frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 09 February 2023
DOI 10.3389/fncom.2023.1053097

OPEN ACCESS

EDITED BY
Si Wu,
Peking University, China

REVIEWED BY
Chung-Chuan Lo,
National Tsing Hua University, Taiwan
Fabian Kloosterman,
Neuroelectronics Research Flanders, Belgium

*CORRESPONDENCE
Yuanxiang Gao
✉ gaoyuanxiang@itp.ac.cn

RECEIVED 25 September 2022
ACCEPTED 16 January 2023
PUBLISHED 09 February 2023

CITATION
Gao Y (2023) A computational model of learning flexible navigation in a maze by layout-conforming replay of place cells. *Front. Comput. Neurosci.* 17:1053097. doi: 10.3389/fncom.2023.1053097

A computational model of learning flexible navigation in a maze by layout-conforming replay of place cells

Yuanxiang Gao[1,2]*

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2]CAS Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China

Recent experimental observations have shown that the reactivation of hippocampal place cells (PC) during sleep or wakeful immobility depicts trajectories that can go around barriers and can flexibly adapt to a changing maze layout. However, existing computational models of replay fall short of generating such layout-conforming replay, restricting their usage to simple environments, like linear tracks or open fields. In this paper, we propose a computational model that generates layout-conforming replay and explains how such replay drives the learning of flexible navigation in a maze. First, we propose a Hebbian-like rule to learn the inter-PC synaptic strength during exploration. Then we use a continuous attractor network (CAN) with feedback inhibition to model the interaction among place cells and hippocampal interneurons. The activity bump of place cells drifts along paths in the maze, which models layout-conforming replay. During replay in sleep, the synaptic strengths from place cells to striatal medium spiny neurons (MSN) are learned by a novel dopamine-modulated three-factor rule to store place-reward associations. During goal-directed navigation, the CAN periodically generates replay trajectories from the animal's location for path planning, and the trajectory leading to a maximal MSN activity is followed by the animal. We have implemented our model into a high-fidelity virtual rat in the MuJoCo physics simulator. Extensive experiments have demonstrated that its superior flexibility during navigation in a maze is due to a continuous re-learning of inter-PC and PC-MSN synaptic strength.

KEYWORDS

flexible navigation, place cells, hippocampal replay, medium spiny neurons, three-factor learning, path planning

1. Introduction

It has been observed that, during sleep or wakeful immobility, hippocampal place cells (PC) spontaneously and sequentially fire, similar to the pattern of activity during movement periods (Lee and Wilson, 2002; Foster and Wilson, 2006; Pfeiffer and Foster, 2013; Stella et al., 2019). Conventionally, studies of such hippocampal replay were restricted to animals moving in simple environments like linear tracks or open fields. Recently, an interesting study (Widloski and Foster, 2022) observed that, for a rat moving in a reconfigurable maze, the replay trajectories conform to the spatial layout and can flexibly adapt to a new layout. Such layout-conforming replay is the key to understanding how the activity of place cells supports the learning of flexible navigation in a maze, a long-standing open question in computational neuroscience.

Existing models (Hopfield, 2009; Itskov et al., 2011; Azizi et al., 2013; Romani and Tsodyks, 2015) of hippocampal replay use a continuous attractor network (CAN) with spike-frequency adaptation or short-term synaptic depression (Romani and Tsodyks, 2015) to generate replay of place cells. In these models, the synaptic strength between a pair of place cells is pre-configured by a decaying function of the Euclidean distance between their place field centers. In a maze,

such Euclidean synaptic strength introduces a strong synaptic coupling between a pair of place cells with firing field centers located at two sides of a thin wall. Accordingly, the activity bump of place cells will pass through the wall without conforming to the layout. Besides, these models lack the modeling of how the replay activity drives downstream circuits, such as striatum, to perform high-level functional roles, such as reward-based learning, planning and navigation. Although reinforcement learning (RL) algorithms (Sutton and Barto, 1998) can serve as candidate models of these functional roles (Brown and Sharp, 1995; Foster et al., 2000; Johnson and Redish, 2005; Gustafson and Daw, 2011; Russek et al., 2017), these algorithms are designed to solve much more general and abstract control problems in engineering domains, falling short of biological plausibility.

In this paper, we propose a computational model for generating layout-conforming replay and explaining how such replay supports the learning of flexible navigation in a maze. First, we model the firing field of a place cell by a function decaying with the shortest path distance from its field center, which conforms to experimental observations (Skaggs and McNaughton, 1998; Gustafson and Daw, 2011; Widloski and Foster, 2022). Then we propose a Hebbian-like rule to learn the inter-PC synaptic strength during exploration. We find that the synaptic strength between a pair of place cells encodes the spatial correlation of their place fields and decays with the shortest path distance between their field centers. With learned inter-PC synaptic strength, the interactions among place cells are modeled by a continuous attractor network with feedback inhibition from hippocampal GABAergic interneurons (Schlingloff et al., 2014; Stark et al., 2014). The drift of the activity bump of place cells under vanishing or non-zero visual inputs models layout-conforming replay during rest and goal-directed navigation, respectively.

During replay in rest, the synaptic strength from place cells to medium spiny neurons (MSN) in striatum is learned by a novel three-factor learning rule based on a replacing trace rule (Singh and Sutton, 1996; Seijen and Sutton, 2014), rather than the conventional accumulating trace rule (Izhikevich, 2007; Gerstner et al., 2018; Sutton and Barto, 2018). This learning rule strengthens a PC-MSN synapse proportional to the co-firing trace of this pair of PC and MSN multiplied by the dopamine release at the synapse (Yagishita et al., 2014; Kasai et al., 2021). After replay, the PC-MSN synaptic strength encodes the geodesic proximity between the firing field center of each place cell and the goal location. As a result, the activity of the MSN population will ramp up when an animal approaches the goal location, which conforms to a line of experimental observations (van der Meer and Redish, 2011; Howe et al., 2013; Atallah et al., 2014; London et al., 2018; Sjulson et al., 2018; O'Neal et al., 2022). During goal-directed navigation, the attractor network periodically generates a series of replay trajectories from the rat's location to lookahead along each explorable path (Johnson and Redish, 2007; Pfeiffer and Foster, 2013), and the trajectory leading to a maximal MSN activity is followed by the rat with a maximal probability.

We have implemented our model into a high-fidelity virtual rat that reproduces the average anatomical features of seven Long-Evans rats in the MuJoCo physics simulator (Todorov et al., 2012; Merel et al., 2020). Extensive experiments have demonstrated that the virtual rat shows behavioral flexibility during navigation in a dynamically changing maze. We have observed that, after the layout changes, the inter-PC synaptic strength is updated to re-encode the adjacency relation between locations in the new layout. As a result, the replay trajectories adapt to the new layout so that after replay the MSN activity ramps up along new paths to the goal location. Periodical planning with respect to the updated MSN activity explains the navigational flexibility of the virtual rat.

2. Model

2.1. Learning inter-PC synaptic strength

For a population of M place cells, the firing rate of the i-th place cell is denoted by r_i. Each place cell has a preferential location, denoted by $\mathbf{x}^{(i)}$, where its firing rate is maximal among all locations. The preferential location of each place cell is uniformly arranged on a regular grid of a square maze without considering grid points that fall into areas under walls. Let J_{ij} denote the strength of the synapse from cell j to cell i. Before learning J_{ij}, the firing rate of each place cell is modeled by a function of the rat's location $\mathbf{x}$ given by,

$$r_i(\mathbf{x}) = \exp(-\mathrm{D}(\mathbf{x}^{(i)}, \mathbf{x})/\sigma), \forall i, \tag{1}$$

Where σ is a scale parameter and $\mathrm{D}(\mathbf{x}^{(i)}, \mathbf{x})$ is the shortest path distance between $\mathbf{x}^{(i)}$ and $\mathbf{x}$. The firing field in Equation (1) is a geodesic place field (Gustafson and Daw, 2011), which conforms to the constraints imposed by walls and is a more accurate model of firing fields observed in a maze (Skaggs and McNaughton, 1998; Gustafson and Daw, 2011; Widloski and Foster, 2022) compared to Gaussian place fields observed in open fields (O'Keefe and Burgess, 1996; Foster et al., 2000).

For a rat randomly exploring a maze with many walls, a pair of place cells with a small shortest path distance between their preferential locations often fire in close temporal order. According to symmetric spike-timing dependent plasticity (STDP) (Isaac et al., 2009; Mishra et al., 2016), the pair of recurrent synapses between this pair of place cells becomes stronger than pairs of place cells with larger shortest path distance between their preferential locations. Accordingly, for a sequence of positions $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_N\}$ visited by a virtual rat during random exploration, the synaptic strength matrix is symmetrically updated by the following Hebbian-like rule,

$$\mathbf{J}^{(n+1)} = \mathbf{J}^{(n)} + \alpha_1(\mathbf{r}(\mathbf{x}_n)^{\mathrm{T}}\mathbf{r}(\mathbf{x}_n) - \mathbf{J}^{(n)}), \tag{2}$$

Where $\mathbf{r}(\mathbf{x}_t) = \{r_i(\mathbf{x}_t), i = 1, ..., M\}$ is the row vector of place cell firing rates at position $\mathbf{x}_t$, α_1 is the learning rate, n is the number of locomotion period during exploration trials, and $\mathbf{J}^{(0)}$ is initialized as a zero matrix. The conventional Hebbian learning rule (Muller et al., 1996; Haykin, 1998; Stringer et al., 2002) increases inter-PC synaptic strength proportional to the product of firing rates of a pair of place cells hence inter-PC synaptic strength increases without bound and fails to converge. The proposed Hebbian-like rule is free from such convergence issues.

The learning rule in Equation (2) is a stochastic approximation process (Nevelson and Hasminskii, 1976) that solves the following equation as $n \to \infty$,

$$J^*_{ij} = \mathbb{E}[r_i(\mathbf{x})r_j(\mathbf{x})], \forall i, j, \tag{3}$$

Where the expectation is taken over possible locations during exploration. As shown by Equation (3), the synaptic strength between a pair of place cells encodes the correlation of their firing fields, which characterizes the geodesic proximity of their field centers.

2.2. Hippocampal replay by a continuous attractor network

After learning the synaptic strength matrix, the activity of place cells can change independently with respect to the rat's current location $\mathbf{x}$, due to synaptic interactions among place cells. Thus, r_i is also a function of time, denoted by $r_i(\mathbf{x}, t)$. Since $\mathbf{x}$ is fixed during replay, $r_i(\mathbf{x}, t)$ is written as $r_i(t)$ for simplicity. $r_i(t)$ obeys following differential equations,

$$\begin{cases} \tau_r \dot{r}_i = -r_i + \left[\sum_{j \neq i} J_{ij} r_j + E_i - I_i - h_0\right]^+, \\ \tau_I \dot{I}_i = -I_i + c_I r_i, \end{cases} \tag{4}$$

Where τ_r and τ_I are time constants, h_0 is a threshold, E_i is the external input to cell i, I_i is the feedback inhibition due to the reciprocal connection between cell i and hippocampal GABAergic interneurons (Pelkey et al., 2017), c_I is the strength of feedback inhibition, and $[x]^+ = \max\{x, 0\}$ is the neural transfer function given by a threshold-linear function. The feedback inhibition in Equation (4) is supported by observations that hippocampal sharp wave ripples (SWRs) containing replay events are generated by feedback inhibition from parvalbumin-containing (PV+) basket neurons (Schlingloff et al., 2014; Stark et al., 2014; Buzsaki, 2015).

