Computational Model of the Hippocampus Drives Exploratory Behaviour in Reinforcement Learning Agents

Student: Sabrina Du[1,2] - Supervisor: Adrien Peyrache[1]

Collaborators: Adel Halawa[1,2], Aleksei Efremov[1,2], Daniel Levenstein[3],

and Blake Richards[1,2,4,5]

10. April 2026

Affiliations: [1] Montreal Neurological Institute, McGill University, Montreal, QC, Canada. [2] Mila, Montreal, QC, Canada. [3] Department of Neuroscience, Yale School of Medicine, New Haven, Connecticut, United States of America. [4] School of Computer Science, McGill University, Montreal, QC, Canada. [5] Learning in Machines and Brains Program, CIFAR, Toronto ON, Canada.

Acknowledgements: We thank all members of the Peyrache Lab for their encouragement and the Richards' Lab for providing access to the Mila cluster. We also acknowledge the use of Claude Code (Anthropic, 2026) to assist in debugging code and optimizing data analysis scripts used in this manuscript. All suggestions were manually reviewed and approved by the authors.

Abstract

The hippocampus supports spatial navigation, memory, and planning through the formation of a cognitive map: a structured environmental representation reflected in its neural activity. These neural dynamics, and the behaviors that arise from them, can be modelled computationally using artificial neural networks (ANNs). However, these models typically leverage reinforcement learning (RL) using external rewards, failing to capture the intrinsic drive of freely exploring animals in the absence of external rewards. This study aims to investigate whether a recurrent neural network (RNN) exhibiting hippocampal-like activity can drive exploratory behavior in reward-free RL agents. We leveraged an existing RNN trained for sensory sequence prediction, exhibiting hippocampal-like activity patterns, and used its prediction error as the intrinsic reward to train an Actor-Critic agent. In a Novel Object Recognition task (NOR), the RL agent occupied the region of interest significantly more than a random control (peak at 37.2% ± 1.8% vs. 15.0% ± 1.2% [mean ± SEM], Welch's t-test, $p < 0.01$). This preference persisted in a multi-room environment, where the agent spent 8.2% ± 0.2% of its time in the most distant, novel room, compared to 0.023% ± 0.017% of the random control ($p < 0.001$). These results demonstrate that cognitive maps derived from sensory prediction can drive exploration and novelty-seeking, providing a computational framework for how hippocampal dynamics guide navigation in the absence of external incentives.

Significance Statement

Hippocampal representations form a cognitive map. However, the way in which cognitive maps are read out to support behavior involving spatial exploration and novelty detection remains unclear. While recent work leverages external-reward-driven reinforcement learning (RL) to answer such questions, we propose combining intrinsically motivated RL and existing hippocampal-like ANNs to address them instead. In reward-absent environments, we show that the hidden state of

an RNN capable of forming a cognitive map is sufficient to drive RL agents towards novel objects in familiar environments and novel locations in unfamiliar ones. The training and architectural scheme that we present therefore opens avenues for investigating how the cognitive map is sampled to guide behavior and how its structure is shaped by exploratory demands.

Introduction

In rodents, the hippocampus is believed to play a key role in novelty detection networks (Knight, 1996; Eichenbaum, 1999; Kafkas and Montaldi, 2018), with both spatial and non-spatial sources of novelty eliciting shifts in hippocampal population activity. For example, the hippocampus was found to support non-spatial novelty encoding, like recognizing novel objects (Tulving et al., 1994; Kanwisher et al., 1996; Stern et al., 1996). In rodents, the Novel Object Recognition (NOR) test is widely used to examine their non-spatial memory because it leverages these animals' tendency to explore and spend more time around novel objects (Broadbent et al., 2010; Barker and Warburton, 2011). While rodent hippocampal function in recognition memory remains controversial, reviews of the literature have established the necessity of the hippocampus for NOR under standard conditions (Broadbent et al., 2010; Antunes and Biala, 2011; Cohen and Jr., 2014). There is instead much more consensus on the hippocampus' contribution to spatial memory and novelty. While hippocampal representations of novel environments emerge rapidly after first exposure (Hill, 1978; Wilson and McNaughton, 1993; Frank et al., 2004; Leutgeb et al., 2004), they differ from the neural activity measured from rodents in familiar environments. In fact, it has been shown that CA1 neurons exhibit higher-amplitude place field spiking (Cohen et al., 2017) and larger changes in place-field structure (Frank et al., 2004) when exploring previously unvisited locations.

In novel environments, animals exhibit incredibly flexible behavior. It has become widely accepted that hippocampal cognitive maps are essential to support such behavior (Tolman, 1948;

O'Keefe and Nadel, 1978; Behrens et al., 2018; O'Keefe, 2025). Cognitive maps are formed from the correlated activity of neuronal populations. Although the properties of such population activity is still debated, there are three key features that persist when discussing neural activity from which emerge cognitive maps: these neural representations (1) convey information about an animal's position in physical space (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978), (2) reflect the environment's topology (Dabaghian et al., 2012; Behrens et al., 2018), and (3) display attractor dynamics (Samsonovich and McNaughton, 1997; Tsodyks, 1999). Despite much previous and ongoing research, we still do not know how - both on the functional and mechanistic levels - information is extracted from neural cognitive maps to inform exploratory and flexible behavior.

To address such broad questions, the cognitive neuroscience field has also begun leveraging artificial neural networks (ANNs) trained to accomplish similar tasks as model organisms like rodents. Specifically, recent work aims to draw insights about biological neural function by studying which training objectives, architecture design choices and data constraints lead to the emergence of ANN activity that is similar to biological neural activity (Richards et al., 2019; Saxe et al., 2021; Doerig et al., 2023). Recurrent neural networks (RNNs) have become the architecture of choice for the computational modelling of hippocampal activity because of the similarity between their inherent feedback loops - involving hidden state vectors - and biological neural circuits. While the training objectives vary in their formulations, they often involve future sensory or state prediction (Recanatesi and others, 2021; Fang and Stachenfeld, 2024; Jensen et al., 2024; Levenstein et al., 2024) and external-reward-based RL (Geerts et al., 2020; Pedamonti et al., 2025). RL grounded in external rewards fails to model the intrinsic drive that powers rodent exploratory behavior, thus limiting the comparisons we can draw between ANNs trained this way and their biological counterparts.

Instead, we propose to use curiosity-driven RL to investigate how hippocampal representations contribute to exploratory behavior. Curiosity-driven RL was designed to solve sparse reward problems, using prediction error as a reward signal and therefore formulating rewards in a way that is *intrinsic* to the agent (Pathak et al., 2017; Burda et al., 2019a; Ladosz et al., 2022). Given this training paradigm and the hippocampus' role in novelty detection, we hypothesized that the previously mentioned predictive RNNs, which exhibit hippocampal-like activity, can drive exploratory behavior in RL agents. Specifically, Levenstein et al. (2024) demonstrated that an RNN trained on sequential sensory prediction can form a cognitive map, and Stachenfeld et al. (2014) showed that a cognitive map representation of space can support efficient reinforcement learning. Combining all this information, we ask whether the representations of an RNN capable of producing a cognitive map are sufficient to drive RL agents towards novel objects and locations in external-reward-free environments (Fig. 1). We find that this is indeed the case and propose that this framework allows for targeted investigations into how the structure of the cognitive map is shaped by, and in turn supports, the flexible sampling strategies that underlie exploration.

Materials and Methods

This investigation aims to study the behavior of RL agents receiving the activity of hippocampal-like networks as input. To do so, experiments require a hippocampal-like network, an RL agent, a custom training scheme, a computational environment where behaviorally relevant tasks can be completed and metrics to analyze exploratory behavior. We describe these components below.

Hippocampal-like network architecture. All experiments were conducted with Levenstein et al. (2024)'s RNN architecture and core codebase. The authors implemented multiple architectures of a predictive recurrent neural network (pRNN), a network trained to predict future sensory observations based on an action input. Originally, the actions were sampled from a fixed distribution (Fig. 2A). Building on top of the *k-masked* pRNN architecture, we introduce an action

network for action selection (Fig. 2B). The action selection strategy defines whether or not we consider the pRNN to be curious or random. Specifically, we consider a pRNN trained with actions sampled from a fixed distribution as a *random pRNN-agent*, and we call a pRNN trained with actions computed by an action neural network a *curious pRNN-agent*.

A k-masked pRNN receives an observation and an action at every $k+1$ timestep. At each timestep t for k timesteps, the network must predict the sensory observation (Fig. 3B) after receiving only an action a_t as input (Fig. 2B). The pRNN follows the following general architecture, with variables described in Table 2:

$$h_t = \text{NormReLU}(W_{\text{rec}} h_{t-1} + W_{\text{in}} o_t + W_{\text{act}} a_t) \tag{1}$$

$$y_t = \text{Sigmoid}(W_{\text{out}} h_t) \tag{2}$$

With activation functions,

$$\text{NormReLU}(\boldsymbol{x}) = \max\left(0, \frac{\boldsymbol{x} - E[\boldsymbol{x}]}{\text{Std}[\boldsymbol{x}] + \varepsilon} + \boldsymbol{b} + \mathcal{N}(0, \eta)\right) \tag{3}$$

$$\text{Sigmoid}(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{x}}}$$

Levenstein et al. (2024) designed multiple metrics to demonstrate that a random pRNN-agent can develop hippocampal-like activity. We used two to ensure that the curious pRNN-agent's hidden states also resembled hippocampal activity: spatial representational similarity analysis (sRSA) (Fig. 3C1) and Sleep-Wake distance (SW-dist) (Fig. 3C2).

sRSA computes the Spearman rank correlation $\rho(D_E(x_t, x_{t'}), D_C(h_t, h_{t'}))$ between a distance in physical space and a distance in pRNN activation space (i.e., neural space). Specifically, $D_E(x_t, x_{t'})$ is the Euclidean distance between all pairs of positions in the environment.

