Capturing rapid learning in an extended successor representation theory of the cognitive map

Suhee Cho[1,*] **and James L. McClelland**[1,*]

[1]Stanford University, Department of Psychology, Stanford, CA 94305, USA
[*]Correspondence: {suheecho,jlmcc}@stanford.edu

ABSTRACT

Animals sometimes adapt their behavior after a single exposure to novel information. Here, we integrate abstract theory with recent mechanistic insights to understand this and other examples of rapid adaptation. We ground our investigation in studies of animal learning in spatial contexts, where it is widely held that learning depends on the formation of a cognitive map reflected in learned representations in hippocampal areas CA1 and CA3. We extend the successor representation (SR) theory of the cognitive map, incorporating the weighting of this representation by perceived salience (PS) signals, and using it to represent future availability of environmental features that allow behavior to be guided flexibly by an animal's goals and needs at later times. To capture rapid adaptation, we rely on behavioral time-scale synaptic plasticity (BTSP), a form of plasticity that captures one-shot formation of new place fields. Through simulations, we show that our hippocampal model network can learn the PS-weighted SR rapidly and set up the ability to replay experiences in the environment after a single exposure through BTSP. Spontaneous replay activity during subsequent offline periods strengthens the PS-weighted SR. Predictions of environmental features also develop rapidly, and behavioral simulations show how these learned representations give rise to need-dependent choice, rapid behavioral adaptation to changes in the availability of reward, and one-trial shock avoidance learning.

Introduction

Animals' behavior often adapts quickly to experience. How is this rapid adaptation achieved? This question is important for both neuroscience and artificial intelligence (AI). Many existing models in both computational neuroscience and AI can capture aspects of learned human and animal abilities, but require orders of magnitude more experience to learn these abilities than humans or animals do. Understanding how the brain can learn so efficiently would therefore address a key limitation in our current understanding of learning both in ourselves and our machines.

Here, we ground our exploration of this question within the context of studies of animals behaving in spatial environments (Fig. 1a). In such settings, animals can show very rapid adaptation, often altering their behavior after a single exposure to a salient environmental feature[1]. We further situate our work in the framework of the "cognitive map," a hypothetical construct that receives locally available information, integrates it with internal state information, and generates signals that guide behavior in spatial environments[2–4]. Recording studies in the hippocampus of behaving animals revealed that the existence of place cells collectively cover all of the locations within a spatial arena[5], such that they can be thought of as a central component of a cognitive map[6]. Furthermore, experiments show that rodents form such a map rapidly, with many new place fields forming during an initial traversal of a novel environment[7]. Our work aims to provide an integrative framework for understanding how this rapid and efficient representation learning comes about and how it then influences behavior.

We build on the seminal account of the cognitive map provided by successor representation (SR) theory[8] and its application to hippocampal representation learning[9]. Rather than encoding environments purely in terms of spatial distance, the SR encodes the expected future occupancy of each location based on the animal's movement history (Fig. 1b, top). Although this approach is supported by experimentally observed properties of hippocampal place cells, including the predictive skewing of place fields[10,11], the original version of this theory[9] has two important limitations.

First, as originally described[9], this account does not explain how factors such as reward[12], aversive outcomes[1], or novelty[13] affect hippocampal representations. These and other factors, collectively termed perceived salience (PS)[14], can modulate hippocampal representations, including increases in the number of place cells encoding salient locations[12,15–18] and accelerated formation of those fields[19]. Together, these findings suggest that the hippocampus integrates PS signals with spatial inputs, producing a PS-weighted SR.

Second, the original SR theory does not address how rapid representational adaptation is achieved. This may reflect the framing of the approach as a normative or computational level[20] account, where the priority has been to capture emergent functional characteristics that the SR enables rather than speed of learning *per se*, and in part from the reliance on learning rules available at the time the work was done. In any case, past simulations showing how the SR may be acquired[9,21] typically rely