In Equation (4), the external input E_i characterizes the association of place cell i with visual cues at the rat's location. E_i is typically a decaying function of the distance between the rat's location and the firing field center of place cell i (Stringer et al., 2002, 2004; Hopfield, 2009). Accordingly, E_i is given by,

$$E_i(\mathbf{x}) = \mathcal{A} \cdot \exp(-\mathrm{D}(\mathbf{x}^{(i)}, \mathbf{x})/\sigma), \forall i, \tag{5}$$

Where $\mathcal{A}$ is the amplitude of the external input. Under Euclidean synaptic strength without feedback inhibition (e.g., $c_I = 0, I_i \equiv 0$), the steady state solution of $\mathbf{r}$, mapped to each preferential location, is a localized Gaussian-like bump centered at the rat's location in an open field (Wu et al., 2008; Fung et al., 2010), due to locally strong excitatory coupling among place cells with firing fields nearby the rat's location. The learned synaptic strength in Equation (4) has similar locally strong coupling so that its steady state solution without feedback inhibition is also a localized Gaussian-like bump centered at the rat's location in the maze.

With both external input and feedback inhibition, the activity bump may move away from the rat's location and its mobility modes are determined by both factors in a competitive way. The external input tends to stabilize the bump at $\mathbf{x}$ and the feedback inhibition tends to deviate the bump from its current location. Under a fixed strength of feedback inhibition, if the external input is strong enough, the feedback inhibition fails to deviate the stable activity bump. In contrast, if the external input vanishes, the activity bump freely drifts along paths in the maze, resembling replay during sleep or rest without sensory drive (Stella et al., 2019). Most interestingly, for an external input with intermediate amplitude, the drift of the activity bump periodically jumps back to the rat's location to regenerate a drift trajectory along other paths, resembling replay during goal-directed navigation (Johnson and Redish, 2007; Pfeiffer and Foster, 2013; Widloski and Foster, 2022).

2.3. Learning PC-MSN synaptic strength

During replay in sleep or brief rest, the synaptic strengths from place cells to a population of medium spiny neurons (MSN) in the striatum are selectively strengthened to store place-reward associations (Lansink et al., 2009; Sjulson et al., 2018; Trouche et al., 2019; Sosa et al., 2020). The learning of PC-MSN synaptic strength is achieved by dopamine-modulated synaptic plasticity of PC-MSN synapses (Floresco et al., 2001; Jay, 2003; Brzosko et al., 2019). Precisely, the co-firing of a place cell and MSN leaves a chemical trace (i.e., Ca^{2+} influx) at the synapse (connection) from the place cell to the population of MSN (Yagishita et al., 2014; Kasai et al., 2021). After co-firing, the Ca^{2+} is gradually taken up by organelles hence the chemical trace decays exponentially at the synapse (Abrams and Kandel, 1988; Yagishita et al., 2014). Before the Ca^{2+} is fully absorbed, if the dopamine concentration increases/decreases from a base level at the synapse, the synapse will be strengthened/depressed proportionally to both the chemical trace and the increase/decrease of the dopamine concentration (Yagishita et al., 2014; Gerstner et al., 2018; Kasai et al., 2021). These experimental observations are modeled as follows.

Let $W_i(t)$ denote the synaptic strength from place cell i to the MSN population at time t. The activity of the MSN population at time t is modeled by Sjulson et al. (2018),

$$V(t) = \sum_i W_i(t) r_i(t). \tag{6}$$

One recent study (Gauthier and Tank, 2018) has found a population of goal cells in hippocampus with their place field centers always tracking the goal location $\mathbf{x}_g$. Let $G(t)$ denote the activity of the population of goal cells, and U_i denote the synaptic strength from place cell i to the goal cell population. Accordingly, the activity of the goal cell population is modeled by,

$$G(t) = \sum_i U_i r_i(t), \tag{7}$$

Where $U_i = \exp(-\mathrm{D}(\mathbf{x}^{(i)}, \mathbf{x}_g)/\xi)$, characterizing the stronger connection between place cells with firing fields nearby the goal location and the goal cell population. Let $z_i(t)$ denote the chemical trace at the synapse from place cell i to the MSN population at time t. Let $\delta(t)$ denote the deviation of dopamine concentration upon a base level at PC-MSN synapses at time t.

During replay in sleep or rest, the PC-MSN synaptic strengths are updated by the following three-factor rule,

$$\begin{cases} \dot{W}_i(t) = \alpha_2 z_i(t) \delta(t), \\ z_i(t) = r_i(t) V(t) I_{[r_i(t)V(t) > q]}, \\ \dot{z}_i(t) = -\frac{z_i(t)}{\tau_z} I_{[r_i(t)V(t) \leq q]}, \\ \delta(t) = G(t) + \dot{V}(t), \end{cases} \tag{8}$$

Where α_2 is the learning rate, $W_i(0)$ is initialized to be zero, q is a small threshold for judging whether PC i and MSN are co-firing,

and $I_{[\cdot]}$ is the indicator function that equals 1 if the condition holds and 0 otherwise. τ_z is a time constant. In Equation (8), the dynamic of $z_i(t)$ characterizes that the concentration of Ca^{2+} at a PC-MSN synapse encodes the instantaneous joint firing rate of PC i and MSN when PC i and MSN are co-firing or decays exponentially otherwise (Helmchen et al., 1996; Wang, 1998). Computationally, this novel trace update rule is a continuous-time generalization of the replacing trace rule (Singh and Sutton, 1996; Seijen and Sutton, 2014) (with $q = 0$ without postsynaptic factors), which has better learning efficiency than the conventional accumulating trace rule (Izhikevich, 2007; Gerstner et al., 2018; Sutton and Barto, 2018). In Equation (8), $\delta(t)$ characterizes the activity of VTA dopamine neurons, which is a summation of the signal on a disinhibitory pathway from the goal cell population (Luo et al., 2011) and the time derivative signal of MSN activity (Kim et al., 2020). The time derivative signal of MSN activity is the summation of the signal on a disinhibitory direct pathway and the signal on an inhibitory indirect pathway from the MSN population (Morita et al., 2012; Keiflin and Janak, 2015; Kim et al., 2020). $\delta(t)$ is a continuous-time generalization of the temporal-difference signal (Schultz et al., 1997; Watabe-Uchida et al., 2017).

2.4. Goal-directed navigation by path planning

During replay in goal-directed navigation, the population vector of place cells $\mathbf{p}(t) = \sum_i r_i(t)\mathbf{x}^{(i)}$ approximately tracks the center of the activity bump hence $\mathbf{p}(t)$ depicts a discontinuous trajectory in the maze. Let $t_0^{(i)}$ denote the instant of the i-th time such that $\mathbf{p}(t)$ leaves the circular region with radius d from the rat's location $\mathbf{x}$, and $t_1^{(i)}$ denote the instant of the i-th time such that $\mathbf{p}(t)$ jumps back into the same circular region. $\{\mathbf{p}(t),\ t_0^{(i)} < t < t_1^{(i)}\}$ defines the i-th sub-trajectory with the initial direction given by $\mathbf{n}^{(i)} = \mathbf{p}(t_0^{(i)}) - \mathbf{x}$.

A line of studies (van der Meer and Redish, 2011; Howe et al., 2013; Atallah et al., 2014; London et al., 2018; Sjulson et al., 2018; O'Neal et al., 2022) has shown that the activity of MSN ramps up when a rat is approaching the goal location. Previous studies (Pfeiffer and Foster, 2013; Xu et al., 2019; Widloski and Foster, 2022) have also observed that rats show tendencies to follow replay trajectories that are approaching the goal location. Accordingly, the maximal MSN activity during the generation of the i-th sub-trajectory serves as the motivation for actually moving along $\mathbf{n}^{(i)}$, which is consistent with observations that the activity of MSN facilitates locomotion (Kravitz et al., 2010; Freeze et al., 2013).

Formally, if K sub-trajectories are generated, the probability of choosing the k-th direction $\mathbf{n}^{(k)}$ to move is given by,

$$P_{\mathbf{x}}(\mathbf{n}^{(k)}) = \frac{\exp(\beta \max_{t\in[t_0^{(k)},\, t_1^{(k)}]} V(t))}{\sum_{i=1}^{K} \exp(\beta \max_{t\in[t_0^{(i)},\, t_1^{(i)}]} V(t))}, \tag{9}$$

Where β characterizes the greediness level of behavior. The locomotion of the virtual rat is divided into periods of movements. During each movement period, the virtual rat first performs a period of replay to evaluate explorable paths, and then samples a direction from Equation (9) to move until the end of this movement period.

3. Experiments

3.1. Experimental setup

3.1.1. Subject

The subject is a virtual rat that reproduces the anatomical features of Long-Evans rats in the MuJoCo physics simulator (Todorov et al., 2012; Merel et al., 2020), as shown in Figure 1A. There are in total 67 bones modeled following average measurements from seven Long-Evans rats. At each simulation step, the proprioceptive input to the virtual rat is an 148-dimension vector that includes the position, velocity and angular velocity of each joint. The motor output from the virtual rat is a 38-dimension vector that contains the torque applied to each joint. Compared to numerical simulations typically used in rodent-navigation modeling literature, the behavior of the MuJoCo rat is much closer to the behavior of an actual rat. Therefore, a computational model that works on the MuJoCo rat might provide a deeper understanding about how neuronal activity relates to external behavior.

3.1.2. Control scheme

Controlling the rat to run is a challenging high-dimensional continuous control task (Schulman et al., 2016), so we use the TD3 deep RL algorithm (Fujimoto et al., 2018) to pretrain the rat in an open field to master several basic locomotion policies, including running forward along the current body direction, turning the body by a certain angle then running forward. The trained turning angles are 45, 90, 135, 180, −45, −90, and −135°. Each basic locomotion policy is represented by a deep neural network (DNN) that maps the proprioceptive input into the correct motor output to generate the desired locomotion. The pretraining details and hyperparameters used for the TD3 algorithm see Appendix, respectively. For moving along a direction sampled from Equation (9), the rat compares its current body direction with the sampled direction and invokes the closest locomotion policy to approximately run along the sampled direction for a period. Such control scheme constitutes a hierarchical controller (Merel et al., 2019) where the deep neural networks are low-level controllers and the CAN and MSN serve as the high-level controller. Figure 2A shows the interaction pattern of the controller with the environment.

For testing the proposed computational model, we build a large scale 10×10 m maze in MuJoCo as shown in Figure 1B. We conduct following four experiments in the maze environment.