$D_C(h_t, h_{t'})$ is the cosine dissimilarity between the hidden states at all corresponding pairs of timepoints h_t and $h_{t'}$. The authors interpreted sRSA as the strength of the cognitive map's alignment with the topology of the environment; a larger sRSA indicated a stronger cognitive map.

SW-dist measures the average cosine dissimilarity between the hidden states during a sleep and wake trial, defined in (Levenstein et al., 2024). Since the hippocampus generates plausible trajectories in neural space during offline periods like sleep, the distance between representations in a sleep-like state and a wakeful state should be small in artificial networks modelling the hippocampus. For further information on trial classification and other metrics, please consult (Levenstein et al., 2024).

Action network architecture. We formulate our RL agent as an Actor-Critic (AC) network (Sutton and Barto, 2018; Zhang and Ma, 2018). The Actor (π_θ) and Critic (V_θ) are both Multi-Layer Perceptrons (MLPs). At each timestep, the agent receives the pRNN's hidden state $\boldsymbol{h}_t$ and selects an action by sampling from the actor's output distribution over the discrete action space $\mathcal{A}$: $a_t \sim \pi_\theta(\cdot \mid \boldsymbol{h}_t)$. The Actor and Critic follow the structure below (See Table 3 for variable definitions) and operate on the action space:

$$\mathcal{A} = \begin{cases} 0 \text{ for turn left} \\ 1 \text{ for turn right} \\ 2 \text{ for move forward} \\ 3 \text{ for stay put} \end{cases}$$

$$\pi(a_t \mid \boldsymbol{h_t}) = \text{softmax}(W_{\pi 2} \tanh(W_{\pi 1} \boldsymbol{h_t} + b_{\pi 1}) + b_{\pi 2}) \tag{4}$$

$$V(\boldsymbol{h_t}) = W_{V2} \tanh(W_{V1} \boldsymbol{h_t} + b_{V1}) + b_{V1} \tag{5}$$

The actor produces logits over the action space ($\mathcal{A}$), which parameterize a categorical distribution for action selection. The critic estimates the value $V(h)$ of each state embedding h, which is used to compute the advantage - the contribution of a specific action relative to the expected return.

Training objectives. The pRNN (parameterized by φ) and Actor-Critic (parameterized by θ) are trained simultaneously but with separate optimizers and no shared gradients, ensuring that π_θ cannot select actions that artificially reduce $\mathcal{L}_{\text{RNN}}$ without genuine exploration. The pRNN parameters φ are updated via supervised next-observation prediction (Mean Squared Error) on each collected rollout of observations. Specifically, the pRNN predicts the next observation $f_\varphi(\boldsymbol{h}_t, a_t) = \hat{o}_t$ and the MSE loss is:

$$\mathcal{L}_{\text{RNN}} = \mathbb{E}_t\left[(\hat{o}_t - o_t)^2\right] \tag{6}$$

This prediction error serves as an intrinsic curiosity reward signal to the action network. As such, states that are harder to predict are more rewarding, encouraging exploration:

$$r_t = \frac{\eta}{2}(\hat{o}_{t+1} - o_{t+1})^2 \tag{7}$$

The action network is trained with Proximal Policy Optimization (Schulman et al., 2017), with advantages estimated via Generalized Advantage Estimation (Schulman et al., 2016). The PPO objective is the following:

$$\mathcal{L}_{\text{AC}} = -\mathbb{E}_t\left[\min\left(\rho_t \hat{A}_t, \text{clip}(\rho_t, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right] + c_V \mathbb{E}_t\left[(V(\boldsymbol{h}_t) - G_t)^2\right] - c_H \mathcal{H}[\pi] \tag{8}$$

where $\rho_t = \frac{\pi_{\theta(a_t \mid \boldsymbol{h}_t)}}{\pi_{\theta_{\text{old}}(a_t \mid \boldsymbol{h}_t)}}$ is the probability ratio, $\hat{A}_t$ the estimated advantage, $\mathcal{H}$ the Shannon entropy, G_t the return, and c_V, c_H are coefficients weighting the value and entropy terms respectively. Moreover, training hyperparameters can be found in Table 1.

Experimental design and behavioral tasks. We trained the pRNN-agents (random and curious) in gridworld environments designed by the Farama Foundation (Chevalier-Boisvert et al., 2023). We designed two tasks with different environments to probe two kinds of novelty - absolute and contextual novelty - because it has been shown that the neural substrates that are recruited for novelty processing depend on the type of novelty at hand (Kafkas and Montaldi, 2018; Frank and Kafkas, 2021). A 16x16 L-Room (Fig. 4A) with colored floor tiles serving as ambiguous visual cues, due to their symmetric structure, was used for the contextual novelty task. A multi-room (Fig. 7) environment with four rooms, similar colored floor tiles and three wall openings, each leading to a subsequent sub-room, was used for the absolute novelty task. Each pRNN-agent executed 256 steps in the environment, which constitutes 1 trajectory.

Contextual novelty and the Novel Object Recognition task. Inspired by the classic NOR task in rodents (Lueptow, 2017), a curious pRNN-agent was trained with a two-phase protocol in a single-room L-Room environment. In Phase 1, the curious pRNN-agent learned the environment without any novel object, allowing the pRNN to develop spatial representations of the familiar room. In Phase 2, a bright green novel object was introduced at one of three target locations within the room (Fig. 4A). After Phase 1, we saved the trained curious pRNN-agent checkpoints and used them to initialize further training of the models on the Phase 2 environment with a novel object, allowing us to measure the curious pRNN-agent's novel object preference.

To isolate novel object preference (behavior of the agent) and learning (predictions of the pRNN), we first compared two curious pRNN-agent pairs: (1) an experimental curious pRNN-agent trained in the familiar L-Room and then in the L-Room with a novel object, and (2) a control curious pRNN-agent solely trained on the L-Room environment for periods equivalent to Phase 1 and Phase 2 training periods. Then, we used a third pRNN-agent, a random one that also underwent Phase 1 and Phase 2 training, as another control. Three novel object locations

- $(7, 11), (7, 2)$ and $(14, 7)$ - were tested with all three pRNN-agent pairs (Fig. 6A, Fig. 6C, Fig. 6E), with at least $n = 20$ independent random seeds per pRNN-agent.

Absolute novelty and the multi-room task. In the multi-room task, a curious pRNN-agent was trained in a four-room MiniGrid environment in which one room (Room 4) was structurally harder to reach than the others, requiring the agent to navigate through openings in otherwise impassable walls (Fig. 7A). The pRNN-agent began each trajectory in a random location with a randomly chosen head direction in Room 1, and its visitation frequency for each novel sub-room was compared against a random pRNN-agent's visitation frequency. Due to computational constraints, n=1 runs were used per agent. Statistical comparisons were therefore performed across the final 8000 trajectories in the environment of each agent, reflecting within-agent behavioral stability rather than across-seed reproducibility.

Statistical analysis for pRNN activity and pRNN-agent behavior. To study pRNN activity, we focused our analysis on the network's hidden state vector (h). Besides the sRSA and SW-dist metrics we previously discussed, we also used the Pynapple package (Viejo et al., 2023) to compute the spatial information (I_i, in bits) of each unit (Skaggs et al., 1992), or neuron, (h_i) of the hidden state, defined as:

$$I_i = \sum_x h_i \frac{x}{h} \log_2 h_i \frac{x}{h} p(x) \tag{9}$$

To study pRNN-agent behavior, we first measure mutual information between agent state (position and head direction) and the action that is taken ($I(S, A)$). Mutual information quantifies how much knowing the agent's state reduces uncertainty about which action it takes. Theoretically, a random pRNN-agent should have zero mutual information $I(S, A)$ because it takes actions independent of state. In our case, the mutual information is nonzero because of limited location

and hidden state sampling. Conversely, a curious pRNN-agent should have a higher $I(S, A)$ because it has learned a state-dependent policy. Formally, mutual information is defined as:

$$\begin{aligned} I(S;A) &= H(S) + H(A) - H(S,A) \\ &= -\sum_{s} p(s)\log_2 p(s) - \sum_{a} p(a)\log_2 p(a) + \sum_{s,a} p(s,a)\log_2 p(s,a) \end{aligned} \tag{10}$$

where H is Shannon entropy, $p(s, a)$ is the joint probability estimated from the visitation count and $p(s)$ and $p(a)$ are marginals obtained by summing over the other variable.

