI tools for polish writing.

REFERENCES

Cornelia I Bargmann. Chemosensation in c. elegans. *WormBook: The online review of C. elegans biology [Internet]*, 2006.

Cornelia I Bargmann and H Robert Horvitz. Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in c. elegans. *Neuron*, 7(5):729–742, 1991.

Matthew Beverly, Sriram Anbil, and Piali Sengupta. Degeneracy and neuromodulation among thermosensory neurons contribute to robust thermosensory behaviors in caenorhabditis elegans. *Journal of Neuroscience*, 31(32):11718–11727, 2011.

Nikhil Bhattasali, Anthony M Zador, and Tatiana Engel. Neural circuit architectural priors for embodied control. *Advances in neural information processing systems*, 35:12744–12759, 2022.

Zhenshan Bing, Claus Meschede, Florian Röhrbein, Kai Huang, and Alois C Knoll. A survey of robotics control based on learning-inspired spiking neural networks. *Frontiers in neurorobotics*, 12:35, 2018.

Jordan H Boyle, Stefano Berri, and Netta Cohen. Gait modulation in c. elegans: an integrated neuromechanical model. *Frontiers in computational neuroscience*, 6:10, 2012.

Beth L Chen, David H Hall, and Dmitri B Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences*, 103(12):4723–4728, 2006.

Zhongyu Chen, Yuguo Yu, and Xiangyang Xue. A connectome-based digital twin caenorhabditis elegans capable of intelligent sensorimotor behavior. *Mathematics*, 11(11):2442, 2023.

Miles Walter Churchland and Jordi Garcia-Ojalvo. Reinforcement learning in densely recurrent biological networks. *iScience*, 29(1), 2026.

Chuntao Dan, Brad K Hulse, Ramya Kappagantula, Vivek Jayaraman, and Ann M Hermundstad. A neural circuit architecture for rapid learning in goal-directed navigation. *Neuron*, 112(15): 2581–2599, 2024.

AV Demin and EE Vityaev. Learning in a virtual model of the c. elegans nematode for locomotion and chemotaxis. *Biologically Inspired Cognitive Architectures*, 7:9–14, 2014.

Romain Franconville, Celia Beron, and Vivek Jayaraman. Building a functional connectome of the drosophila central complex. *Elife*, 7:e37017, 2018.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Christopher V Gabel, Harrison Gabel, Dmitri Pavlichin, Albert Kao, Damon A Clark, and Aravinthan DT Samuel. Neural circuits mediate electrosensory behavior in caenorhabditis elegans. *Journal of Neuroscience*, 27(28):7586–7596, 2007.

Padraig Gleeson, David Lung, Radu Grosu, Ramin Hasani, and Stephen D Larson. c302: a multiscale framework for modelling the nervous system of caenorhabditis elegans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1758):20170379, 2018.

Andrew Gordus, Navin Pokala, Sagi Levy, Steven W Flavell, and Cornelia I Bargmann. Feedback from network states generates variability in a probabilistic olfactory circuit. *Cell*, 161(2):215–227, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. A natural lottery ticket winner: Reinforcement learning with ordinary neural circuits. In *international conference on machine learning*, pp. 4082–4093. PMLR, 2020.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7657–7666, 2021.

Nicolas Heess, Greg Wayne, Yuval Tassa, Timothy Lillicrap, Martin Riedmiller, and David Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv:1610.05182*, 2016.

Atil Iscen, Ken Caluwaerts, Jie Tan, Tingnan Zhang, Erwin Coumans, Vikas Sindhwani, and Vincent Vanhoucke. Policies modulating trajectory generators. In *Conference on robot learning*, pp. 916–926. PMLR, 2018.

Eyal Itskovits, Rotem Ruach, Alexander Kazakov, and Alon Zaslaver. Concerted pulsatile and graded neural dynamics enables efficient chemotaxis in c. elegans. *Nature communications*, 9(1): 2866, 2018.

Eduardo J Izquierdo and Randall D Beer. Connecting a connectome to behavior: an ensemble of neuroanatomical models of c. elegans klinotaxis. *PLoS computational biology*, 9(2):e1002890, 2013.

Eduardo J Izquierdo and Randall D Beer. An integrated neuromechanical model of steering in c. elegans. In *Artificial Life Conference Proceedings*, pp. 199–206. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2015.

Eduardo J Izquierdo and Shawn R Lockery. Evolution and analysis of minimal neural circuits for klinotaxis in caenorhabditis elegans. *Journal of Neuroscience*, 30(39):12908–12917, 2010.

Reece Keller, Alyn Kirsch, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Intrinsic goals for autonomous agents: Model-based exploration in virtual zebrafish predicts ethological behavior and whole-brain dynamics. *Advances in Neural Information Processing Systems*, 38: 35872–35903, 2026.

Jimin Kim and Eli Shlizerman. Deep reinforcement learning for neural control. *arXiv preprint arXiv:2006.07352*, 2020.