3.1.3. Goal-fixed experiment

In the goal-fixed experiment, a location in the maze is set as the goal location, as shown by the green point in Figure 1B. The rat first randomly explores the maze for 50 trials starting from random initial positions to learn the inter-PC synaptic strength by Equation (2). Every exploration trial contains 6,000 simulation steps, corresponding to 120 s simulated time. Before every 150 simulation steps (i.e., 3 s), the rat randomly chooses a turning angle and runs along this direction in next 150 simulation steps. The place cell firing rates collected during exploration are used to update inter-PC synaptic strength as shown in Figure 2B. After exploration trials, the rat enters a brief rest period to learn the PC-MSN synaptic strength with Equation (8) by performing 60 s replay. The schematic diagram to illustrate the learning of the PC-MSN synapses is shown

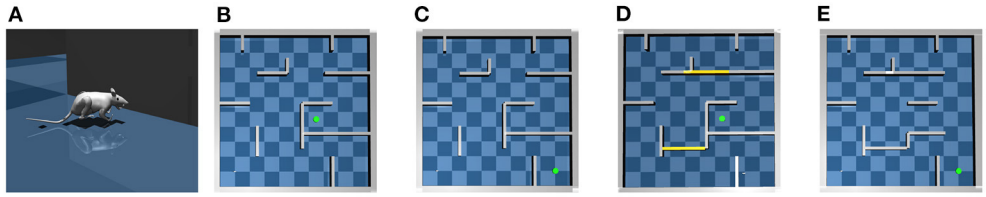

FIGURE 1
The virtual rat and maze environments. **(A)** The virtual rat used in experiments. **(B)** The maze environment used in the goal-fixed experiment. **(C)** The maze environment used in the goal-changing experiment. **(D)** The maze environment used in the detour experiment. **(E)** The maze environment used in the shortcut experiment.

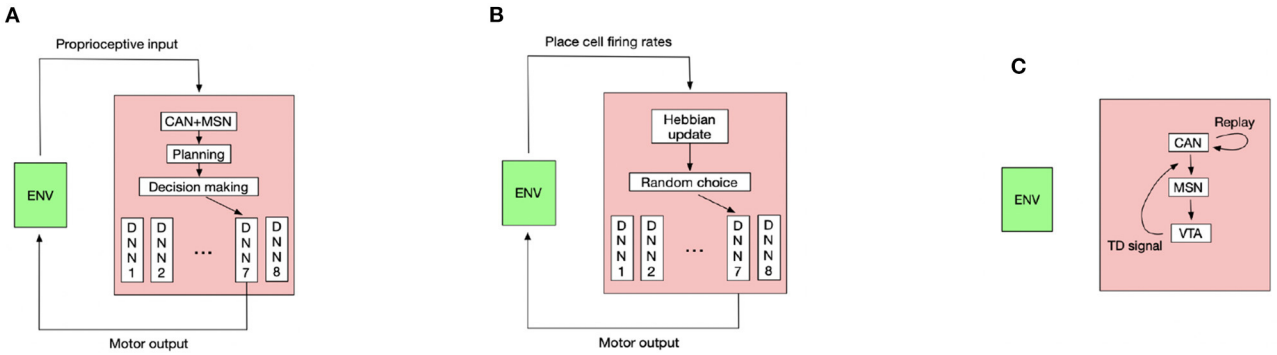

FIGURE 2
The diagram of key components in our model. The schematic figure for **(A)** planning and decision making during goal-directed navigation. **(B)** Learning inter-PC synapses during exploration. **(C)** Learning PC-MSN synapses during replay (ENV means environment and DNN means deep neural network).

in Figure 2C. After replay during rest, the rat performs 100 test trials starting from each integer grid point in the maze (i.e., (1,1), (1,2),...). Every test trial contains at most 6,000 simulation steps. Before every 150 simulation steps (i.e., 3 s), the rat performs 1 s awake replay to evaluate the MSN activity along possible paths. Then, it samples a new direction from Equation (9) and runs along this direction for next 100 simulation steps (i.e., 2 s). Once the rat enters a 1×1 m circle enclosing the goal location, this test trial is completed and treated as a successful trial.

3.1.4. Goal-changing experiment

In the goal-changing experiment, the rat first completes the time course of the goal-fixed experiment. After test trials, the goal location is suddenly changed as shown in Figure 1C. We define a dopaminergic reward function $h(\mathbf{x})$, which equals one if the rat is located within a 1×1 m circle enclosing the goal location and equals zero otherwise. After goal changing, the rat randomly explore the maze and the synaptic strength from place cells to the goal cell population is updated by a reward-modulated learning rule $\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \alpha_3(\mathbf{r}(\mathbf{x})^{\mathrm{T}} \cdot 1 - \mathbf{U}^{(n)}) \cdot h(\mathbf{x})$. After that, the rat re-enters a brief rest period to update the PC-MSN synaptic strength by replay with the updated goal cell population. Then the rat performs another 100 test trials.

3.1.5. Detour experiment

In the detour experiment, the rat first completes the time course of the goal-fixed experiment. After test trials, two critical passages in the maze are closed as shown in Figure 1D. After spatial layout changing, the rat reperforms 50 exploration trials to update the inter-PC synaptic strength. After that, the rat reperforms 120 s replay to update the PC-MSN synaptic strength. Then the rat performs another 100 test trials.

3.1.6. Shortcut experiment

In the shortcut experiment, the rat first completes the time course of the detour experiment. After test trials, three walls in the right side of the detour maze are removed as shown in Figure 1E. After the removement of walls, the rat once again randomly re-explores the maze for 50 trials and then reperforms 120 s replay to update the PC-MSN synaptic strength. After that, the rat performs another 100 test trials.

3.1.7. Experimental details

In above experiments, the preferential locations of 2,500 place cells are arranged on a regular grid of the maze without considering grid points that fall into areas under walls. The shortest path distance between pairs of preferential locations is computed by the Lee algorithm (Lee, 1961), based on breadth-first search (BFS) (Cormen et al., 2009). For Hebbian-like learning during exploration, σ is 0.3 and α_1 is 0.001. To simulate the dynamic of the CAN, Δt is 1 ms, h_0 is 0, τ_r is 2 ms, τ_I is 0.5 s, c_I is 10, and the global inhibitory strength added on the inter-PC synaptic strength is −0.3. For replay during the rest period, α_2 is 0.01, q is 0.1, τ_z is 0.5 s, ξ is 0.3 and the duration is 60 s. The initial condition of the CAN during the rest

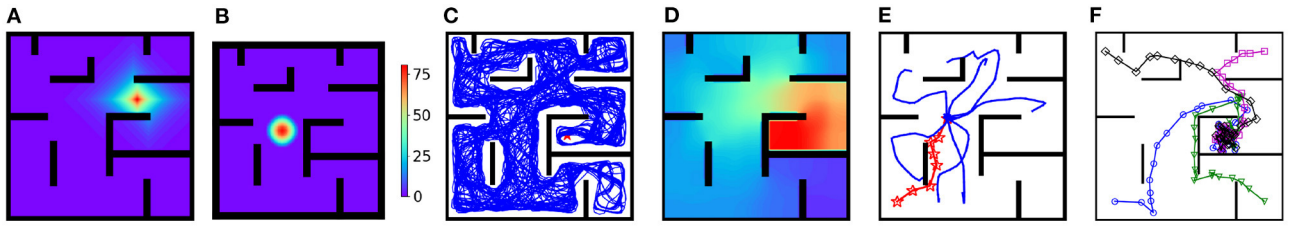

FIGURE 3
Results for the goal-fixed experiment. **(A)** Inter-PC synaptic strength after exploration. **(B)** A stationary activity bump of the CAN under a strong external input. **(C)** The replay trajectory during the rest period. **(D)** The activity of the MSN population. **(E)** The replay sub-trajectories during planning (the red trajectory is the rat's movement trajectory, and blue trajectories are replay sub-trajectories starting from the rat's location). **(F)** Exemplary movement trajectories during test trials (Markers are locations at which planning is performed).

period is a transient external input with amplitude 10 centered at the goal location for 10 ms. For replay during goal-directed navigation, the external input is persistent with amplitude 50, the radius d is 0.5 m, β is 10, and the duration is 1 s.

3.2. Experimental results

3.2.1. Goal-fixed experiment

Figure 3A shows the strength of synapses projected from the place cell with preferential location (7, 6.4) to other place cells after exploration. The synaptic strength decays from the location (7, 6.4) along shortest paths to other locations, so the geodesic proximity between preferential locations of place cells is encoded by inter-PC synaptic strength. Figure 3B shows a stationary activity bump centered at the rat's location under a strong external input with amplitude 100. After removing the external input during the rest period, the activity bump starts to deviate from its current position, and Figure 3C shows the trajectory of the population vector of place cells. Interestingly, the trajectory can go around walls and well cover the whole maze. After replay during rest, the activity of the MSN, as a function of the location of the rat with a stationary activity bump, is shown in Figure 3D. Interestingly, the activity of MSN ramps up when the rat approaches the goal location from any initial locations, which conforms to a series of observations about MSN (van der Meer and Redish, 2011; Howe et al., 2013; Atallah et al., 2014; London et al., 2018; Sjulson et al., 2018; O'Neal et al., 2022). Accordingly, the MSN activity encodes the geodesic proximity between each location and the goal location. Figure 3E shows the sub-trajectories generated during a planning period in test trials. Strikingly, without any randomization mechanisms involved in the dynamic, the sub-trajectories can nearly uniformly explore each possible path to a considerable depth. This is especially beneficial if the MSN activity at a local area is either flat (uninformative) or with a wrong shape (misguiding). The success rate of test trials is 100% and Figure 3F shows several exemplary locomotion trajectories.

To measure the performance of the virtual rat during test trials, we define the normalized latency of a successful trial as the time taken to reach the goal location divided by the shortest path distance between the initial position and the goal location. The normalized latency eliminates the influence of different initial positions on the time required to reach the goal. It measures the average time taken to approach the goal location per meter hence it characterizes the intrinsic efficiency of the navigation independent of initial positions.

To show the influence of parameters on performance, the goal-fixed experiment is repeated under varying values of a considered parameter while keeping other parameters as default values. As Figure 4A shows, the success rate of test trials improves when the number of trials for exploration increases but the success rate nearly keeps at 1 as long as at least 50 exploration trials are performed. As shown in Figure 4B, the normalized latency reduces with an increasing number of exploration trials and fluctuates around 5 s/m after 50 exploration trials. With an inadequate number of exploration trials, the spatial experiences of the virtual rat can't fully cover the maze. As a result, the synaptic strength between place cells with preferential locations at those under-explored areas are weak, which prevents the replay during rest to fully explore the maze and impairs the learning of PC-MSN synaptic strength consequently.

As Figure 4C shows, the success rate increases with longer duration for replay during rest and saturates at nearly 1 as long as at least 7 s are allowed. Although the MSN activity after only 7 s of replay has only a rough and unsmooth trend to ramp up (not shown), the planning trajectories lookahead over a long distance, enough to overcome the local unsmoothness and utilize the long-range trend to find the correct direction to the goal location. Such rapid learning is also observed in rodent experiments (Rosenberg et al., 2021) and it might be very important for the survival of rodents in a changing environment. In contrast, the three-factor rule using the accumulating trace update $\dot{z}_i(t) = -z_i(t)/\tau_z + r_i(t)V(t)$ (Izhikevich, 2007; Gerstner et al., 2018) fails to improve the success rate with an increasing duration for replay. The reason for the inferior performance is that the accumulating trace overly strengthens the PC-MSN synapses of frequently activated place cells even if their firing field centers are far away from the goal location. Such frequency-dependence leads to local peaks of the MSN activity that might attract the rat at places without rewards. As shown in Figure 4D, the normalized latency of successful trials fluctuates around 5 s/m and shows little dependence on the duration for replay during rest.