For the NOR contextual novelty task, we measure the probability of occupying the region of interest (ROI), an L2 radius of 3 cells around the novel object, as exposure to the novel object increases. This probability is defined as:

$$P_{\text{ROI}} = \frac{\sum_{(x,y)\in \text{ ROI}} \text{occupancy\_count}(x, y)}{\sum_{(x,y)} \text{occupancy\_count}(x, y)} \tag{11}$$

We run 20 separate seeds for the experimental curious pRNN-agent, control curious pRNN-agent and random pRNN-agent in the NOR task. We use Welch's two-sample t-test to determine the significance of the difference between P_{ROI} of both pairwise experimental-control condition combinations: Curious vs. Curious Control and Curious vs. Random. Confidence intervals for the probabilities in each condition are computed with a bootstrap 95% CI with $n = 1000$ resamples.

For the multi-room absolute novelty task, we measured the percentage of time each agent spent in each sub-room, averaged over the last 8000 training trajectories of a single run per agent, once the curious pRNN-agent's policy had stabilized. Due to computational constraints, only a single seed was used per agent; statistical comparisons were therefore performed across these final trajectories, reflecting within-agent behavioral stability rather than across-seed reproducibility. Room visitation frequencies were compared between agents using Welch's two-sided t-tests.

Software and code availability. Training and analysis code will be released at https://github.com/SabrinaDu7/RL_for_pRNN. For Farama-Minigrid environments, we used https://github.com/SabrinaDu7/minigrid. pRNN implementations can be found at https://github.com/LevensteinLab/pRNN.

Results

NOR contextual novelty task. As described in the Methods, we trained both random and curious pRNN-agents in an L-Room to accomplish Phase 1 of the task, where agents learn the environment before novel object placement. Levenstein et al. (2024) showed that random pRNN-agents can develop representations with hippocampal-like properties. Phase 1 training validates that the curious pRNN-agent can develop such spatial representations as well (Fig. 3). Both pRNN-agents learned to predict future sensory observations after receiving action alone, as shown in the decreasing MSE loss curves (Fig. 3 C1). The curious pRNN-agent's loss is higher than the random's because the former is selecting actions that maximize pRNN prediction error. However, spatial representational similarity analysis (sRSA) over the hidden states of both agents reveals that the correlation between physical position and hidden state representations is weaker in the curious pRNN-agent than in the random agent early in training, likely reflecting the random agent's greater spatial coverage (Fig. 3 C3). Both agents' sRSAs only converge late in training, becoming statistically indistinguishable at 115482 steps (0.631 ± 0.028 vs. 0.698 ± 0.020, Welch's t-test, $t(5.92) = 1.6551, p = 0.1497$) (Fig. 3 C2). We further characterized the hippocampal-like properties of the curious pRNN-agent's hidden state representations by measuring spatial information content and Sleep-Wake distance (cosine dissimilarity) at the end of Phase 1 training. We found no significant difference in Sleep-Wake distance between the curious and random agents (0.086 ± 0.034 vs 0.089 ± 0.016; Welch's t-test, $t(5.66) = 0.061, p = 0.953$) (Fig. 3 C4), nor in spatial information (1.155 ± 0.230 vs 0.916 ± 0.002; Welch's t-test, $t(4.07) = 0.927, p =$

0.406) (Fig. 3 C5). Taken together, the absence of significant differences in spatial information, Sleep-wake Distance, and sRSA between the two agents indicates that the curious pRNN-agent develops hippocampal-like hidden state representations to a similar degree as the random agent during Phase 1 training, consistent with the findings of (Levenstein et al., 2024).

We also found that the mutual information between the curious pRNN-agent's position (x-y location and head direction) and its chosen action was higher than the constant mutual information of the random baseline (Fig. 3 C6). This result indicate that the curious agent developed a more state-dependent policy, where specific locations and orientations systematically influence action selection, motivating us to study its behavior under novel environmental conditions.

In Phase 2, a novel green object was introduced at one of three possible target locations (Fig. 4 A). We first validated the curious pRNN-agent's predictions of the green object by inspecting the change in red, green and blue pixel values separately at the target location (Fig. 4B, top). Then, we measured the goal modulation: the change in green pixel value at the target location minus the change in green pixel value at three separate control locations. We found that the curious agent had significantly higher goal modulation than the random agent (Fig. 4B, bottom). This result suggests that the curious agent learned to predict the novel object more robustly than the random baseline.

Secondly, we visualize the occupancy heatmaps of the three pRNN-agents during Phase 2 training: curious, random, and curious control Fig. 5. At the start of training, the curious pRNN-agent has not yet developed a preference for the novel object. However, after 1408 trajectories in Phase 2, its position clusters around the novel object. To verify that this clustering reflected a genuine preference for the novel object rather than an inherent spatial bias of the agent, we tested three separate target locations: $(7, 2), (7, 11)$ and $(14, 7)$ in a 16×16 L-Room. L-Room, and observed

consistent clustering around the novel object across all three. This preference is absent in the random pRNN-agent. To further validate that clustering was driven by the novel object and not simply by continued exploration of the Phase 2 environment, we introduced a curious control agent: an agent that underwent identical Phase 1 training but, during Phase 2, continued training in the standard L-Room without a novel object. This control agent similarly showed no clustering, confirming that the curious pRNN-agent's novelty preference is specifically attributable to the presence of the novel object and its learned policy. Moreover, this positional clustering around the object likely underpins the increased goal modulation.

To quantify this preference, we measured the probability of each agent remaining within a radius-3 region of interest (ROI) around the novel object (Fig. 6A, C, E) as a function of Phase 2 training exposure. The curious pRNN-agent's ROI occupancy probability (Fig. 6 B, D, F, purple) was significantly higher than both baseline agents for the majority of Phase 2 training across all object locations. For example, for the $(7, 11)$ target location, the curious and curious control pRNN-agents both reached a peak ROI probability at step 1008. The curious reached $37.2\% \pm 1.8\%$ (mean $\pm$ SEM), compared to $26.6\% \pm 1.8\%$ for the curious control (Welch's t-test, $t(55.64) = 4.03, p < 0.001$) and $15.0\% \pm 1.2\%$ for the random control (Welch's t-test, $t(47.72) = 10.00, p < 0.001$) . As expected, no such difference was observed early in training: at the random agent's peak ROI step (step 208), the curious agent's ROI probability of $20.2 \pm 0.7\%$ was not significantly different from the random agent's $17.9 \pm 1.4\%$ (Welch's t-test, $t(17.50) = 1.41$, p=ns), consistent with the curious agent not yet having learned to seek out the novel object. Significant separation between the curious agent and both baselines was similarly observed for the $(7, 2)$ and $(14, 7)$ target locations. Notably, this separation emerged more slowly for the $(14, 7)$ location (Fig. 6F), which we attribute to this location being situated in a corner of the L-Room, making it inherently less accessible during early exploration.

Multi-room absolute novelty task. In the multi-room task, we assessed whether the curious pRNN-agent would explore beyond its starting room into increasingly novel, distal rooms (Fig. 7A, B). We computed the mean percentage of time each agent spent in each room over the last 8000 trajectories of training. The random agent spent the vast majority of its time in Room 1 ($86.8\% \pm 0.6\%$), with visitation dropping sharply across rooms ($12.2\% \pm 0.6\%$, $1.0\% \pm 0.1\%$ and $0.02\% \pm 0.02\%$ for Rooms 2, 3, and 4, respectively), indicating that it rarely ventured beyond the starting room. By contrast, the curious pRNN-agent distributed its time more evenly across rooms ($44.2\% \pm 0.4\%, 31.3\% \pm 0.2\%, 16.3\% \pm 0.2\%$ and $8.2\% \pm 0.2\%$ for Rooms 1 through 4). These differences were significant across all four rooms (Welch's t-tests; $p < 0.001$ for Rooms 1 to 3, and $p = 0.0103$ for Room 4) (Fig. 7C). These results suggest that the curious agent can seek absolute novelty as well as contextual novelty.

Discussion

We have shown that the hidden state of a predictive RNN with hippocampal-like properties is sufficient to drive novelty-seeking behavior in reward-free RL agents: (1) in a contextual novelty task, the curious pRNN-agent exhibits a significantly higher spatial preference for novel objects compared to random and curious control baselines, and (2) in an absolute novelty task, the curious pRNN-agent visits increasingly distal, novel rooms more frequently than controls. These results demonstrate that representations that are hippocampal-like and support cognitive map formation can guide exploration in the absence of external reward. Below, we discuss these two findings and compare them to rodent exploratory behavior.

After the curious pRNN-agent familiarized itself with the L-Room environment and began Phase 2 training in the Novel Object Recognition task, we found that the curious pRNN-agent remained consistently near the green object we introduced (Fig. 5, Fig. 6). This result aligns with the hippocampus' known role in contextual novelty detection, particularly when the introduction of

a novel stimulus into a familiar environment elicits a mismatch signal in CA1 (Duncan et al., 2011; Kafkas and Montaldi, 2018). The prediction error used to train the action selection network serves as a computational analog of this mismatch signal: after learning the L-Room environment in Phase 1, the pRNN generates elevated prediction errors when the novel object is in view, driving the agent toward it. Furthermore, the elevated goal modulation (Fig. 4B) shows that the curious agent has spent enough time around the novel object to learn to predict it.