Jimin Kim, William Leahy, and Eli Shlizerman. Neural interactome: Interactive simulation of a neuronal system. *Frontiers in computational neuroscience*, 13:8, 2019.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Hirofumi Kunitomo, Hirofumi Sato, Ryo Iwata, Yohsuke Satoh, Hayao Ohno, Koji Yamada, and Yuichi Iino. Concentration memory-dependent synaptic plasticity of a taste circuit regulates salt concentration chemotaxis in caenorhabditis elegans. *Nature communications*, 4(1):2210, 2013.

Mathias Lechner, Radu Grosu, and Ramin M Hasani. Worm-level control through search-based reinforcement learning. *arXiv preprint arXiv:1711.03467*, 2017.

Mathias Lechner, Ramin Hasani, Manuel Zimmer, Thomas A Henzinger, and Radu Grosu. Designing worm-inspired neural networks for interpretable robotic control. In *2019 international conference on robotics and automation (ICRA)*, pp. 87–94. IEEE, 2019.

Mathias Lechner, Ramin Hasani, Alexander Amini, Thomas A Henzinger, Daniela Rus, and Radu Grosu. Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence*, 2(10): 642–652, 2020.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Timothy Paul Lillicrap, Jonathan James Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daniel Pieter Wierstra. Continuous control with deep reinforcement learning, September 15 2020. US Patent 10,776,692.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Linjiao Luo, Quan Wen, Jing Ren, Michael Hendricks, Marc Gershow, Yuqi Qin, Joel Greenwood, Edward R Soucy, Mason Klein, Heidi K Smith-Parker, et al. Dynamic encoding of perception, memory, and movement in a c. elegans chemotaxis circuit. *Neuron*, 82(5):1115–1128, 2014.

Liqun Luo. Architectures of neuronal circuits. *Science*, 373(6559):eabg7285, 2021.

Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, and Nicolas Heess. Neural probabilistic motor primitives for humanoid control. *arXiv preprint arXiv:1811.11711*, 2018.

Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. *arXiv preprint arXiv:1911.09451*, 2019.

BA Mitchell and Linda R Petzold. Control of neural systems at multiple scales using model-free, deep reinforcement learning. *Scientific reports*, 8(1):10721, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Keita Mori, Naohiro Yamauchi, Haoyu Wang, Ken Sato, Yu Toyoshima, and Yuichi Iino. Probabilistic generative modeling and reinforcement learning extract the intrinsic features of animal behavior. *Neural networks*, 145:107–120, 2022.

Jonathan T Pierce-Shimomura, Thomas M Morse, and Shawn R Lockery. The fundamental role of pirouettes in caenorhabditis elegans chemotaxis. *Journal of Neuroscience*, 19(21):9557–9569, 1999.

Joelle Pineau, Arthur Guez, Robert Vincent, Gabriella Panuccio, and Massimo Avoli. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *International journal of neural systems*, 19(04):227–240, 2009.

Libânia Queirós, Carlos Marques, Joana L Pereira, Fernando JM Gonçalves, Michael Aschner, and Patrícia Pereira. Overview of chemotaxis behavior assays in caenorhabditis elegans. *Current protocols*, 1(5):e120, 2021.

Pavan Ramdya and Auke Jan Ijspeert. The neuromechanics of animal locomotion: From biology to robotics and back. *Science Robotics*, 8(78):eadg0279, 2023.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Eli Shlizerman, Jeffrey A Riffell, and J Nathan Kutz. Data-driven inference of network connectivity for modeling the dynamics of neural codes in the insect antennal lobe. *Frontiers in computational neuroscience*, 8:70, 2014.

Zu Soh, Kazuma Sakamoto, Michiyo Suzuki, Yuichi Iino, and Toshio Tsuji. A computational model of internal representations of chemical gradients in environments for chemotaxis of caenorhabditis elegans. *Scientific Reports*, 8(1):17190, 2018.

Balázs Szigeti, Padraig Gleeson, Michael Vella, Sergey Khayrulin, Andrey Palyanov, Jim Hokanson, Michael Currie, Matteo Cantarelli, Giovanni Idili, and Stephen Larson. Openworm: an open-science approach to modeling caenorhabditis elegans. *Frontiers in computational neuroscience*, 8:137, 2014.

Jiaxin Wang, Mengxiao Zhang, Kejun Wang, Lijun Kang, Liang Feng, Gang Pan, and Huajin Tang. Neurosimworm: A multisensory framework for modeling and simulating neural circuits of caenorhabditis elegans. *Neurocomputing*, 638:130055, 2025.

John Graham White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode caenorhabditis elegans. *Philosophical transactions of the royal society of london. b, biological sciences*, 314(1165):1–340, 1986.

Shoichiro Yamaguchi, Honda Naoki, Muneki Ikeda, Yuki Tsukada, Shunji Nakano, Ikue Mori, and Shin Ishii. Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS computational biology*, 14(5):e1006122, 2018.

Mengdi Zhao, Ning Wang, Xinrui Jiang, Xiaoyang Ma, Haixin Ma, Gan He, Kai Du, Lei Ma, and Tiejun Huang. An integrative data-driven model simulating c. elegans brain, body and environment interactions. *Nature Computational Science*, 4(12):978–990, 2024.

Junfeng Zuo, Ying Nian Wu, Si Wu, and Wenhao Zhang. The motion planning neural circuit in goal-directed navigation as lie group operator search. 2024.