As shown by Figures 4E, F, the success rate generally decreases and the normalized latency generally increases when the period between decision making is prolonged, because infrequent decision making fails to timely re-adjust the movement direction that has became suboptimal over a distance of locomotion. In contrast, too frequent decision making (e.g., 2 s) also harms the performance

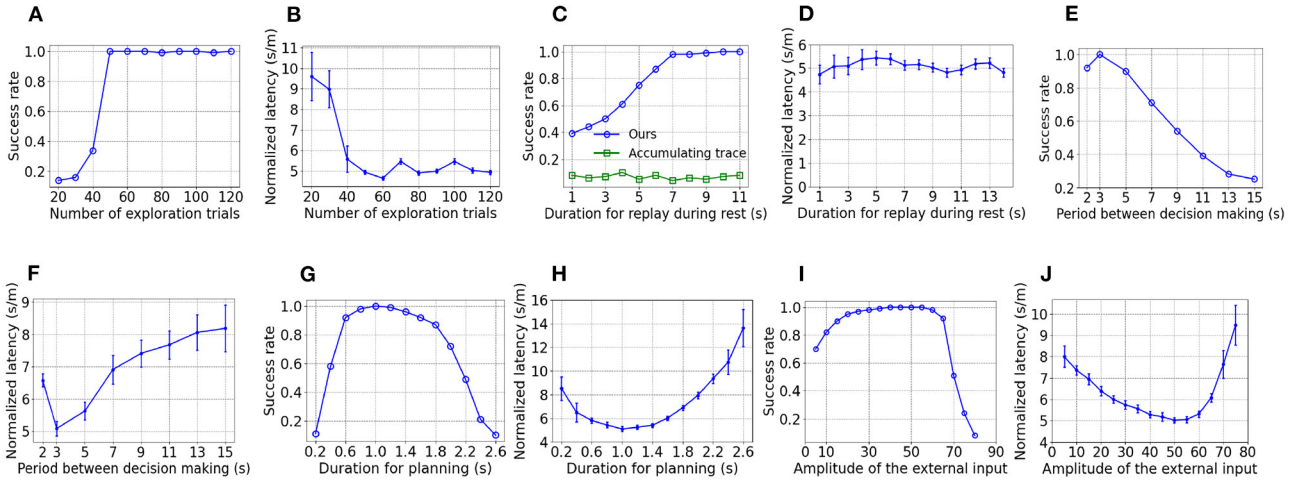

FIGURE 4
The influence of parameters on performance. **(A, B)** are the influence of the number of exploration epochs. **(C, D)** are the influence of duration for replay during rest. **(E, F)** are the influence of the period between decision making. **(G, H)** are the influence of duration for planning. **(I, J)** are the influence of the external input.

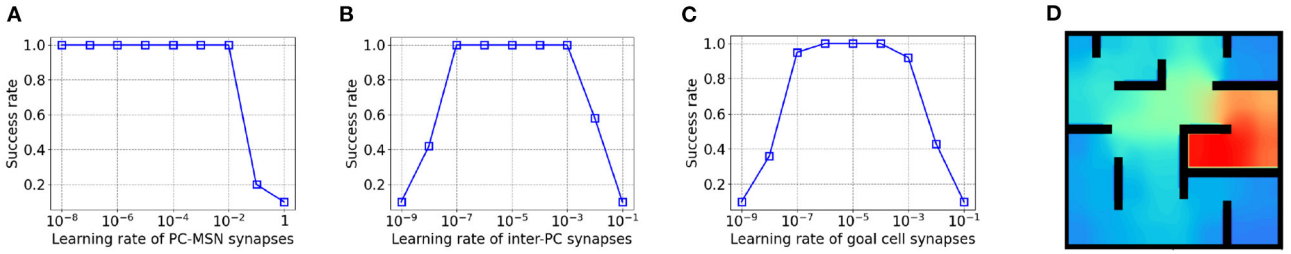

FIGURE 5
The influence of learning rates on performance in goal-fixed experiment. **(A)** The PC-MSN synaptic learning rate. **(B)** The inter-PC synaptic learning rate. **(C)** The PC to goal cell synaptic learning rate. **(D)** The MSN activity with the same transfer function and threshold for place cells and inhibitory neurons.

because the overhead of planning (e.g., 1 s/2 s = 50%) increases so that the time used for locomotion is reduced, which slowdowns the progress toward the goal location.

As Figures 4G, H show, when the duration for planning is shorter than 1 s, the performance improves with a longer duration for planning due to the ability to evaluate more directions. In contrast, when the duration for planning is longer than 1 s, the performance deteriorates due to the increasing overhead of planning and the shortened time for locomotion. As Figures 4I, J show, when the amplitude of the external input is smaller than 50, the performance improves with a larger external input, due to a reduced length of sub-trajectories leading to an increasing number of sub-trajectories during planning. However, when the amplitude of the external input is larger than 50, the performance collapses because it becomes more difficult to deviate the activity bump so that the number of sub-trajectories reduces rapidly.

As shown in Figure 5A, the success rate keeps at 1 even for a very small PC-MSN synaptic learning rate but the performance collapses for a learning rate larger than 10^{-2} due to divergence. Differently, as Figure 5B shows, either too large or too small inter-PC synaptic learning rate can significantly impair the performance due to divergence or slow learning, respectively. Similar trend is observed for PC to goal cell synaptic learning rate, as shown in Figure 5C. We have re-performed the goal-fixed experiment using the same transfer function and threshold for both place cells and inhibitory neurons. As shown in Figure 5D, the learned MSN activity under such settings still shows the desired ramping pattern as for inhibitory neurons without using a transfer function and threshold.

Figure 6 shows the time trace of place cell firing rates, the chemical trace and PC-MSN synaptic strengths at 4, 8, 12, 16, 20, and 24 s during replay period. At each time point, the firing rate vector $\mathbf{r}(t)$ is a localized activity bump in the maze, possibly truncated by the barriers of the maze (Figure 6A). The chemical trace at each time point decays along the most recent trajectory of the current activity bump, so that a dopamine signal can effectively strengthen those recently activated PC-MSN synapses (Figure 6B). At an early stage of replay, the PC-MSN synaptic strength is quite unsmooth and does not show trend to ramp up (Figure 6C, 4 s). With more replay, the PC-MSN synaptic strength becomes more and more smooth and finally converges to the desired ramping shape (Figure 6C, 24 s). As Figure 7A shows, the goal cell is activated at time points when the activity bump visits the goal location. Despite the goal cell firing events are sparse, the temporal difference signal fluctuates frequently to create more learning signals

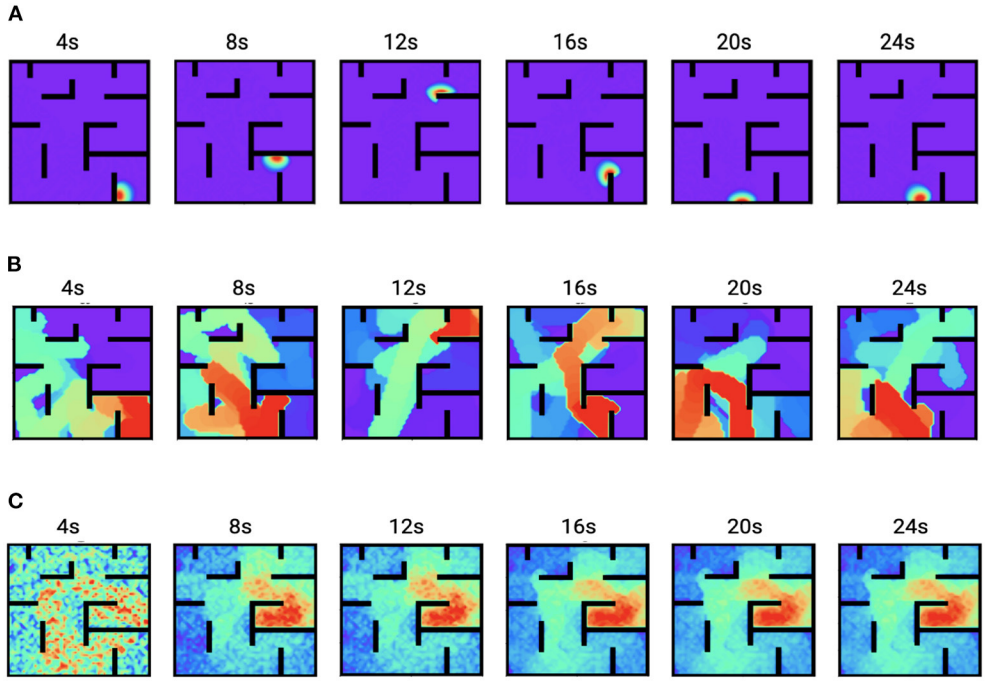

FIGURE 6
Time traces of variables during the replay period. **(A)** The firing rates $\mathbf{r}(t)$. **(B)** The chemical trace $\mathbf{z}(t)$. **(C)** The PC-MSN synaptic strength $\mathbf{W}(t)$.

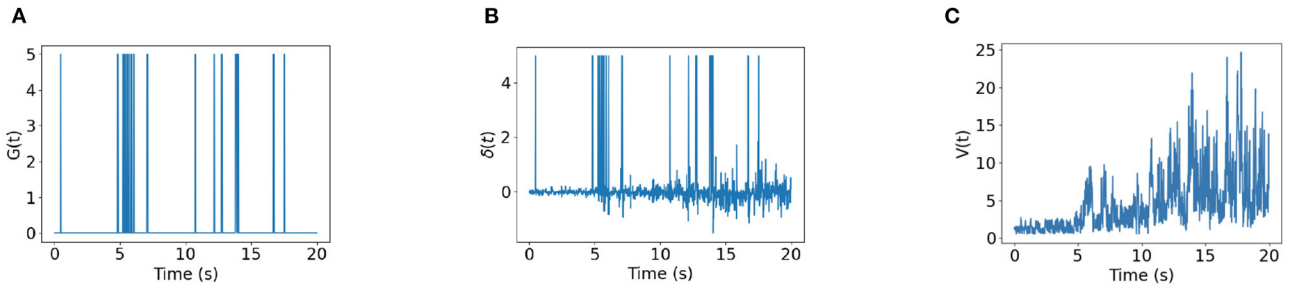

FIGURE 7
Time traces of variables. **(A)** The goal cell activity $G(t)$. **(B)** The temporal difference $\delta(t)$. **(C)** The MSN activity $V(t)$.

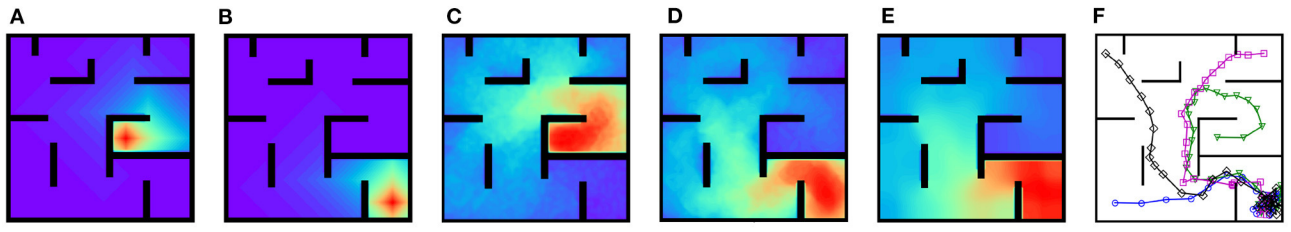

FIGURE 8
Results for the goal-changing experiment. The place field of the goal cell population before **(A)** and after **(B)** goal changing. The PC-MSN synaptic strength before **(C)** and after **(D)** reperforming replay during rest. **(E)** The activity of the MSN population. **(F)** Exemplary movement trajectories during test trials (Markers are locations at which planning is performed).