However, we observe that the curious pRNN-agent's preference for the novel object did not diminish over time, as it became more familiar with it (Fig. 6). This observation contrasts with the habituation that can be observed in rodent NOR paradigms, when exploration of a novel object decreases as the animal incorporates it into memory and as the memory trace of the familiar objects degrades (Antunes and Biala, 2011; Cohen and Jr., 2014; Lueptow, 2017). This persistent preference also reflects a limitation of the current experimental design: the novel object remained at a fixed location throughout Phase 2 training. In fact, the curious pRNN-agent's behavior may reflect the learning of a specific memorized location encoded in the pRNN hidden state and linked to high prediction error, rather than the learning of a generalized novelty signal from the hidden state.

This result (Fig. 3, Fig. 6) also raises a deeper question: what constitutes exploratory behavior in a fully familiar environment? Once an agent has learned the statistics of its surroundings, the sources of novelty become vanishingly small (Pathak et al., 2017; Ladosz et al., 2022). Instead, a curiosity-driven agent may exploit its own policy to generate novelty: chasing minute sensory fluctuations or adopting behaviors such as turning in place, since each rotation yields substantially different observations. This pathological behavior is facilitated by the fact that prediction error is computed in raw pixel space, making the intrinsic reward sensitive to low-level perceptual changes rather than actual environmental novelty. In fact, Pathak et al. (2017), Burda et al.

(2019a) and Burda et al. (2019b) propose that errors should be computed in a learned feature space that is invariant to small and ultimately behaviorally irrelevant perceptual changes. The choice of prediction error space is non-trivial, as the hidden state space or the space generated by the convolutional neural network processing the sensory inputs are both valid options.

In the multi-room task, the curious pRNN-agent visited all four rooms, spending progressively more time in rooms further from its starting location (room 1). However, the random agent, which remained confined to Room 1, represents a weak baseline for this comparison. A stronger validation would involve comparing against another intrinsically motivated RL agent, such as one based on Random Network Distillation (Burda et al., 2019b), which would serve as a positive control and allow us to disentangle the contribution of the hippocampal-like representations from other curiosity-driven RL strategies. Nonetheless, the elevated prediction errors generated upon entry into each novel room appear to drive the agent progressively outward into unexplored space. Although we can draw parallels between the prediction error in the NOR task and CA1 mismatch signals (Duncan et al., 2011; Kafkas and Montaldi, 2018), the prediction error in the absolute novelty task could reflect something different: the rapid formation of new place field representations upon entry into a novel room (O'Keefe and Nadel, 1978; Wilson and McNaughton, 1993; Frank et al., 2004), generating high prediction error globally rather than at a specific mismatching location. The way in which the action selection network leverages the pRNN's hidden state may therefore differ fundamentally between contextual and absolute novelty, a distinction that future work could probe directly through targeted analysis of the pRNN's hidden state across both task conditions.

Beyond comparing hidden states across task conditions, future extensions of this work can pair different pRNN architectures with the current RL pipeline. In particular, leveraging the rollout pRNN, which produces theta-sweep activity (Levenstein et al., 2024), would increase the biolog-

ical plausibility of the hippocampal-action-selection model while better aligning with frontier self-supervised RL architectures such as DreamerV3 (Hafner et al., 2023), known for achieving robust exploration-exploitation tradeoffs. A rollout-based pRNN would also allow us to test whether the cognitive map supports prospective simulation of novel trajectories, in the spirit of hippocampal replay (Gorriz et al., 2023). Generating, probing and comparing hidden states from different pRNN-architectures (rollout versus masked) and different tasks (contextual versus absolute novelty) will ultimately allow us to examine *how* the cognitive map is read out to guide exploration. The latter is the central question that this study opens up. By demonstrating that the activity of a hippocampal-like RNN is sufficient to drive exploratory behavior in the absence of external reward, we provide a starting point for asking how the hippocampal cognitive map is sampled to guide behavior and how its structure is shaped by exploratory demands.

References

Antunes M, Biala G (2011). The novel object recognition memory: neurobiology, test procedure, and its modifications. Cognitive Processing 13:93–110. https://doi.org/10.1007/s10339-011-0430-z.

Barker GR, Warburton EC (2011). When Is the Hippocampus Involved in Recognition Memory?. Journal of Neuroscience 31:10721–10731. https://doi.org/10.1523/JNEUROSCI.6413-10.2011.

Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron 100:490–509. https://doi.org/10.1016/j.neuron.2018.10.002.

Broadbent NJ, Gaskin S, Squire LR, Clark RE (2010). Object recognition memory and the rodent hippocampus. Learning & Memory 17:5–11. https://doi.org/10.1101/lm.1650110.

Burda Y, Edwards H, Pathak D, Storkey A, Darrell T, Efros AA (2019a). Large-Scale Study of Curiosity-Driven Learning. In: International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1808.04355.

Burda Y, Edwards H, Storkey A, Klimov O (2019b). Exploration by Random Network Distillation. In: International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1810.12894.

Chevalier-Boisvert M, Dai B, Towers M, Lazcano R de, Willems L, Lahlou S, Pal S, Castro PS, Terry J (2023). Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. In: Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2306.13831.

Cohen JD, Bolstad M, Lee AK (2017). Experience-dependent shaping of hippocampal CA1 intracellular activity in novel and familiar environments. eLife 6:e23040. https://doi.org/10.7554/eLife.23040.

Cohen SJ, Jr. RWS (2014). Assessing Rodent Hippocampal Involvement in the Novel Object Recognition Task: A Review. Behavioural Brain Research 285:105–117. https://doi.org/10.1016/j.bbr.2014.08.002.

Dabaghian Y, Mémoli F, Frank L, Carlsson G (2012). A topological paradigm for hippocampal spatial map formation. PLOS Computational Biology 8:e1002581. https://doi.org/10.1371/journal.pcbi.1002581.

Doerig A, Sommers RP, Seeliger K, others (2023). The neuroconnectionist research programme. Nature Reviews Neuroscience 24:431–450. https://doi.org/10.1038/s41583-023-00705-w.

Duncan K, Ketz N, Inati S, Davachi L (2011). Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. Hippocampus 22:389–398. https://doi.org/10.1002/hipo.20933.

Eichenbaum H (1999). The hippocampus: The shock of the new. Current Biology 9:R482–R484. https://doi.org/10.1016/s0960-9822(99)80301-7.

Fang C, Stachenfeld KL (2024). Predictive auxiliary objectives in deep RL mimic learning in the brain. In: International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2310.06089.

Frank D, Kafkas A (2021). Expectation-driven novelty effects in episodic memory. Neurobiology of Learning and Memory 183:107466. https://doi.org/10.1016/j.nlm.2021.107466.

Frank LM, Stanley GB, Brown EN (2004). Hippocampal plasticity across multiple days of exposure to novel environments. Journal of Neuroscience 24:7681–7689. https://doi.org/10.1523/JNEUROSCI.1958-04.2004.

Geerts JP, Chersi F, Stachenfeld KL, Burgess N (2020). A general model of hippocampal and dorsal striatal learning and decision making. Proceedings of the National Academy of Sciences 117:31427–31437. https://doi.org/10.1073/pnas.2007981117.

Gorriz MH, Takigawa M, Bendor D (2023). The role of experience in prioritizing hippocampal replay. Nature Communications 14:8157. https://doi.org/10.1038/s41467-023-43939-z.

Hafner D, Pašukonis J, Ba J, Lillicrap T (2023). Mastering Diverse Domains through World Models. arXiv. https://doi.org/10.48550/arXiv.2301.04104.

Hill AJ (1978). First occurrence of hippocampal spatial firing in a new environment. Experimental Neurology 62:282–297. https://doi.org/10.1016/0014-4886(78)90058-4.

Jensen KT, Hennequin G, Mattar MG (2024). A recurrent network model of planning explains hippocampal replay and human behavior. Nature Neuroscience 27:1340–1348. https://doi.org/10.1038/s41593-024-01675-7.

Kafkas A, Montaldi D (2018). How do memory systems detect and respond to novelty?. Neuroscience Letters 680:60–68. https://doi.org/10.1016/j.neulet.2018.01.053.

Kanwisher N, Chun MM, McDermott J, Ledden PJ (1996). Functional imaging of human visual recognition. Cognitive Brain Research 5:55–67. https://doi.org/10.1016/s0926-6410(96)00041-9.

Knight R (1996). Contribution of human hippocampal region to novelty detection. Nature 383:256–259. https://doi.org/10.1038/383256a0.

Konda VR, Tsitsiklis JN (1999). Actor-Critic Algorithms. In: Advances in Neural Information Processing Systems, pp 1008–1014. MIT Press.

Ladosz P, Weng L, Kim M, Oh H (2022). Exploration in deep reinforcement learning: A survey. Information Fusion 85:1–22. https://doi.org/10.1016/j.inffus.2022.03.003.

Leutgeb S, Leutgeb JK, Treves A, Moser M-B, Moser EI (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. Science 305:1295–1298. https://doi.org/10.1126/science.1100265.