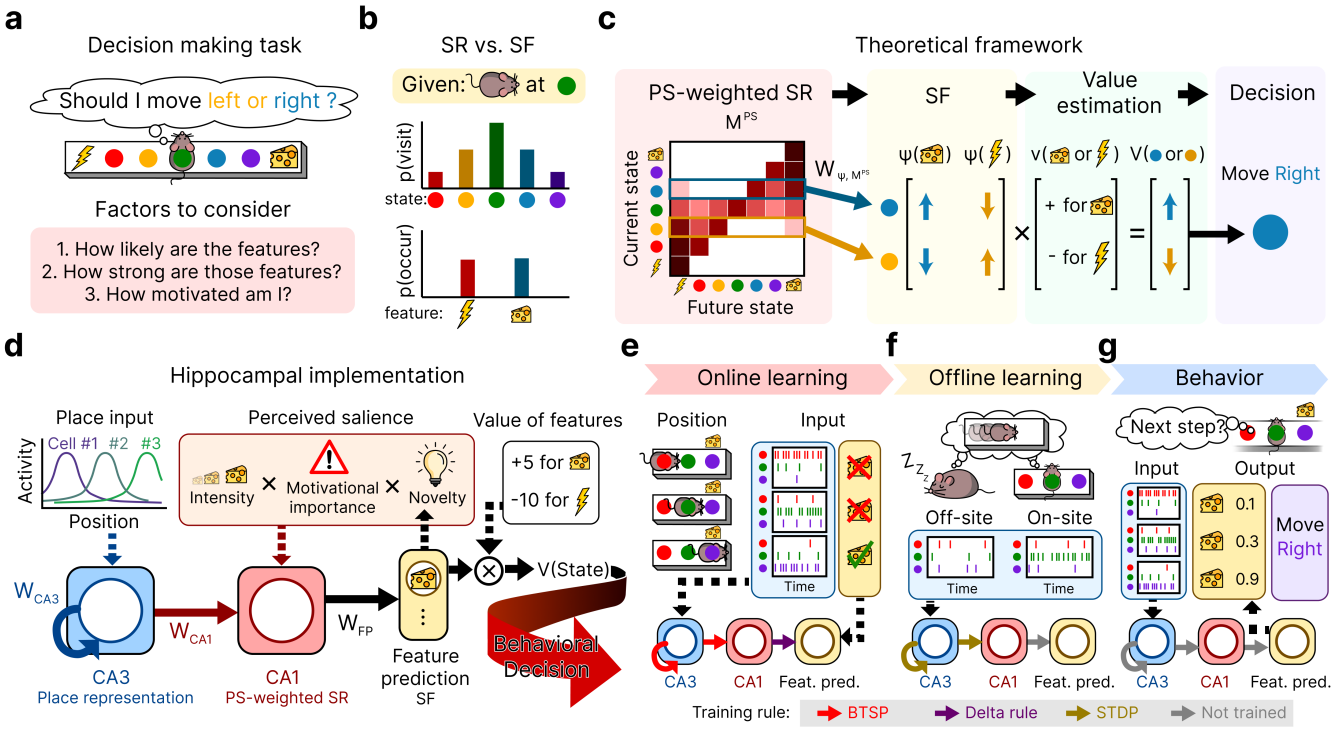


Figure 1. **a**, Rodent decision-making in a maze. **b**, Two ways of representing a state: successor representation (SR) and successor features (SF). **c**, Schematic of the theoretical framework. A PS–weighted SR is relayed to predict successor features, which are then used to compute value and guide decisions. **d**, Hippocampal implementation of the model. CA3 receives place input and encodes place representations. CA1 receives input from CA3 along with P, and encodes the PS-weighted SR. This is passed to the feature prediction layer, which encodes SFs and is ultimately used to determine the subjective value of states. **e–g**, Three distinct phases of our simulations. **e**, *Online learning*: the animal actively explores the environment and encounters features. Place input varies with position, and hippocampal weights are updated using the BTSP rule, while the feature-prediction weights are trained using an error-correcting delta rule. **f**, *Offline learning*: simulating rest or sleep. Place input is held constant, and hippocampal weights are updated using symmetric STDP while the feature-prediction weights are fixed. **g**, *Behavior simulation*: Value-based decision making in the learned track. No weights are updated. The learned connection weights are used to calculate feature predictions, which are assigned value based on current motivational state to guide behavior toward higher-valued outcomes.

on reinforcement-learning-theory based learning rules[22] or longstanding biologically proposed plasticity rules[23], both of which can require many (up to thousands) of traversals through an environment to learn.