(Figure 7B). As a result, the MSN activity pattern is gradually built up through the learning of the PC-MSN synaptic strengths (Figure 7C).

3.2.2. Goal-changing experiment

Figures 8A, B show the place fields of the goal cell population before and after goal changing, respectively. The place fields are

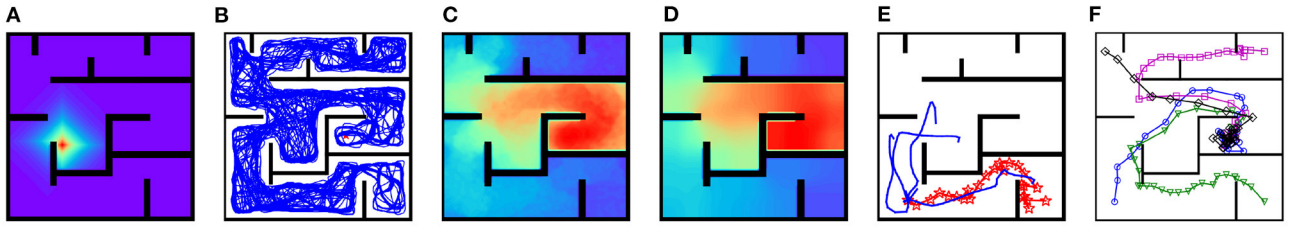

FIGURE 9
Results for the detour experiment. **(A)** Inter-PC synaptic strength after re-exploration. **(B)** The replay trajectory after introducing new walls. **(C)** The PC-MSN synaptic strength after replay during rest. **(D)** The activity of the MSN population. **(E)** Replay sub-trajectories during planning. **(F)** Exemplary movement trajectories during test trials (Markers are locations at which planning is performed).

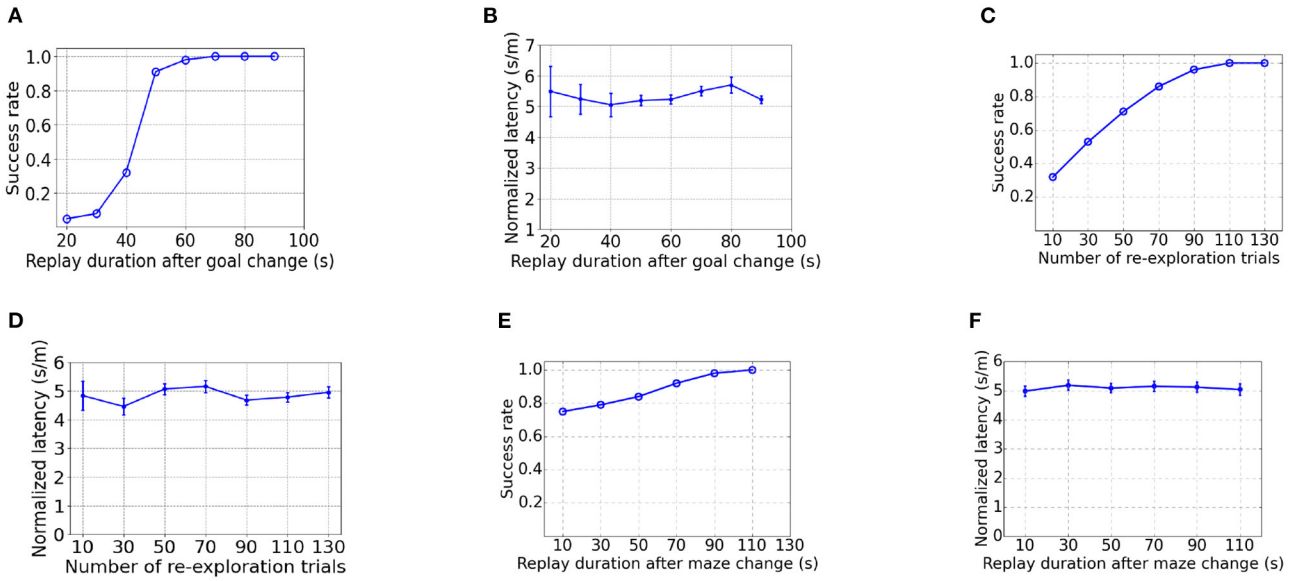

FIGURE 10
The influence of parameters on performance in the goal-changing or detour experiment. **(A, B)** are the influence of the duration for replay during rest after goal change in the goal-changing experiment. **(C, D)** are the influence of the number of re-exploration trials after walls are introduced in the detour experiment. **(E, F)** are the influence of the duration for replay during rest after walls are introduced in the detour experiment.

constrained by walls and decay along paths away from the center. Before goal changing, the PC-MSN synaptic strength encodes the geodesic proximity between each location and the old goal location (Figure 8C). After reperforming replay during rest, the PC-MSN synaptic strength now re-encodes the geodesic proximity between each location and the new goal location (Figure 8D). As a result, the MSN activity ramps up along paths to the new goal location (Figure 8E). The success rate of test trials after goal changing is still 100% and Figure 8F shows several exemplary locomotion trajectories.

3.2.3. Detour experiment

After re-exploration of the detour maze, the inter-PC synaptic strengths are updated and the strength of synapses projected from the place cell with preferential location (3, 4) to other place cells is shown in Figure 9A. The inter-PC synaptic strength re-encodes the updated adjacency relation between locations in the detour maze. Under updated inter-PC synaptic strength, the replay trajectory during rest is blocked by new walls (Figure 9B), signifying an adaptation to the new layout. After replay, both the PC-MSN synaptic strength and the MSN activity ramps up along detours to the goal location (Figures 9C, D). Interestingly, the replay sub-trajectories during planning explore the area behind the new wall and evaluate the MSN activity there (Figure 9E). As a result, the virtual rat achieves a 100% success rate during test trials in the detour maze and Figure 9F shows several exemplary movement trajectories. Starting from initial positions far away from the goal location, the rat can find a detour to the goal location. Such detour behavior resembles that observed in rodents (Tolman and Honzik, 1930; Alvernhe et al., 2011), which has been considered a hallmark of cognition (Tolman, 1948; Widloski and Foster, 2022).

As Figure 10A shows, the success rate increases with a longer duration for replay after goal changing. It takes a significantly longer time to achieve 100% success rate than learning from scratch (Figures 10A vs. 4C), due to the time taken to overwrite the PC-MSN synaptic strengths built in the goal-fixed experiment. As shown by Figure 10B, the normalized latency of successful trials is insensitive to the duration for replay during rest after goal changing. After new walls are introduced in the detour experiment, it requires more

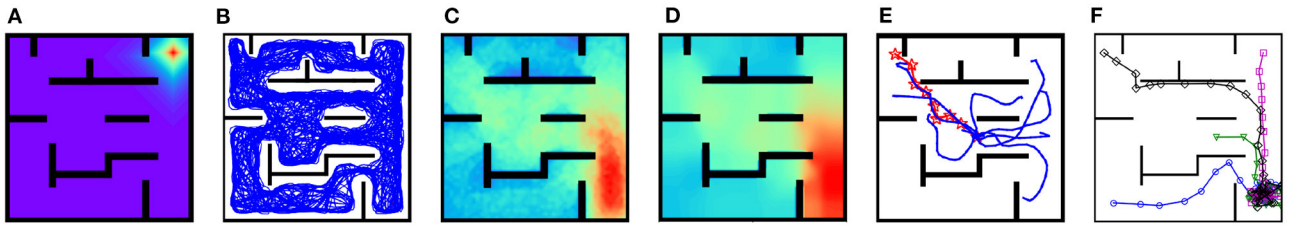

FIGURE 11
Results for the shortcut experiment. **(A)** Inter-PC synaptic strength after re-exploring the shortcut maze. **(B)** The replay trajectory after removing walls. **(C)** The PC-MSN synaptic strength after replay during rest. **(D)** The activity of the MSN population. **(E)** Replay sub-trajectories during planning. **(F)** Exemplary movement trajectories during test trials (Markers are locations at which planning is performed).

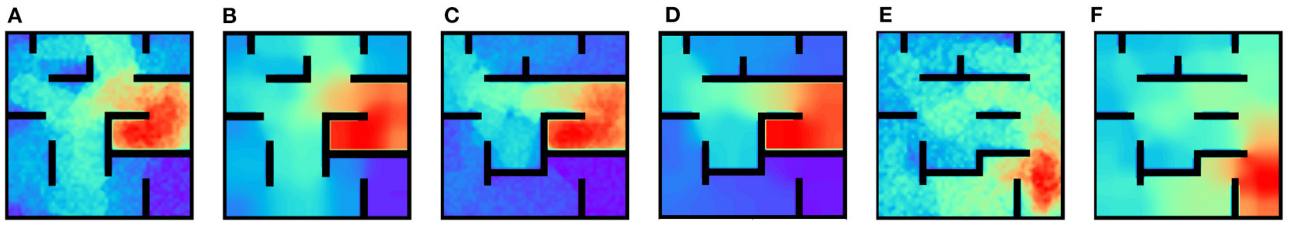

FIGURE 12
Results with simultaneously updated inter-PC and PC-MSN synapses. **(A, B)** The PC-MSN synaptic strength and MSN activity in the goal-fixed experiment. **(C, D)** The PC-MSN synaptic strength and MSN activity in the detour experiment. **(E, F)** The PC-MSN synaptic strength and MSN activity in the shortcut experiment.

re-exploration trials to achieve 100% success rate (Figures 10C vs. 4A), due to the need to update the inter-PC synaptic strength built in the goal-fixed experiment. Similarly, the normalized latency of successful trials is insensitive to the number of re-exploration trials (Figure 10D). As Figure 10E shows, even a short duration for replay (e.g., 10 s) is adequate to achieve a success rate close to 80%, because the MSN activity pattern in the goal-fixed experiment is similar to the desired MSN activity pattern in the detour experiment (Figures 9D vs. 3D). However, a longer duration for replay is still required to achieve a 100% success rate. Still, the normalized latency of successful trials is insensitive to the duration for replay (Figure 10F).

3.2.4. Shortcut experiment

After re-exploring the shortcut maze, the strength of synapses projected from the place cell with preferential location (9, 9) to other place cells decays along newly available shortcuts (Figure 11A). As a result, the replay trajectory during rest freely traverses these shortcuts (Figure 11B). After replay, both the PC-MSN synaptic strength and the MSN activity ramp up along newly available shortcuts (Figures 11C, D). The replay sub-trajectories during planning can lookahead along newly available shortcuts (Figure 11E). The success rate during test trials in the shortcut maze is 100% and Figure 11F shows several exemplary locomotion trajectories, all following the shortcut to the goal location.

3.2.5. Updating inter-PC and PC-MSN synapses simultaneously

In this set of experiments, we relax the requirement of learning inter-PC synaptic strength only during exploration phase and learning PC-MSN synaptic strength only during replay phase. We re-perform the goal-fixed, detour and shortcut experiment with inter-PC and PC-MSN synapses simultaneously updated during both exploration and replay periods. Specifically, during exploration phase, we also use Equation (8) to update PC-MSN synaptic strength. During replay phase, we further use Equation (2) to update inter-PC synaptic strength in Equation (4). As shown in Figures 12A–F, the PC-MSN synaptic strength and the MSN activity can be successfully learned in all experiments. Guided by the learned MSN activity, the virtual rat can successfully navigate to the goal location in all experiments.