Levenstein, Efremov, Eyono, Peyrache A, Richards B (2024). Sequential predictive learning is a unifying theory for hippocampal representation and replay. bioarxiv. https://doi.org/https://doi.org/10.1101/2024.04.28.591528 .

Lueptow LM (2017). Novel Object Recognition Test for the Investigation of Learning and Memory in Mice. Journal of Visualized Experiments. https://doi.org/10.3791/55718.

O'Keefe J (2025). How the Hippocampal Cognitive Map Supports Flexible Navigation. Annual Review of Neuroscience 48:331–344. https://doi.org/10.1146/annurev-neuro-112723-023341.

O'Keefe J, Dostrovsky J (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. Brain Research 34:171–175. https://doi.org/10.1016/0006-8993(71)90358-1.

O'Keefe J, Nadel L (1978). The Hippocampus as a Cognitive Map. Clarendon Press.

Pathak D, Agrawal P, Efros AA, Darrell T (2017). Curiosity-driven Exploration by Self-supervised Prediction. In: International Conference on Machine Learning (ICML), pp 2778–2787. https://doi.org/10.5555/3305890.3305968.

Pedamonti D, Mohinta S, Dimitrov MV, Malagon-Vina H, Ciocchi S, Costa RP (2025). Hippocampus supports multi-task reinforcement learning under partial observability. Nature Communications 16:9619. https://doi.org/10.1038/s41467-025-64591-9.

Recanatesi S, others (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nature Communications 12:1417. https://doi.org/10.1038/s41467-021-21696-1.

Richards BA, Lillicrap TP, Beaudoin P, others (2019). A deep learning framework for neuroscience. Nature Neuroscience 22:1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

Samsonovich A, McNaughton BL (1997). Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model. Journal of Neuroscience 17:5900–5920. https://doi.org/10.1523/JNEUROSCI.17-15-05900.1997.

Saxe A, Nelli S, Summerfield C (2021). If deep learning is the answer, what is the question?. Nature Reviews Neuroscience 22:55–67. https://doi.org/10.1038/s41583-020-00395-8.

Schulman J, Moritz P, Levine S, Jordan MI, Abbeel P (2016). High-Dimensional Continuous Control Using Generalized Advantage Estimation. In: International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1506.02438.

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017). Proximal Policy Optimization Algorithms. arXiv preprint arXiv:170706347. https://doi.org/10.48550/arXiv.1707.06347.

Skaggs WE, McNaughton BL, Gothard KM, Markus EJ (1992). An Information-Theoretic Approach to Deciphering the Hippocampal Code. In: Advances in Neural Information Processing Systems (NeurIPS), pp 1030–1037.

Stachenfeld KL, Botvinick MM, Gershman SJ (2014). Design Principles of the Hippocampal Cognitive Map. In: Advances in Neural Information Processing Systems (NeurIPS), pp 2528–2536. https://doi.org/10.5555/2969033.2969109.

Stern CE, Corkin S, González RG, Guimaraes AR, Baker JR, Jennings PJ, Carr CA, Sugiura RM, Vedantham V, Rosen BR (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. Proceedings of the National Academy of Sciences USA 93:8660–8665. https://doi.org/10.1073/pnas.93.16.8660.

Sutton RS, Barto AG (2018). Reinforcement Learning: An Introduction, 2nd ed. MIT Press.

Tolman EC (1948). Cognitive maps in rats and men. Psychological Review 55:189–208. https://doi.org/https://doi.org/10.1037/h0061626.

Tsodyks M (1999). Attractor neural network models of spatial maps in hippocampus. Hippocampus 9:481–489. https://doi.org/10.1002/(SICI)1098-1063(1999)9:4<481::AID-HIPO14>3.0.CO;2-S.

Tulving E, Markowitsch HJ, Kapur S, Habib R, Houle S (1994). Novelty encoding networks in the human brain: positron emission tomography data. NeuroReport 5:2525–2528. https://doi.org/10.1097/00001756-199412000-00030.

Viejo G, Levenstein D, Carrasco SS, Mehrotra D, Mahallati S, Vite GR, Denny H, Sjulson L, Battaglia FP, Peyrache A (2023). Pynapple, a toolbox for data analysis in neuroscience. eLife. https://doi.org/10.7554/eLife.85786.2.

Wilson MA, McNaughton BL (1993). Dynamics of the hippocampal ensemble code for space. Science 261:1055–1058. https://doi.org/10.1126/science.8351520.

Zhang X, Ma H (2018). Pretraining Deep Actor-Critic Reinforcement Learning Algorithms With Expert Demonstrations. arXiv preprint arXiv:180110459. https://doi.org/10.48550/arXiv.1801.10459.

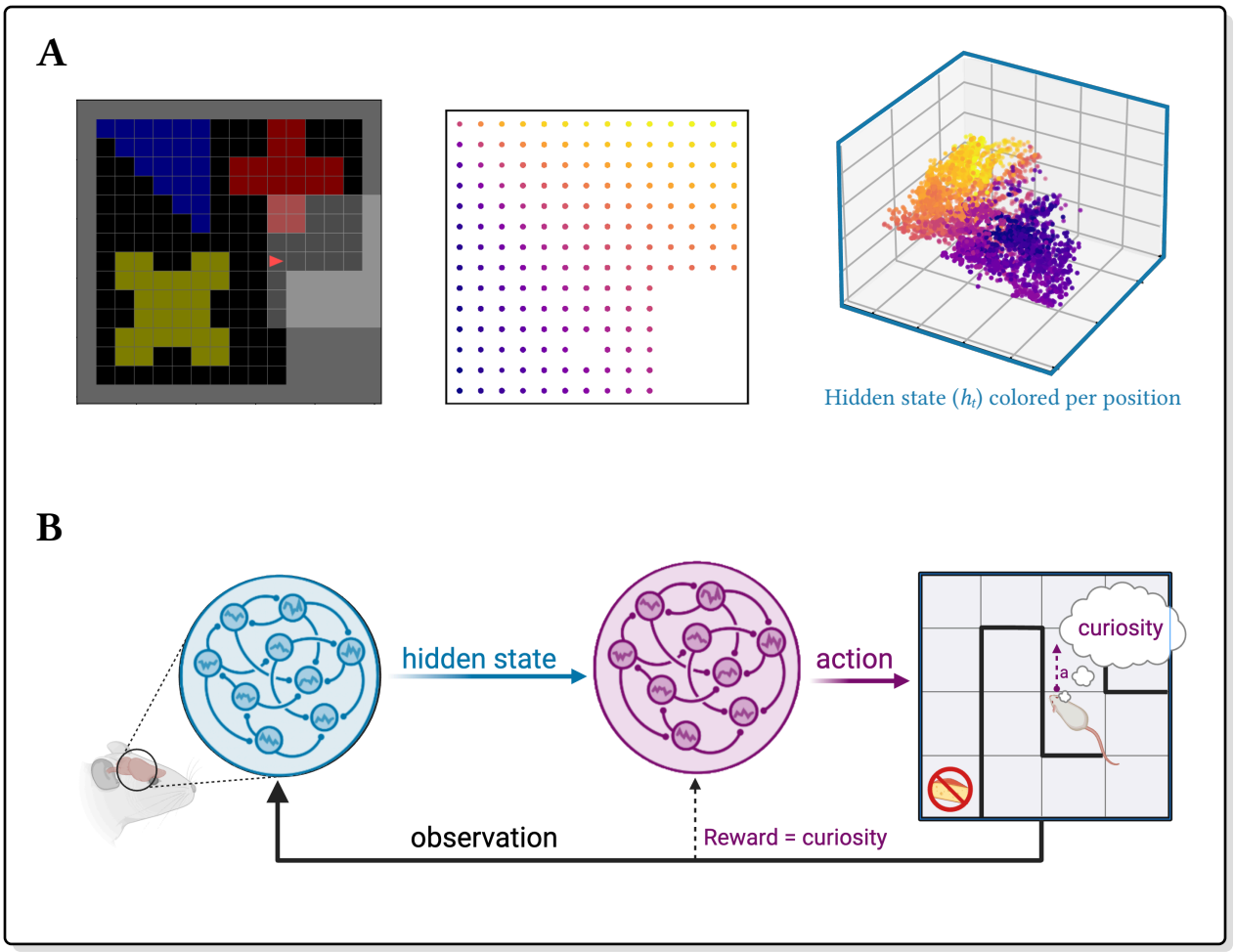

Figure 1: A predictive recurrent neural network (pRNN) forms a cognitive map that can be used for curiosity-driven RL. (A, left). Example environment that the Reinforcement Learning (RL) pRNN-agent must navigate. (A, middle) Possible positions that it can occupy in the environment. At each position, the pRNN-agent can have four possible head directions (N, W, S, E). (A, right) Isomap visualization of the population activity of the pRNN (i.e., 500-dimensional hidden state vectors of the pRNN), projected along first three Isomap components and colored per position. 5000 hidden states sampled from 688 trajectories of 256 steps each were used for the Isomap visualisation. (B) Overview of the curiosity-driven RL framework. The pRNN (blue) receives visual observations from the environment and maintains a hidden state h_t that encodes a cognitive map of the agent's surroundings. This hidden state is passed to an Actor-Critic network (purple), which selects actions. The pRNN's prediction error serves as an intrinsic curiosity reward in the absence of any external reward signal, driving the agent to seek out novelty.