The discovery of behavioral time-scale synaptic plasticity (BTSP) provides a candidate mechanism that makes it possible to work toward treating rapid adaptation as an important functional characteristic to be addressed. First described nearly a decade ago[12], BTSP occurs when neurons undergo dendritic plateau potentials that produce large synaptic changes linking neural activity across a time scale of seconds, leading to one-trial formation of new place fields[12,24]. BTSP is coming to be recognized as a potentially viable basis for rapid neural adaptation[25,26], capturing formation and change of place fields better than alternative spike-time dependent learning rules, which are too weak to capture these changes within realistic parameter ranges[27]. In hippocampal area CA1, BTSP creates place fields that exhibit SR-like anticipatory predictions of future states[12] and has recently been used to model rapid formation of the classical (non–PS-weighted) SR and reward-directed navigation[28]. Here, we extend this approach to incorporate PS-weighting of predictive representations in CA1, consistent with findings that BTSP in this region is modulated by novelty, stimulus intensity, and reward[19,29–31].

Our work further highlights how BTSP can initiate a virtuous cycle of adaptation by enabling hippocampal replay after a single exposure to a novel environment or a motivationally salient event. Hippocampal replay was first observed during periods of sleep[32] and identified as occurring during bursts of activity during which hippocampal place cells spontaneously reactivate in patterns that recapitulate trajectories through an environment[33]. This phenomenon has long been thought to contribute to the strengthening, stabilization, and/or consolidation of learning acquired during physical exploration of the environment[34–37]. More recent studies show evidence of replay during waking behavior, sometimes while animals are at rest after consuming a reward[38], and sometimes during or prior to exploration[1,39]. We rely on BTSP in hippocampal area CA3 to rapidly create connections among CA3 place cells to enable replay events that occur in subsequent offline periods[27,40,41]. Remarkably, these sequences can represent trajectories that the animal has never physically traversed[42], suggesting that replay helps complete the internal spatial map[7,43], and we demonstrate how this can occur through further synaptic plasticity thought to occur during replay. In this way, replay works synergistically with BTSP, which forms the synaptic connections that allows replay to occur

after as little as a single exposure to novel information, helping to extend new learning without requiring additional exposure to environmental contingencies.

Beyond incorporating PS weighting into the theory and using BTSP as the key learning mechanism, we also extend the way the SR theory links the cognitive map to environmental features. The original SR theory links environmental states (places in an environment) to a simple scalar estimate of value based on reward, limiting its generality and adaptive flexibility[44], since value of an external stimulus can vary with the animal's needs and transient goals. For example, it would be adaptive to seek states associated with water when thirsty and to seek states associated with food when hungry. More generally, it would be a powerful extension of the SR theory to be able to use a SR to determine how to navigate toward or away from arbitrarily specified environmental features when they become goal relevant. To address this, we incorporate successor features (SFs) previously proposed in a machine learning context[45] (another recent extension of the SR theory[46] incorporates SFs to address issues different from those considered here). Unlike the SR, which predicts the *future occupancy of states*, SFs predict *the future occurrence of features* (Fig. 1b, bottom) that occur in these states. We propose that the PS-weighted SR is used to predict SFs in brain areas downstream from the hippocampus. As we shall see below, a simple error-correcting learning rule for learning feature predictions based on the PS-weighted SR speeds learning to predict salient features relative to others with lesser salience. The resulting feature predictions can then be assigned value based on the animal's motivational state (or, more generally, its transient goals) during later behavioral episodes, allowing behavior to be guided toward places with current motivation-relevant properties.

Combining these components, we propose an extended integrative framework for the cognitive map, its neural implementation, and its consequences for rapid adaptation of neural activity and behavior (Fig. 1c-g). In this framework (Fig. 1c) the hippocampus integrates spatial input with PS signals to form a PS-weighted SR, which is then conveyed to downstream areas that compute SFs and combine them with motivational signals to estimate state values (Fig. 1d). To demonstrate how this representation can emerge during behavior, further strengthen during offline periods, and support rapid behavioral adaptation, our simulations employ three distinct phases: an online exploration phase, during which BTSP updates hippocampal connectivity (Fig. 1e); an offline replay phase, during which spontaneous replays propagate predictive structure and further enhance learning through synaptic plasticity (Fig. 1f); and a behavioral phase in which the learned predictive map supports flexible value computation and guides behavior (Fig. 1g).