3.2.6. Using the same governing equation during exploration and replay phases

In this set of experiments, we use Equations (4), (5) to replace Equation (1) during the exploration phase. The amplitude of the external input is set as 80 and other parameters are set as default values given above. Figures 13A–D show the strength of synapses projected to other place cells from the place cells with preferential locations (7, 6.4), (9, 9), (3, 2), (2, 8), respectively, after exploration phase. Using Equation (4) with Equation (2) during exploration can still learn the desired synaptic strength patterns, despite less regular than using Equation (1) with Equation (2) (Figures 13A vs. 3A). After the exploration phase, the same governing (Equation 4) with vanishing external input is used to learn the PC-MSN synaptic strength during replay (Figure 13E). After replay, the MSN activity ramps up correctly so that the virtual rat can navigate to the goal location from arbitrary initial locations (Figure 13F).

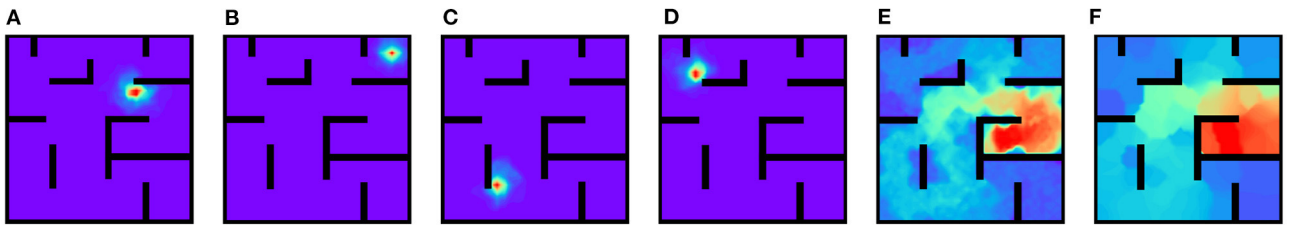

FIGURE 13
Results using the same governing equation during exploration and replay phases. **(A–D)** The inter-PC synaptic strength projected from several place cells after exploration. **(E, F)** The PC-MSN synaptic strength and MSN activity after replay.

4. Discussion

This paper proposed a computational model that uses a continuous attractor network with feedback inhibition to generate layout-conforming replay for achieving reward-based learning and planning, two key functions that support flexible navigation in a dynamically changing maze. We have shown that these functions are achieved by a single CAN under a different strength of external inputs. Reward-based learning is implemented by three-factor learning of PC-MSN synaptic strength during replay under vanishing external inputs. Planning is achieved by evaluating the MSN activity along lookahead trajectories during replay under non-zero external inputs. Combining these two functions, this paper sheds a light on how the learning of flexible navigation in a maze is achieved by the activity of an ensemble of interacting place cells. Besides, our model is beneficial to the design of neuroscience-inspired artificial intelligence (Hassabis et al., 2017). We have shown that incorporating state-of-the-art neuroscientific insights about learning, memory and motivation into the design loop of an artificial agent is important toward developing artificial intelligence that matches the performance of animals.

Many works use attractor states of a CAN to model the population activity of head-direction cells (Skaggs et al., 1995; Blair, 1996; Zhang, 1996) or place cells (Samsonovich and McNaughton, 1997; Battaglia and Treves, 1998; Tsodyks, 1999; Stringer et al., 2002; McNaughton et al., 2006). In these models, the synaptic strength is assumed to be a negative exponential function of the Euclidean distance between preferential angles or locations of two cells, possibly with periodic boundary conditions. This assumption is valid for head-direction cells in a ring attractor space or place cells in a barrier-free plane attractor space but it ignores the influence of barriers in non-Euclidean spaces, such as a maze.

Next, we discuss previous computational models for the generation of replay with a CAN. Two computational models (Hopfield, 2009; Azizi et al., 2013) use integrate-and-fire cells with spike-frequency adaptation (SFA) to simulate drifts of the activity bump. In these models, the SFA is modeled by an adaptive inhibitory current fed into each place cell, which is increased after every spike of a place cell. Despite functionally similar to the feedback inhibition in our model, the adaptive inhibitory current is used to model the dynamic blockage of Ca^{2+}-dependent K^{+} channels (Shao et al., 1999; Faber and Sah, 2003; Hopfield, 2009) rather than interactions between place cells and hippocampal GABAergic interneurons. One model (Itskov et al., 2011) and another model (Romani and Tsodyks, 2015) use adaptive thresholds and short-term depression, respectively, to generate drifts of the activity bump. However, the generation of hippocampal replay requires interactions between place cells and inhibitory interneurons (Schlingloff et al., 2014; Stark et al., 2014; Buzsaki, 2015), which is not captured by these two models. Other models (Zhang, 1996; Spalla et al., 2021) generate bump drifts by introducing an asymmetric component into the inter-PC synaptic strength. However, it remains unclear how such a systematic perturbation of inter-PC synaptic strength can be implemented in the hippocampus. Besides, the above models pre-configure the inter-PC synaptic strength as a function decaying with the Euclidean distance between preferential locations, so the bump drifts generated by these models violate the constraints imposed by walls of a maze. Among the aforementioned models, some of them (Hopfield, 2009; Azizi et al., 2013; Romani and Tsodyks, 2015; Spalla et al., 2021) consider multiple possible mapping from place cells to place field centers, called multi-chart, to model a global remapping of place cells across different environments (Alme et al., 2014). However, it's observed that the place field centers are largely stable under different layout of a maze (Alvernhe et al., 2011; Widloski and Foster, 2022). Accordingly, we keep the mapping of place cells consistent across different layout.

The early works (Muller et al., 1991, 1996) first proposed that the inter-PC synaptic strength should decay with the Euclidean distance between their firing field centers due to the long-term potentiation (Isaac et al., 2009). In these works (Muller et al., 1991, 1996), the hippocampus is modeled as a weighted graph where place cells are nodes and synapses are edges with synaptic resistance (reciprocal synaptic strength) as edge weights. To navigate to the goal location, this model uses the Dijkstra's algorithm (Cormen et al., 2009) to search a shortest path in the graph from the currently activated place cell to the place cell activated at the goal location. However, it remains unclear how such complex path search computations are implemented in rodent brains. Our model has shown that the learning of PC-MSN synaptic strength by replay during rest encodes the geodesic proximity between each place field center and the goal location, which is read out subsequently during awake replay to plan a path toward the goal location.

If the firing rate vector of place cells is interpreted as a feature vector of the rat's location, then the activity of MSN can be interpreted as a linear value function approximation (Sutton and Barto, 2018) of the rat's location. Such an interpretation relates our model to a line of spatial navigation models based on reinforcement learning (RL). In one model (Brown and Sharp, 1995), the synaptic strengths from PC to motor neurons are learned using policy gradient like reinforcement learning (Williams, 1992; Sutton et al.,

1999) to build correct place-response associations by trial-and-errors. Another line of models (Foster et al., 2000; Gustafson and Daw, 2011; Banino et al., 2018) use the firing rates of place cells as the set of basis functions for representing the policy and the value function, which are trained by the actor-critic algorithm (Sutton and Barto, 1998) during interactions with the environment. Unfortunately, navigation behavior learned with such model-free RL algorithms is inflexible to changes of the goal location or the spatial layout.

Next, we discuss computational models of spatial navigation based on model-based RL. Inspired from the Dyna-Q algorithm (Sutton, 1990), a computational model (Johnson and Redish, 2005) learns a *Q*-value function during offline replay of a network of place cells. This model interprets the inter-PC synaptic strength matrix as a state transition matrix that sequentially activates a place cell with a maximal synaptic strength with the currently activated place cell. Such one-by-one replay of place cells with binary activity can hardly capture the bump-like activity during hippocampal replay (Pfeiffer and Foster, 2013; Stella et al., 2019; Widloski and Foster, 2022). Besides, the Q-value function is learned by the biologically unrealistic one-step TD(0) algorithm (i.e., Q-learning) without using a chemical trace. The SR-Dyna model (Russek et al., 2017; Momennejad, 2020) learns a successor representation (SR) matrix by TD(0) during one-by-one replay of binary place cells. After replay, the *Q*-value function is computed simply by a dot product of the SR matrix and the reward function. However, the neural substrate of the SR remains unclear and controversial (Russek et al., 2017; Stachenfeld et al., 2017).

Next, we discuss computational models of spatial navigation without using a value representation. Two studies (Blum and Abbott, 1996; Gerstner and Abbott, 1997) proposed that the long-term potentiation (LTP) of inter-PC synapses during exploration slightly shifts the location encoded by the population vector away from the rat's actual location. Such shifts are used as a vector field tending toward the goal location to navigate the rat to the goal. Our model similarly uses a shift of the population vector during awake replay to define the directional vector of each replay sub-trajectory. However, our model does not construct a static vector field toward the goal but relies on online planning with respect to the MSN activity to choose a direction to follow. Another model (Burgess and O'Keefe, 1996) directly uses the place field of a putative goal cell to navigate to the goal by moving along the gradient direction of the place field. However, the gradient of the place field vanishes at locations far away from the goal location due to the limited range of the goal cell's place field. In our model, the activity of MSN ramps up even at locations far away from the goal location. Furthermore, planning with respect to the MSN activity can avoid wrong gradient directions due to local unsmoothness of the MSN activity.

The work (Ponulak and Hopfield, 2013) uses a similar CAN as in Hopfield (2009) to generate a wavefront propagation of activity on the sheet of place cells after visiting the goal location. With anti-STDP, the inter-PC synaptic strength will be biased toward the goal location after wavefront propagation, which can be interpreted as a synaptic vector field to guide the navigation toward the goal. However, the activity propagation during hippocampal replay is sequential and follows particular directions (Pfeiffer, 2017) rather than propagating over all directions as in the above model. Another model (Gonner et al., 2017) uses a learned goal location representation of place cells in dentate gyrus (DG) as the input of place cells in CA3 to move the activity bump toward the goal location. The end point of the moving bump is used for vector-based navigation. However, the bump trajectories generated by this model always converge to the goal location, falling short of explaining the diversity of directions and end points of replay trajectories.

Our model suggests that the hippocampal goal cells, but not hippocampal place cells, contribute to a dopamine temporal difference signal through the disinhibitory hippocampus-lateral septum-VTA pathway. Anatomically, it's possible that other place cells might project to the lateral septum (LS). However, some experiments (Wirtshafter and et al., 2019) have shown that most LS cells show ramping activity when a rat is approaching the goal location, quite similar to the firing pattern of a goal cell. The reason for such observations might be that the synaptic efficacy is stronger for projections from goal cells to the LS than projections from other place cells to the LS, possibly because the goal cell to LS projections have received more dopamine release at the goal location.

Next, we discuss limitations of our model. Our model assumes place cells fire independently following Equation (1) without interacting with each other during the exploration phase. This assumption holds at the beginning of the exploring phase because the inter-PC connections are too weak to induce synaptic transmissions among place cells. However, as learning proceeds, some connections become stronger so that interactions among place cells can not be neglected. Our experiments have shown that adopting the more realistic Equation (4) instead of Equation (1) is able to learn inter-PC synaptic strengths without disabling interactions among place cells during the exploration phase. We leave a formal analysis of the unified governing Equation (4) during different phases for our future work. Our model defines the dynamic of the chemical trace $\mathbf{z}(t)$ in Equation (8) separately for two cases according to whether the joint firing rate of place cells and MSN exceeding a threshold. It's possible to unify these two cases using the dynamic equation $\dot{z}_i(t) = -z_i(t)/\tau_z + f[r_i(t)V(t)]$, where the transfer function $f[x]$ equals zero if $x \leq q$ and equals x otherwise. However, such dynamic requires an amount of time to relax to the joint firing rate, possibly leading to performance issues. We leave the derivation of a unified dynamic equation of the chemical trace for our future study.