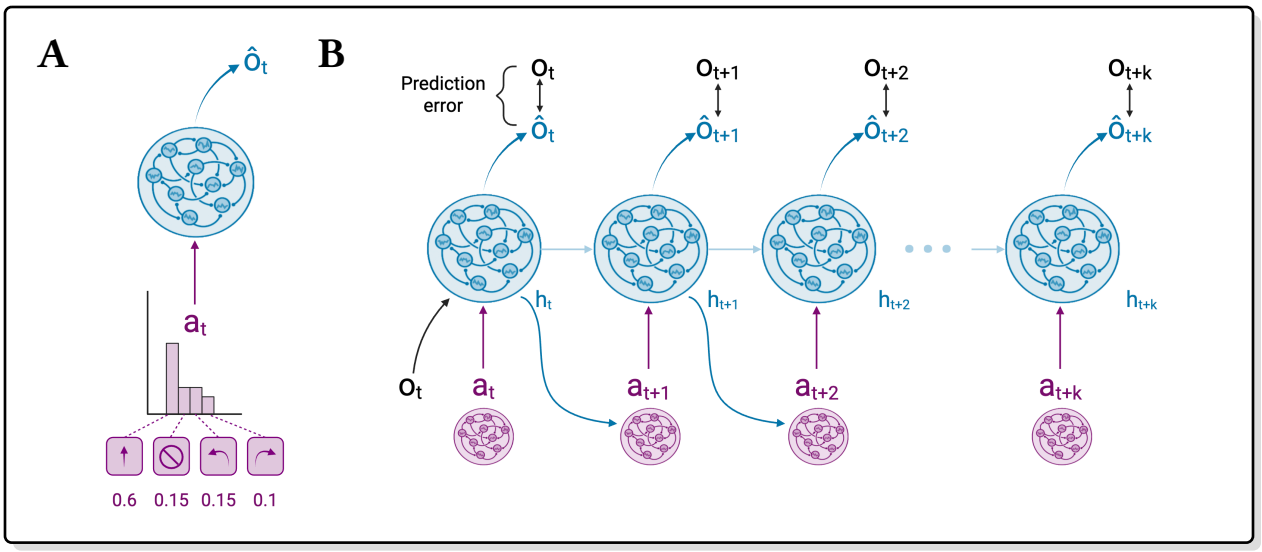

Figure 2: Architecture diagram for the curious pRNN-agent. (A) The recurrent neural network (RNN) (blue) is an architecture taken from (Levenstein et al., 2024), where the actions are sampled from a fixed distribution (purple). We call this pair the random pRNN-agent. (B) The RNN is a k-masked network trained to predict the observation at the current time step given an action. This predictive RNN (pRNN) only receives the true observation every k steps. It must therefore predict k-1 observations with only an action as input. The agent, or action network (purple), consists of two multi-layer perceptrons forming an Actor-Critic network (Konda and Tsitsiklis, 1999). The action network is trained to maximize a reward, formulated as the pRNN's prediction error. We call this pair the curious pRNN-agent.

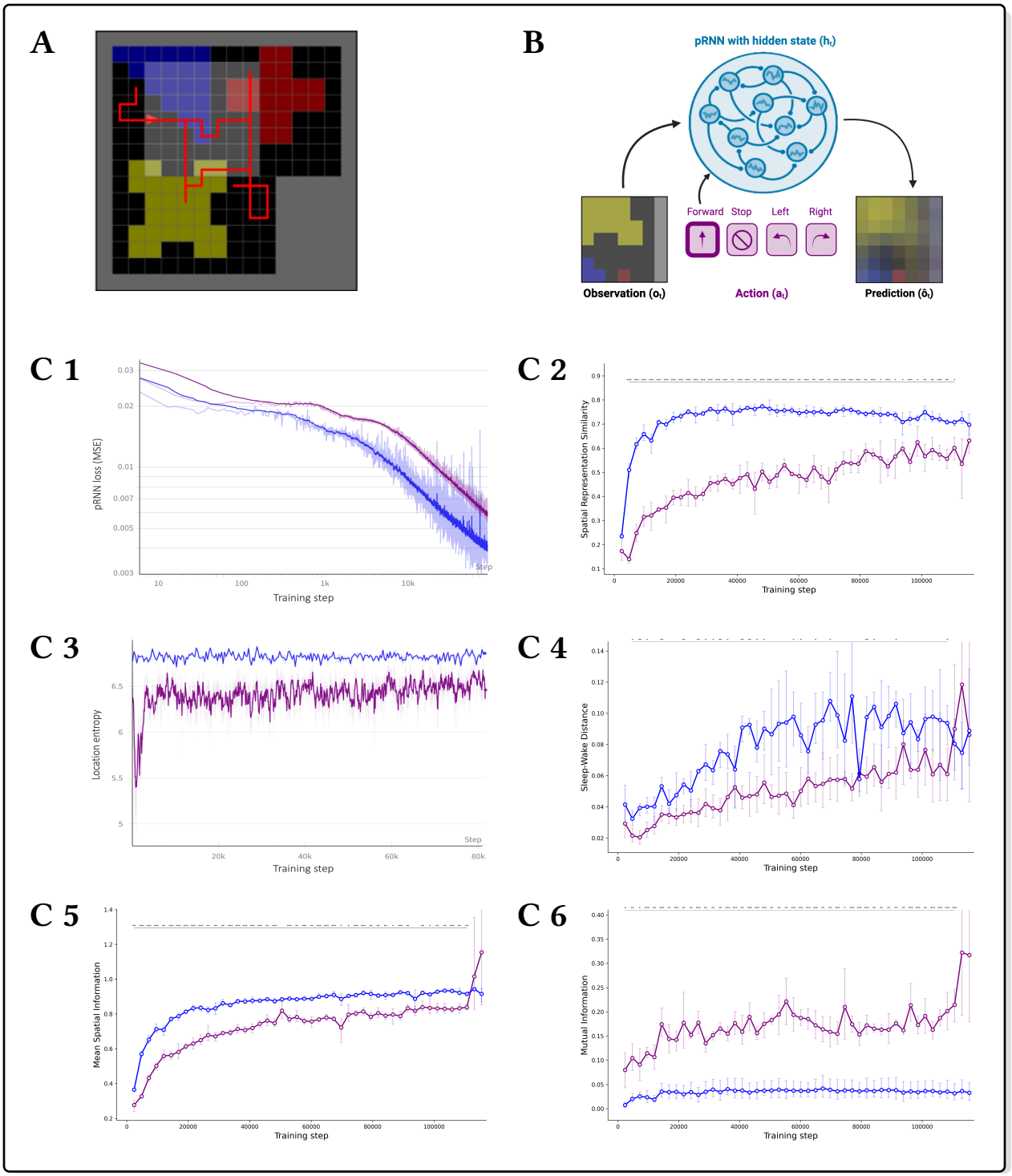

Figure 3: Validation of curious pRNN-agent's performance and hippocampal-like representations in the L-Room Phase 1 training. (A) Example 256-step trajectory taken by the curious pRNN-agent. (B) Illustration of the observation-action-prediction loop, adapted from (Levenstein et al., 2024). (C) Blue: Actions fed into the pRNN are sampled from a fixed distribution (random agent). Purple: Actions are computed by the Actor-Critic action network. Metrics were introduced in (Levenstein et al., 2024).

(C1) pRNN loss curves. (C2) Spatial representational similarity analysis: Spearman rank correlation between pairwise distances in pRNN hidden state space and corresponding pairwise Euclidean distances between positions in the environment. (C3) Location entropy: Shannon entropy of the agent's visited position distribution, reflecting the uniformity of spatial coverage across the environment. (C4) Cosine distance in pRNN hidden state space between sleep and wake manifolds. See (Levenstein et al., 2024) and Methods for futher details. (C5) Mean spatial information: spatial information score across pRNN hidden units, quantifying how much individual unit activations predict the agent's position in the environment (Skaggs et al., 1992). (C6) Mutual information between the pRNN-agent state (position and orientation) and the selected action.