Results

Formal statement of theory

Following the proposal of Dayan[8], the SR for a starting state s and a target state s' is the expected discounted future occupancy of s' across time t, given $s_0 = s$:

$$M(s,s') \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{I}(s_t = s') \,\middle|\, s_0 = s\right], \tag{1}$$

where $\gamma \in [0,1]$ is a discount factor. Thus $M(s,s')$ weights nearer states more heavily than distant ones. The Matrix M of current-state, future-state pairs summarizes the SR for a given environment.

The SR can be extended to characterize successor features, which correspond to discounted future predictions of environmental features (e.g., cues, rewards, or hazards)[45]. The SF estimate $\psi(s,f)$ for a feature f at start state s is defined as the expected discounted future occurrence of that feature:

$$\psi(s,f) \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, \phi(s_t,f) \,\middle|\, s_0 = s\right] \tag{2}$$

Which is simply the future occupancy of s' times the estimated probability $\phi(s',f)$ of the feature's occurrence at state s':

$$\psi(s,f) \;=\; \sum_{s'} M(s,s')\,\phi(s',f). \tag{3}$$

We next incorporate perceived salience (PS) into our framework[14,44]. The PS of state s, $\omega(s)$, is a positive quantity that depends on the sum, across features of a state, of the PS of each feature of the state, which in turn depends on the intensity, novelty, and motivational significance of each feature. We can then define the PS-weighted SR,

$$M^{\mathrm{PS}}(s,s') = \omega(s')\,M(s,s'). \tag{4}$$

With this scaling, the SF representation $\psi(s,f)$ of feature f in state s can be computed by a linear mapping $W_{\psi,M^{\mathrm{PS}}}(s',f)$ from M^{PS}:

$$\psi(s,f) = \sum_{s'} M^{\mathrm{PS}}(s,s') \cdot W_{\psi,M^{\mathrm{PS}}}(s',f), \tag{5}$$

Note that this equation would compute the same successor feature probabilities as Equation 2 if the mapping weight $W_{\psi,M^{\mathrm{PS}}}(s',f)$ of PS-weighted state occupancy probabilities to feature predictions is set equal to $\phi(s',f)$ divided by $\omega(s')$:

$$W_{\psi,M^{\mathrm{PS}}}(s',f) = \frac{\phi(s',f)}{\omega(s')}. \tag{6}$$

In our proposed implementation, the learning rule discussed below for estimating $W_{\psi,M^{\mathrm{PS}}}(s',f)$ implicitly performs this division, while allowing predictions of features of salient states to be learned more quickly.

Finally, the value of state s at at some time t is defined as the sum of successor feature probabilities at s weighted by their value at the time of assessment

$$V(s,t) = \sum_{f} \psi(s,f)\, v(f,t). \tag{7}$$

This allows the same predictive representation ψ to flexibly support different decisions, depending on the agent's assessed value of different features when making decisions.

Linking the theory to its neural implementation and behavior

We link these theoretical quantities to neural processes and behavior as follows:

- An estimate of the **PS-weighted SR**, $M^{\mathrm{PS}}(s,s')$, is acquired through BTSP in simulated neural populations representing hippocampal areas CA3 and CA1 during exploration and is further reinforced through replay-driven plasticity.
- **Perceived salience**, $\omega(s)$, thought to be computed outside of the hippocampus[14], is computed as a simple function on the intensity, novelty, and motivational significance of features present at state s and is incorporated into the CA1 representation by modulating BTSP.
- The **mapping from PS-weighted SR to SF**, $W_{\psi,M^{\mathrm{PS}}}$, is instantiated by learned synaptic weights from the CA1 area to downstream feature-encoding populations. In our implementation, the learning rule is a simple error correcting learning rule[47]. As we discuss below (see Fig. 3), the up-weighting of the neural representation of the successor representation by PS speeds up the learning of features associated with highly salient states, contributing to rapid adaptation, while error correction prevents over-prediction.
- **State value**, $V(s)$, is thought to be computed in extra-hippocampal circuits integrating feature predictions with feature-specific value signals $v(f)$, which we treat as scalar variables that can vary depending on the animal's needs and goals at different times.
- The resulting $V(s)$ guides **decision making**: when selecting actions, we simulate behavior by assuming animals preferentially transition to states with larger $V(s)$.