In our model, awake replay serves the planning and evaluation purpose. However, there is evidence that awake replay also contributes to reward learning (Jadhav et al., 2012), possibly through (reverse) replay at reward sites. Since the replay dynamic is driven by the same CAN, it's direct to use the Equation (8) during awake replay to learn PC-MSN synaptic strength without needing for an extended sleep period. However, learning PC-MSN synaptic strength during awake replay requires the behavioral policy of the rat to be explorative initially so that regions near the goal location can be visited by chance. Since our model encodes the goal location information by the MSN activity, our modeling of place fields does not capture some weak off-center goal-related activity observed in Hok et al. (2007). Besides, the ubiquitous existence of such goal-related activity of place cells remains controversial and it might introduce ambiguity to the coding of place cells (Poucet and Hok, 2007). Strictly speaking, the most detailed model of symmetric STDP (Mishra et al., 2016) should consider spike trains of pre- and post-synaptic neurons. Since the pre- and post-synaptic spike trains of a pair of place cells can be modeled as independent Poisson processes before learning the inter-PC synaptic strength, the synaptic update under the rate-based Hebbian-like rule equals the average synaptic updates under symmetric STDP using spike trains (Kempter et al., 1999). Our model

requires the exploration performed by the virtual rat fully covers the whole maze. However, animals can navigate in unexplored areas based on landmarks. Our model of the feedback inhibition in the CAN is a simplified abstraction of the interaction between place cells and interneurons without modeling mutual interactions among interneurons (Schlingloff et al., 2014; Stark et al., 2014; Buzsaki, 2015). Our model of the MSNs does not capture observations that the activity of some MSN assemblies encodes other behavioral information, such as locomotion initialization (Kravitz et al., 2010; Freeze et al., 2013) or speed (Kim et al., 2014; Fobbs et al., 2020).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was supported by Academy of Military Sciences (No. 22TQ0904ZT01025).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abrams, T., and Kandel, E. (1988). Is contiguity detection in classical conditioning a system or a cellular property? learning in aplysia suggests a possible molecular site. *Trends Neurosci.* 11, 128–135. doi: 10.1016/0166-2236(88)90137-3

Alme, C., Miao, C., Jezek, K., Treves, A., Moser, E., and Moser, M. (2014). Place cells in the hippocampus: eleven maps for eleven rooms. *Proc. Natl. Acad. Sci. U.S.A.* 111, 18428–18435. doi: 10.1073/pnas.1421056111

Alvernhe, A., Save, E., and Poucet, B. (2011). Local remapping of place cell firing in the Tolman detour task. *Eur. J. Neurosci.* 33, 1696–1705. doi: 10.1111/j.1460-9568.2011.07653.x

Atallah, H., McCool, A., Howe, M., and Graybiel, A. (2014). Neurons in the ventral striatum exhibit cell-type-specific representations of outcome during learning. *Neuron* 82, 1145–1156. doi: 10.1016/j.neuron.2014.04.021

Azizi, A., Wiskott, L., and Cheng, S. (2013). A computational model for preplay in the hippocampus. *Front. Comput. Neurosci.* 7, 161. doi: 10.3389/fncom.2013.00161

Banino, A., Barry, C., Uria, B., and et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 429–433. doi: 10.1038/s41586-018-0102-6

Battaglia, F., and Treves, A. (1998). Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Phys. Rev.* 58, 7738–7753. doi: 10.1103/PhysRevE.58.7738

Blair, H. (1996). Simulation of a thalamocortical circuit for computing directional heading in the rat. *Adv. Neural Inf. Process. Syst.* 8, 152–158.

Blum, K., and Abbott, L. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Comput.* 8, 85–93. doi: 10.1162/neco.1996.8.1.85

Brown, M., and Sharp, P. (1995). Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. *Hippocampus* 5, 171–188. doi: 10.1002/hipo.450050304

Brzosko, Z., Mierau, S., and Paulsen, O. (2019). Neuromodulation of spike-timing-dependent plasticity: past, present, and future. *Neuron Rev.* 103, 563–581. doi: 10.1016/j.neuron.2019.05.041

Burgess, N., and O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus* 6, 749–762. doi: 10.1002/(SICI)1098-1063(1996)6:6andlt;749::AID-HIPO16andgt;3.0.CO;2-0

Buzsaki, G. (2015). Hippocampal sharp wave-ripple: a cognitive biomarker for episodic memory and planning. *Hippocampus* 25, 1073–1188. doi: 10.1002/hipo.22488

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2009). *Introduction to Algorithms, 3rd Edn*. Cambridge, MA: MIT Press.

Faber, E., and Sah, P. (2003). Ca^2+-activated K^+ (BK) channel inactivation contributes to spike broadening during repetitive firing in the rat lateral amygdala. *J. Physiol.* 552, 483–497. doi: 10.1113/jphysiol.2003.050120

Floresco, S., Blaha, C., Yang, C., and Phillips, A. (2001). Modulation of hippocampal and amygdalar-evoked activity of nucleus accumbens neurons by dopamine: cellular mechanisms of input selection. *J. Neurosci.* 21, 2851–2860. doi: 10.1523/JNEUROSCI.21-08-02851.2001

Fobbs, W., Bariselli, S., Licholai, J., and et al. (2020). Continuous representations of speed by striatal medium spiny neurons. *J. Neurosci.* 40, 1679–1688. doi: 10.1523/JNEUROSCI.1407-19.2020

Foster, D., Morris, R., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* 10, 1–16. doi: 10.1002/(SICI)1098-1063(2000)10:1andlt;1::AID-HIPO1andgt;3.0.CO;2-1

Foster, D., and Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440, 680–683. doi: 10.1038/nature04587

Freeze, B., Kravitz, A., Hammack, N., and et al. (2013). Control of basal ganglia output by direct and indirect pathway projection neurons. *J. Neurosci.* 33, 18531–18539. doi: 10.1523/JNEUROSCI.1278-13.2013

Fujimoto, S., Hoof, H., and Meger, D. (2018). "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning (ICML)* (Stockholm).

Fung, C., Wong, K., and Wu, S. (2010). A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Comput.* 22, 752–792. doi: 10.1162/neco.2009.07-08-824

Gauthier, J., and Tank, D. (2018). A dedicated population for reward coding in the hippocampus. *Neuron*. 99, 179–193. doi: 10.1016/j.neuron.2018.06.008

Gerstner, W., and Abbott, L. (1997). Learning navigational maps through potentiation and modulation of hippocampal place cells. *J. Comput. Neurosci.* 4, 79–94. doi: 10.1023/A:1008820728122

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: experimental support of neoHebbian three-factor learning rules. *Front. Neural Circ.* 12, 1–16. doi: 10.3389/fncir.2018.00053

Gonner, L., Vitay, J., and Hamker, F. (2017). Predictive place-cell sequences for goal-finding emerge from goal memory and the cognitive map: a computational model. *Front. Comput. Neurosci.* 11, 84. doi: 10.3389/fncom.2017.00084

Gustafson, N., and Daw, N. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Comput. Biol.* 7, 1–14. doi: 10.1371/journal.pcbi.1002235

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Hoboken, NJ: Prentice Hall.

Helmchen, F., Imoto, K., and Sakmann, B. (1996). Ca^2+ buffering and action potential-evoked Ca^2+ signaling in dendrites of pyramidal neurons. *Biophys. J.* 70, 1069–1081. doi: 10.1016/S0006-3495(96)79653-4

Hok, V., Lenck-Santini, P., Roux, S., and et al. (2007). Goal-related activity in hippocampal place cells. *J. Neurosci.* 27, 472–482. doi: 10.1523/JNEUROSCI.2864-06.2007

Hopfield, J. (2009). Neurodynamics of mental exploration. *Proc. Natl. Acad. Sci.* 107, 1648–1653. doi: 10.1073/pnas.0913991107

Howe, M., Tierney, P., Sandberg, S., and et al. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* 500, 575–579. doi: 10.1038/nature12475

Isaac, J., Buchanan, K., Muller, R., and Mellor, J. (2009). Hippocampal place cell firing patterns can induce long-term synaptic plasticity *in vitro*. *J. Neurosci.* 29, 6840–6850. doi: 10.1523/JNEUROSCI.0731-09.2009

Itskov, V., Curto, C., Pastalkova, E., and Buzsaki, G. (2011). Cell assembly sequences arising from spike threshold adaptation keep track of time in the hippocampus. *J. Neurosci.* 31, 2828–2834. doi: 10.1523/JNEUROSCI.3773-10.2011

Izhikevich, E. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* 10, 2443–2452. doi: 10.1093/cercor/bhl152

Jadhav, S., Kemere, C., German, P., and Frank, L. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science* 336, 1454–1458. doi: 10.1126/science.1217230

Jay, T. (2003). Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. *Progr. Neurobiol.* 69, 375–390. doi: 10.1016/S0301-0082(03)00085-6

Johnson, A., and Redish, A. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Netw.* 18, 1163–1171. doi: 10.1016/j.neunet.2005.08.009

Johnson, A., and Redish, A. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* 27, 12176–12189. doi: 10.1523/JNEUROSCI.3761-07.2007

Kasai, H., Ziv, N., Okazaki, H., and et al. (2021). Spine dynamics in the brain, mental disorders and artificial neural networks. *Nat. Rev. Neurosci.* 22, 407–422. doi: 10.1038/s41583-021-00467-3

Keiflin, R., and Janak, P. (2015). Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron* 88, 247–263. doi: 10.1016/j.neuron.2015.08.037

Kempter, R., Gerstner, W., and van Hemmen, J. (1999). Hebbian learning and spiking neurons. *Phys. Rev. E* 59, 4498–4514. doi: 10.1103/PhysRevE.59.4498

Kim, H., Malik, A., Mikhael, J., and et al. (2020). A unified framework for dopamine signals across timescales. *Cell* 183, 1600–1616. doi: 10.1016/j.cell.2020.11.013

Kim, N., Barter, J., Sukharnikova, T., and et al. (2014). Striatal firing rate reflects head movement velocity. *Eur. J. Neurosci.* 40, 3481–3490. doi: 10.1111/ejn.12722

Kravitz, A., Freeze, B., Parker, P., and et al. (2010). Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature* 466, 622–629. doi: 10.1038/nature09159

Lansink, C., Goltstein, P., Lankelma, J., McNaughton, B., and Pennartz, C. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* 7, e1000173. doi: 10.1371/journal.pbio.1000173

Lee, A., and Wilson, M. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* 36, 1183–1194. doi: 10.1016/S0896-6273(02)01096-6

Lee, C. (1961). An algorithm for path connections and its applications. *IRE Trans. Electron. Comput.* 10, 346–365. doi: 10.1109/TEC.1961.5219222

London, T., Licholai, J., and Szczot, I. (2018). Coordinated ramping of dorsal striatal pathways preceding food approach and consumption. *J. Neurosci.* 38, 3547–3558. doi: 10.1523/JNEUROSCI.2693-17.2018

Luo, A., Tahsili-Fahadan, P., Wise, R., and et al. (2011). Linking context with reward: a functional circuit from hippocampal CA3 to ventral tegmental area. *Science* 333, 353–357. doi: 10.1126/science.1204622

McNaughton, B., Battaglia, F., Jensen, O., Moser, E., and Moser, M. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience* 7, 663–678. doi: 10.1038/nrn1932

Merel, J., Aldarondo, D., Marshall, J., and et al. (2020). "Deep neuroethology of a virtual rodent," in *International Conference on Learning Representations (ICLR)* (Addis Ababa).