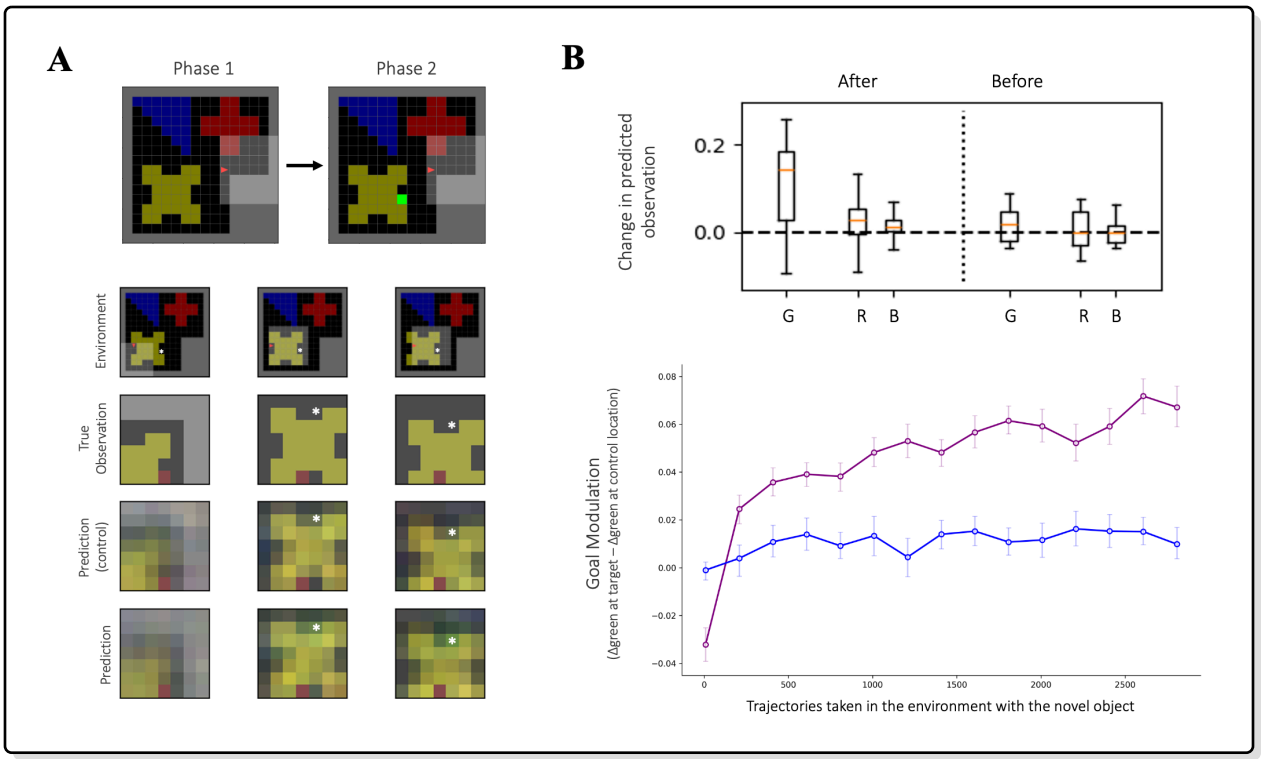

Figure 4: Novel Object Recognition task setup and curious pRNN-agent performance. (A) The curious pRNN-agent was trained to convergence in a Farama-Minigrid L-Room before a novel green object was introduced at one of three possible locations. Here, we visualize position $(7, 11)$. Rows show the environment, true observation, control pRNN prediction, and curious pRNN prediction at three viewpoints near the target location (white star). Predictions are collected in the original environment without the novel object, so that the predicted green pixel value reflects memory of the object rather than its current sensory input. (B, top) Change in predicted green (G), red (R), and blue (B) pixel values at position $(7, 11)$ before and after novel object exposure (after 1808 trajectories in Phase 2), showing a selective increase in predicted green value after exposure. (B, bottom) Goal modulation (change in predicted green pixel at target minus control locations) is significantly higher in the curious agent (purple) than the random baseline (blue) across all Phase 2 training time points shown. (Welch's t-test, t-sided, $p < 0.001$). Evaluation procedure follows (Levenstein et al., 2024) Figure S21.

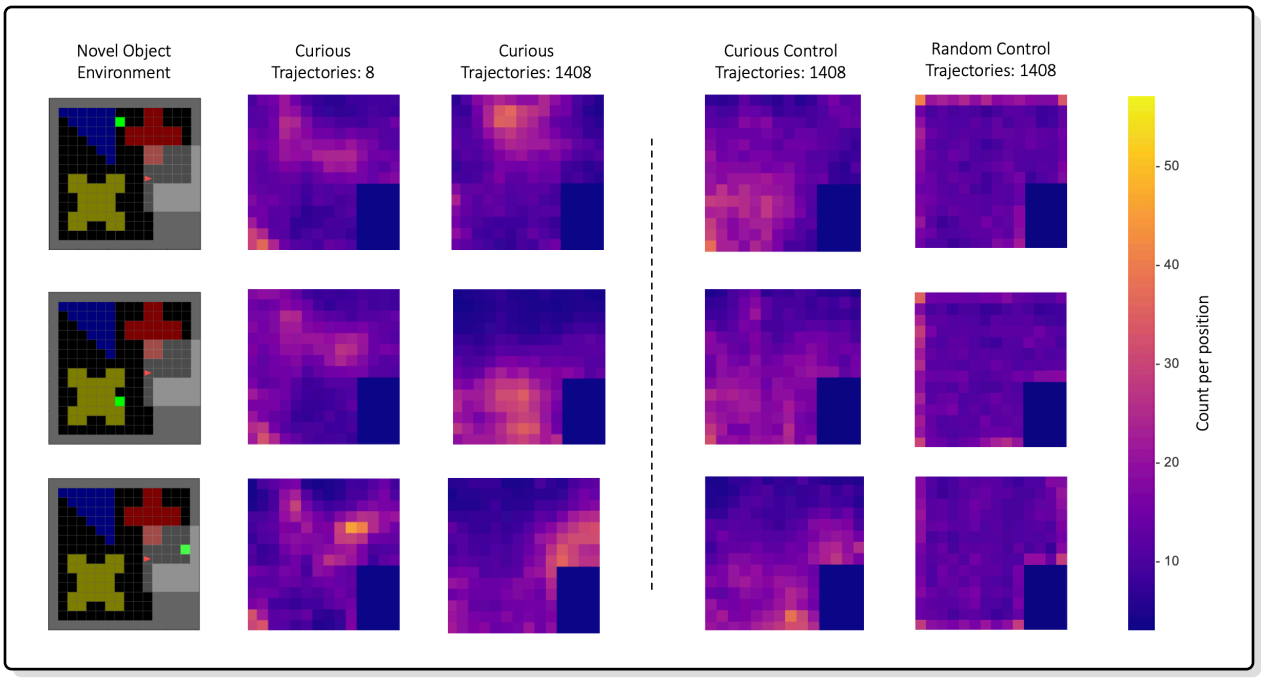

Figure 5: pRNN-agent occupancy heatmaps for different novel object locations in the NOR task. (First column) Three novel object environments used to evaluate curious pRNN-agent behavior. Novel object locations are $(7, 11)$ (top), $(7, 2)$ (middle) and $(14, 7)$ (bottom). pRNN-agents were trained separately in each novel object environment but the same initial pRNN-agent checkpoints were used from the novel-object-absent L-Room pretraining. (Second column) Curious pRNN-agent occupancy heatmap after training on eight 256-step trajectories in novel object environment. Occupancy heatmaps are all average over all four possible head directions per position. (Third column) Curious pRNN-agent occupancy heatmaps after training on 1408 trajectories in the novel object L-Room. (Fourth column) Using the same initial pRNN-agent checkpoints, we continued training in the novel-object-absent L-Room to visualize "typical" curious pRNN-agent occupancy, which does not cluster near novel object locations yet is not completely uniform over the environment either. (Fifth column) Occupancy heatmap under random action selection.

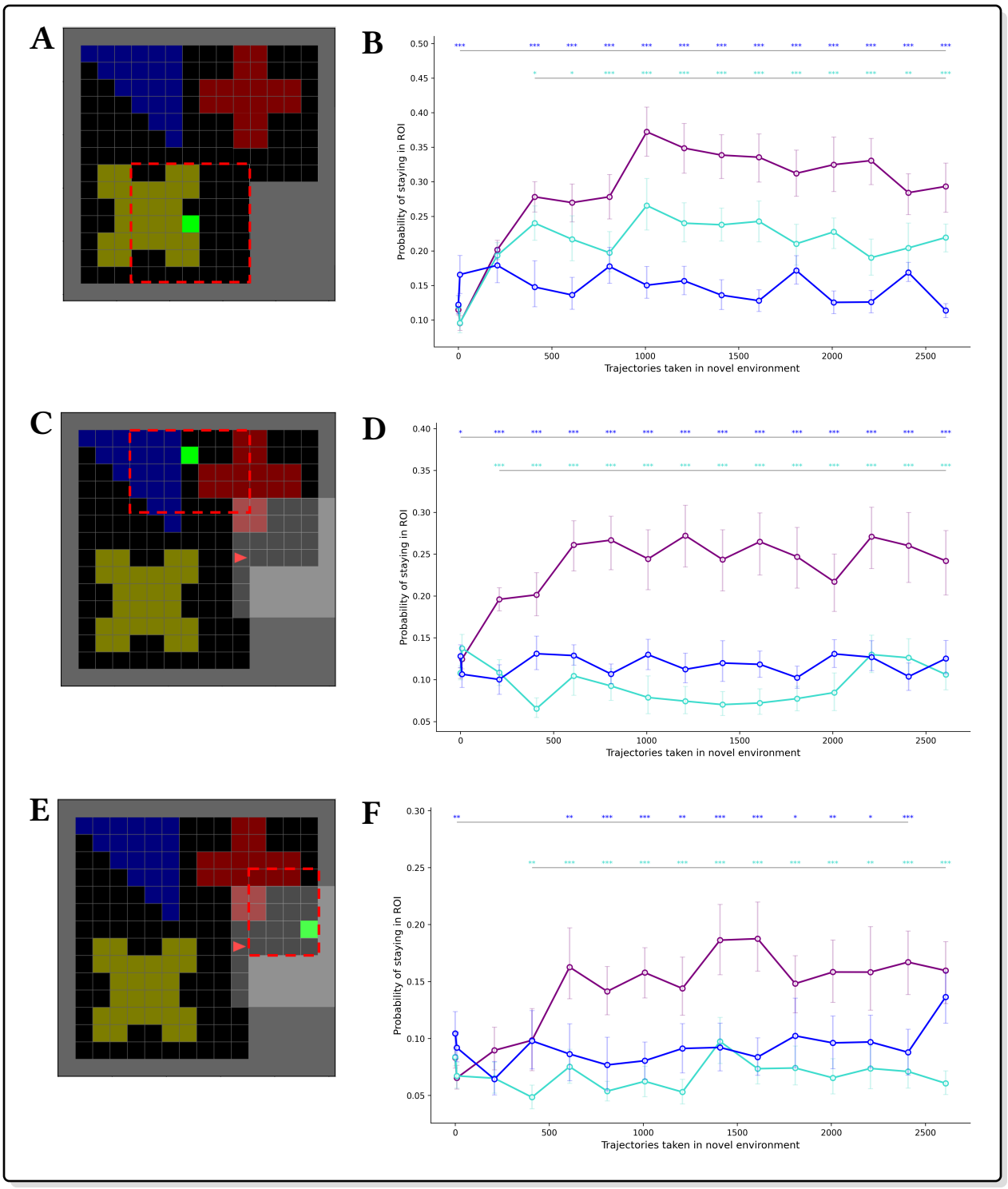

Figure 6: Probability of remaining near (radius of 3) the novel object during the NOR task. (Left column) Visualization of the region of interest near the novel object (red). (Right column) Probability of remaining inside the region of interest for the curious pRNN-agent (purple), random pRNN-agent (navy blue) and curious control pRNN-agent (turquoise) as the number of training trajectories in a given novel object environment increases. Significance was measured with Welch's two-sided t-test.

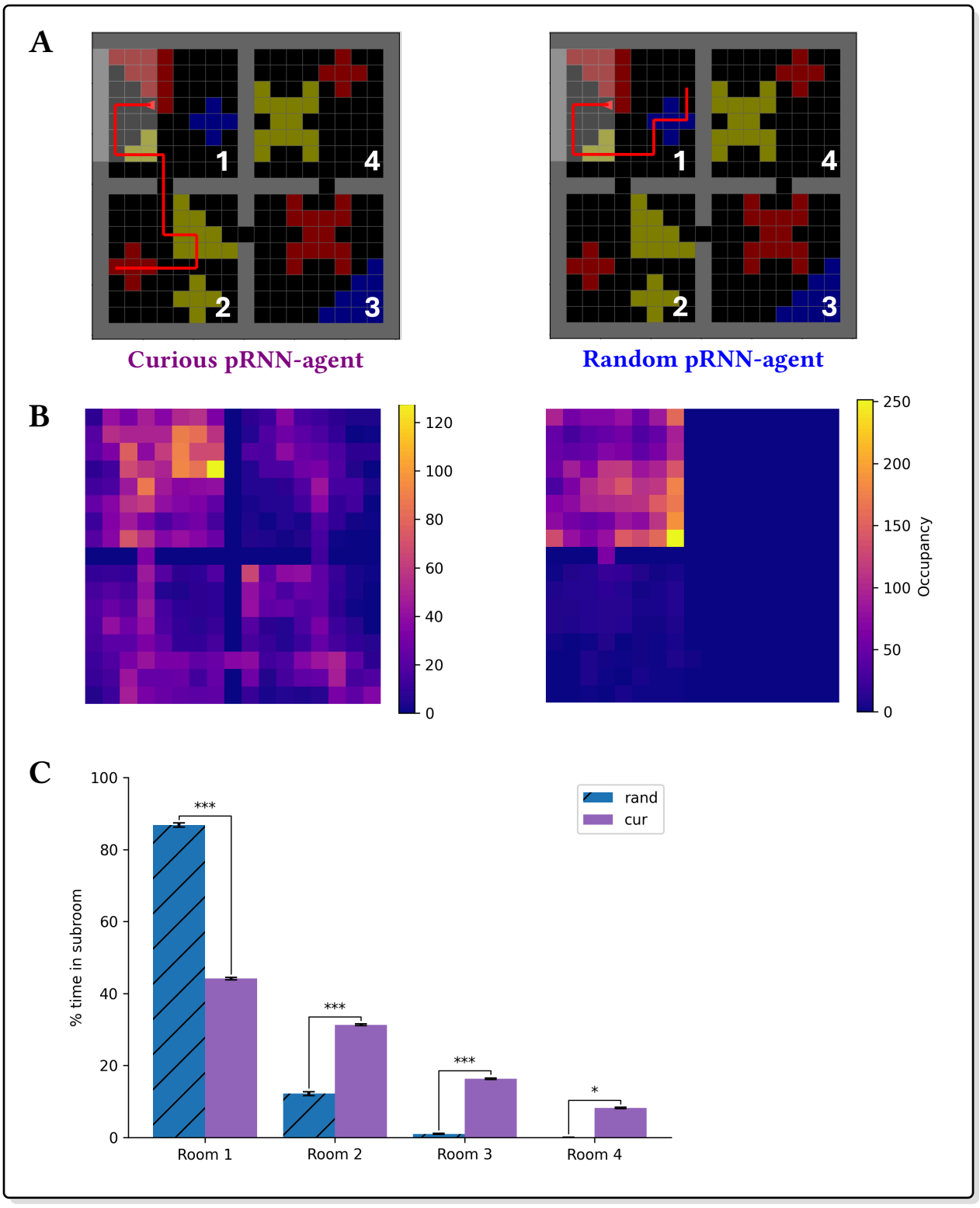

Figure 7: Multi-room task and sub-room visitation frequency. (A) Curious and random pRNN-agent example trajectories in the multi-room environment. (B) Occupancy heatmaps at the end of training of pRNN-agents. (C) Percentage of time spent in each sub-room for the last 8000 training trajectories, as that is when the curious pRNN-agent behavior has stabilized.

Hyperparameter	Value	Description
Learning rate	10^{-3}	Adam optimizer
Batch size	64	Training batch
Hidden size	128	RNN hidden dim
Discount (γ)	0.99	RL discount factor
RL (PPO)		
Total training steps	2.048×10^7	Environment frames
Trajectories per batch	8	Sequences per gradient update
Trajectory length	256	Steps in environment per episode
Frames per update	2048	trajs $\times$ seq. duration
Discount (γ)	0.98	RL discount factor
Learning rate	3×10^{-4}	Adam optimizer
GAE λ	0.95	GAE variance reduction factor
Entropy coefficient	0.0	Policy entropy bonus
PPO epochs	4	Gradient steps per batch
PPO clip ε	0.2	Policy ratio clipping
Max gradient norm	0.5	Gradient clipping
pRNN: Phase 1 of NOR and multi-room tasks		
Hidden size	500	RNN hidden dim
Learning rate	3×10^{-3}	Adam optimizer
Weight decay	3×10^{-3}	L2 regularization
Dropout	0.15	
Hidden state noise σ	0.05	Additive Gaussian noise
Phase 2 of NOR		
Training trajectories	3000	Number of trajectories in environment.
Learning rate multiplier	2	Learning rate scale during Phase 2

Table 1: Training hyperparameters Unless stated otherwise, these are the standard hyperparameters used in all experiments.

Variable	Description
$h_t \in \mathbb{R}^n$	Hidden state of the pRNN, considered as the "activity" of the pRNN.
$b \in \mathbb{R}^n$	Neuron-specific bias in the pRNN
$W_{\text{rec}} \in \mathbb{R}^{N \times N}$	Recurrent weight matrix
$W_{\text{in}} \in \mathbb{R}^{N \times N_{\text{obs}}}$	Input weight matrix for the typically 7×7 observation
$W_{\text{act}} \in \mathbb{R}^{N \times N_{\text{act}}}$	Input weight matrix for the action
$W_{\text{out}} \in \mathbb{R}^{N_{\text{obs}} \times N}$	Output weight matrix of the pRNN
η	Standard deviation of uncorrelated noise injection to the pRNN

Table 2: pRNN architecture variables. Variable definitions taken from (Levenstein et al., 2024)

Variable	Description
$h_t \in \mathbb{R}^n$	Hidden state of the pRNN, considered as the "activity" of the pRNN.
$b \in \mathbb{R}^n$	Neuron-specific bias
$W_{\pi 1} \in \mathbb{R}^{N \times 64}$	Input weight matrix for the actor network
$W_{V1} \in \mathbb{R}^{N \times 64}$	Input weight matrix for the critic network
$W_{\pi 2} \in \mathbb{R}^{N \times N_{\text{action}}}$	Output weight matrix for the actor network
$W_{V2} \in \mathbb{R}^{N \times 1}$	Output weight matrix for the critic network (maps to dim = 1 because the critic network outputs a scalar value for a state)

Table 3: Actor-Critic network architecture variables.