Merel, J., Botvinick, M., and Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nat. Commun.* 10, 1–12. doi: 10.1038/s41467-019-13239-6

Mishra, R., Kim, S., Guzman, S., and Jonas, P. (2016). Symmetric spike timing-dependent plasticity at CA3-CA3 synapses optimizes storage and recall in autoassociative networks. *Nat. Commun.* 7, 1–11. doi: 10.1038/ncomms11552

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* 32, 155–166. doi: 10.1016/j.cobeha.2020.02.017

Morita, K., Morishima, M., Sakai, K., and et al. (2012). Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends Neurosci.* 35, 457–467. doi: 10.1016/j.tins.2012.04.009

Muller, R., Kubie, J., and Saypoff, R. (1991). The hippocampus as a cognitive graph (abridged version). *Hippocampus* 1, 243–246. doi: 10.1002/hipo.450010306

Muller, R., Stead, M., and Pach, J. (1996). The hippocampus as a cognitive graph. *J. Gen. Physiol.* 107, 663–694. doi: 10.1085/jgp.107.6.663

Nevelson, M., and Hasminskii, R. (1976). *Stochastic Approximation and Recursive Estimation*. Providence, RI: American Mathematical Society.

O'Keefe, J., and Burgess, N. (1996). Geometrical determinants of the place fields of hippocampal neurons. *Nature* 381, 425–428. doi: 10.1038/381425a0

O'Neal, T., Bernstein, M., MacDougall, D., and Ferguson, S. (2022). A conditioned place preference for heroin is signaled by increased dopamine and direct pathway activity and decreased indirect pathway activity in the nucleus accumbens. *J. Neurosci.* 42, 2011–2024. doi: 10.1523/JNEUROSCI.1451-21.2021

Pelkey, K., Chittajallu, R., Craig, M., and et al. (2017). Hippocampal GABAergic inhibitory interneurons. *Physiol. Rev.* 97, 1619–1747. doi: 10.1152/physrev.00007.2017

Pfeiffer, B. (2017). The content of hippocampal "replay". *Hippocampus* 30, 6–18. doi: 10.1002/hipo.22824

Pfeiffer, B., and Foster, D. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79. doi: 10.1038/nature12112

Ponulak, F., and Hopfield, J. (2013). Rapid, parallel path planning by propagating wavefronts of spiking neural activity. *Front. Comput. Neurosci.* 7, 98. doi: 10.3389/fncom.2013.00098

Poucet, B., and Hok, V. (2007). Remembering goal locations. *Curr. Opin. Behav. Sci.* 17, 51–56. doi: 10.1016/j.cobeha.2017.06.003

Romani, S., and Tsodyks, M. (2015). Short-term plasticity based network model of place cells dynamics. *Hippocampus* 25, 94–105. doi: 10.1002/hipo.22355

Rosenberg, M., Zhang, T., Perona, P., and Meister, M. (2021). Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *Elife* 10, 1–30. doi: 10.7554/eLife.66175

Russek, E., Momennejad, I., Botvinick, M., Gershman, S., and Daw, N. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* 13, e1005768. doi: 10.1371/journal.pcbi.1005768

Samsonovich, A., and McNaughton, B. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci.* 17, 5900–5920. doi: 10.1523/JNEUROSCI.17-15-05900.1997

Schlingloff, D., Kali, S., Freund, T., and et al. (2014). Mechanisms of sharp wave initiation and ripple generation. *J. Neurosci.* 34, 11385–11398. doi: 10.1523/JNEUROSCI.0867-14.2014

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2016). "High-dimensional continuous control using generalized advantage estimation," in *International Conference on Learning Representations (ICLR)* (San Juan).

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Seijen, H., and Sutton, R. (2014). "True online TD(λ)," in *International Conference on Machine Learning (ICML)* (Beijing).

Shao, L., Halvorsrud, R., Borg-Graham, L., and Storm, J. (1999). The role of BK-type Ca^2+-dependent K^+ channels in spike broadening during repetitive firing in rat hippocampal pyramidal cells. *J. Physiol.* 521, 135–146. doi: 10.1111/j.1469-7793.1999.00135.x

Singh, S., and Sutton, R. (1996). Reinforcement learning with replacing eligibility traces. *Mach. Learn.* 22, 123–158. doi: 10.1007/BF00114726

Sjulson, L., Peyrache, A., Cumpelik, A., Cassataro, D., and Buzsaki, G. (2018). Cocaine place conditioning strengthens location-specific hippocampal coupling to the nucleus accumbens. *Neuron* 98, 926–934. doi: 10.1016/j.neuron.2018.04.015

Skaggs, W., Knierim, J., Kudrimoti, H., and McNaughton, B. (1995). A model of the neural basis of the rat's sense of direction. *Adv. Neural Inf. Process. Syst.* 7, 173–180.

Skaggs, W., and McNaughton, B. (1998). Spatial firing properties of hippocampal CA1 populations in an environment containing two visually identical regions. *J. Neurosci.* 18, 8455–8466. doi: 10.1523/JNEUROSCI.18-20-08455.1998

Sosa, M., Joo, H., and Frank, L. (2020). Dorsal and ventral hippocampal sharp-wave ripples activate distinct nucleus accumbens networks. *Neuron* 105, 725–741. doi: 10.1016/j.neuron.2019.11.022

Spalla, D., Cornacchia, I., and Treves, A. (2021). Continuous attractors for dynamic memories. *Elife* 10, 1–28. doi: 10.7554/eLife.69499

Stachenfeld, K., Botvinick, M., Gershman, S., and et al. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* 20, 1643–1653. doi: 10.1038/nn.4650

Stark, E., Roux, L., Eichler, R., Senzai, Y., and et al. (2014). Pyramidal cell-interneuron interactions underlie hippocampal ripple oscillations. *Neuron* 83, 467–480. doi: 10.1016/j.neuron.2014.06.023

Stella, F., Baracskay, P., O'Neill, J., and Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling Brownian diffusion. *Neuron* 102, 450–461. doi: 10.1016/j.neuron.2019.01.052

Stringer, S., Rolls, E., and Trappenberg, T. (2004). Self-organising continuous attractor networks with multiple activity packets, and the representation of space. *Neural Netw.* 17, 5–27. doi: 10.1016/S0893-6080(03)00210-7

Stringer, S., Rolls, E., Trappenberg, T., and Araujo, I. (2002). Self-organizing continuous attractor networks and path integration: two-dimensional models of place cells. *Netw. Comput. Neural Syst.* 13, 429–446. doi: 10.1088/0954-898X_13_4_301

Sutton, R. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *International Conference on Machine Learning (ICML)* (Austin, TX).

Sutton, R., and Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sutton, R., and Barto, A. (2018). *Reinforcement Learning: An Introduction, 2nd Edn*. Cambridge, MA: MIT Press.

Sutton, R., McAllester, D., Singh, S., and Mansour, Y. (1999). "Policy gradient methods for reinforcement learning with function approximation," in *Neural Information Processing Systems (NeurIPS)* (Denver, CO).

Todorov, E., Erez, T., and Tassa, Y. (2012). "Mujoco: a physics engine for model-based contro,". in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vilamoura).

Tolman, E. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208. doi: 10.1037/h0061626

Tolman, E., and Honzik, C. (1930). "Insight" in rats. *Univer. California Publicat. Psychol.* 4, 215–232.

Trouche, S., Koren, V., Doig, N., and et al. (2019). A hippocampus-accumbens tripartite neuronal motif guides appetitive memory in space. *Cell* 176, 1393–1406. doi: 10.1016/j.cell.2018.12.037

Tsodyks, M. (1999). Attractor neural network models of spatial maps in hippocampus. *Hippocampus* 9, 481–489. doi: 10.1002/(SICI)1098-1063(1999)9:4andlt;481::AID-HIPO14andgt;3.0.CO;2-S

van der Meer, M., and Redish, A. (2011). Theta phase precession in rat ventral striatum links place and reward information. *J. Neurosci.* 31, 2843–2854. doi: 10.1523/JNEUROSCI.4869-10.2011

Wang, X. (1998). Calcium coding and adaptive temporal computation in cortical pyramidal neurons. *J. Neurophysiol.* 79, 1549–1566. doi: 10.1152/jn.1998.79.3.1549

Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* 40, 373–394. doi: 10.1146/annurev-neuro-072116-031109

Widloski, J., and Foster, D. (2022). Flexible rerouting of hippocampal replay sequences around changing barriers in the absence of global place field remapping. *Neuron* 110, 1547–1558. doi: 10.1016/j.neuron.2022.02.002

Williams, R. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* 8, 229–256. doi: 10.1007/BF00992696

Wirtshafter, H., and et al. (2019). Locomotor and hippocampal processing converge in the lateral septum. *Curr. Biol.* 19, 3177–3192. doi: 10.1016/j.cub.2019.07.089

Wu, S., Hamaguchi, K., and Amari, S. (2008). Dynamics and computation of continuous attractors. *Neural Comput.* 20, 994–1025. doi: 10.1162/neco.2008.10-06-378

Xu, H., Baracskay, P., O'Neill, J., and et al. (2019). Assembly responses of hippocampal CA1 place cells predict learned behavior in goal-directed spatial tasks on the radial eight-arm maze. *Neuron* 101, 119–132. doi: 10.1016/j.neuron.2018.11.015

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G., and et al. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* 345, 1616–1620. doi: 10.1126/science.1255514

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* 16, 2112–2126. doi: 10.1523/JNEUROSCI.16-06-02112.1996

Appendix

Pretraining details

In order to train the policy that turns θ° then runs forward, we use the following reward function,

$$r = vel_g + 0.1\vec{h}_g \cdot \vec{h}_r + 0.1\vec{n}_z \cdot \vec{n}_r. \quad \text{(A1)}$$

Here vel_g is the velocity of the rat along the targeted direction. $\vec{h}_g$ is the unit vector along the targeted direction. $\vec{h}_r$ is the unit vector along the body direction of the rat. $\vec{n}_z$ is the unit vector along the z-axis in the global coordinate frame. $\vec{n}_r$ is the unit vector along the z-axis in the local coordinate frame of the rat. To obtain high rewards, the rat requires to turn into the targeted direction and run as fast as possible while keeping $\vec{n}_r$ upward to avoid falling down. The TD3 algorithm trains each policy network from scratch to maximize the accumulated rewards collected from a behavior trajectory. Training each policy takes several hours on a GPU server.

Hyperparameters of the TD3 algorithm

TABLE A1 Hyperparameters and their values.

Hyperparameter	Value
Replay buffer size	100000
Learning rate	0.0001
Exploration coefficient	0.2
Discount factor	0.98
Mini-batch size	256
Number of layers	5
Target net update rate	0.01
Noise standard deviation	0.2
Noise clip coefficient	0.5