In the following sections, we modeled neural activity and behavior after simulated exposure to environments paralleling those used in published experiments. We used distinct simulation paradigms for the online exploration, offline replay, and behavior phases of these simulations (Fig. 1e–g) as described more fully in *Methods*.

Representations in CA3 and CA1 regions emerge rapidly through BTSP, supporting replay and prediction of successor features

We extended a spiking neural network modeling hippocampal areas CA3 and CA1[48] and added circuitry to learn SF representations. The model includes a CA3-like population of 3200 simulated neurons with recurrent weights W_{CA3} and a CA1-like population of the same size with feedforward weights W_{CA1} from CA3 to CA1. As in previous work[48], each CA3 neuron receives place-specific input in the form of a pre-specified Gaussian bump, with the centers of the bumps evenly covering a simulated linear track. On the other hand, CA1 neurons only received input via W_{CA1} from CA3. We added a downstream "feature-prediction" population of neuron-like units, each of which was assigned to correspond to a specific environmental

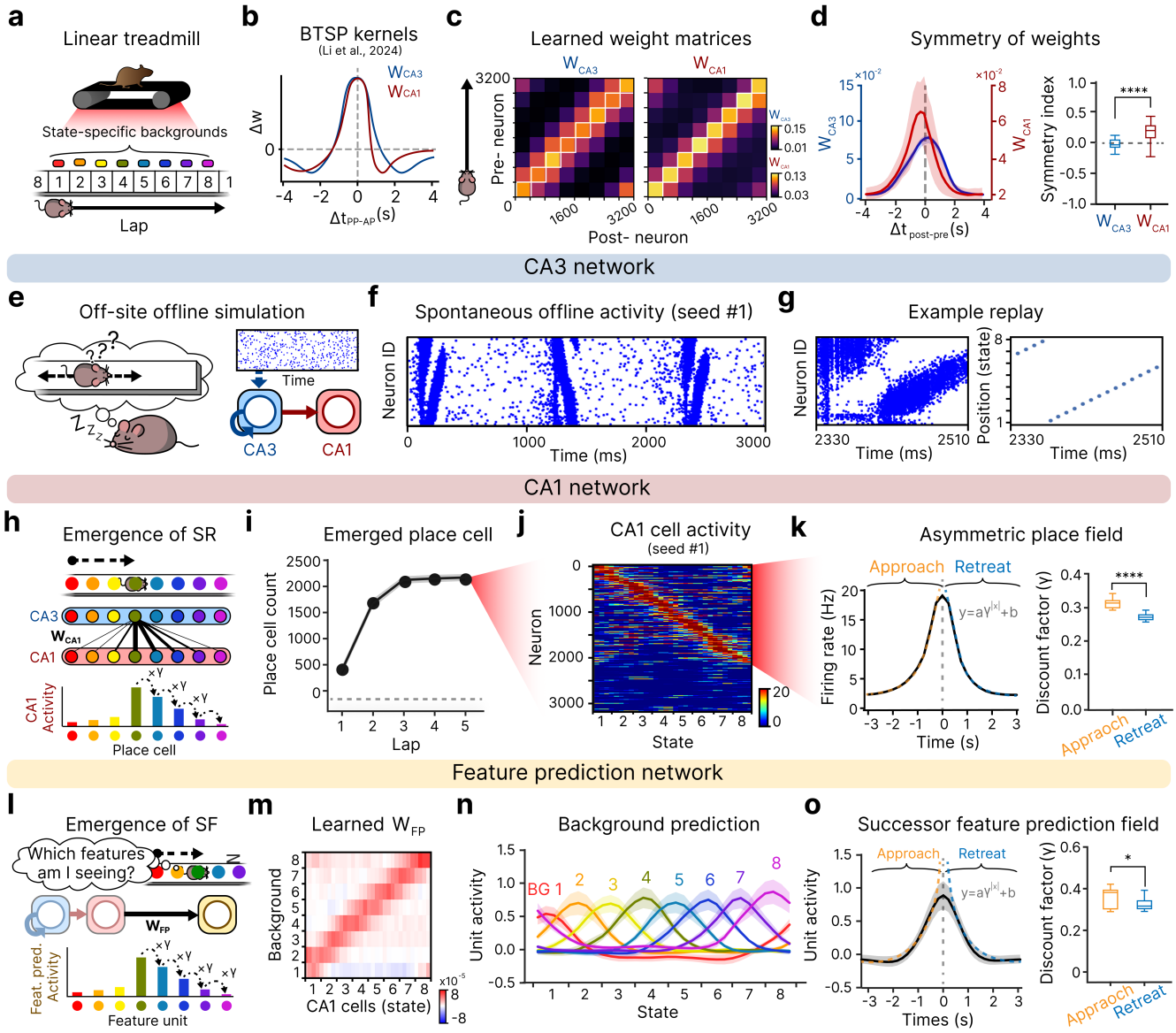


Figure 2. **a**, Linear treadmill simulation setup. **b**, BTSP weight update kernels for CA3 and CA1. **c**, Weight matrices after lap 5 with place cells sorted by field location, binned into 8 groups and averaged. Left: W_{CA3}. Right: W_{CA1}. **d**, Left: Averaged synaptic weights of postsynaptic neurons aligned by presynaptic index, for W_{CA3} (blue) and W_{CA1} (red). The curve is divided into weights from presynaptic neurons with smaller indices ("behind") and larger indices ("ahead"). Scales for curves are chosen to aid visualization of differences in shape. Right: Symmetry index of W_{CA3} and W_{CA1}, calculated by normalizing the area of behind vs. ahead weights. Paired t-test, $N_{seed} = 10$, ****$P = 4.58 \times 10^{-36}$. **e–g**, Results from CA3. **e**, Offline simulation schematic. **f**, Example spontaneously generated offline activity. **g**, Left: Example replay activity. Right: Decoded position. **h–k**, Results from CA1. **h**, SR emerges in CA1 due to asymmetry of W_{CA1}. **i**, Number of CA1 place cells formed after each lap. **j**, CA1 place fields after lap 5. **k**, Left: Average firing rate across place fields. Right: Fitted discount factors for approach vs. retreat trajectories relative to field centers. Paired t-test, $N_{seed} = 10$, ****$P = 7.02 \times 10^{-5}$. **l–o**, Results from the feature prediction layer. **l**, Learning W_{FP} to predict SFs. **m**, W_{FP} after lap 5. **n**, Predictions for each background feature across positions after lap 5. **o**, Average predicted presence for the current state vs. non-current states. Paired t-test, $N_{seed} = 10$, ***$P = 8.13 \times 10^{-4}$. All data are presented as mean±s.d.

feature, with weights W_{FP} projecting to it from CA1. We divided the linear treadmill into eight consecutive states (states 1–8), each of which was associated with a unique neutral "background" feature to distinguish spatial locations (Fig. 2a).

We modeled BTSP-driven online learning through 5 contiguous laps of navigation— uniform movement through the assigned CA3 place field centers. During the online learning, W_{CA3} was trained using the BTSP learning rule with a symmetric time kernel[12,31], whereas W_{CA1} was trained using an asymmetric kernel (Fig. 2b)[40]. W_{FP} was trained using an error correcting delta rule[47] to predict the presence of the eight background features associated with the track states (Fig. 2l).

We first examined the learning of W_{CA3} and W_{CA1}. As expected, the distinction in learning kernels resulted in different structures in these connections. To illustrate this, we separately identified CA3 and CA1 neurons by rank-ordering them according to the position at which each neuron exhibited its peak firing rate, from left to right along the track. Then, we plotted W_{CA3} and W_{CA1} after lap 5, arranging both pre-synaptic and post synaptic neurons by their IDs (Fig. 2c). The results showed that W_{CA3} encoded spatial relationships between place fields in an allocentric, movement direction-independent manner, resulting in a symmetric structure (Fig. 2d, blue). In contrast, W_{CA1} displayed a skewed distribution: connections from place cells representing positions behind the animal's current location were more potentiated than those ahead (Fig. 2d, red).

Next, we examined the functional consequences that ensued from these distinct representations. We first showed that online learning through the symmetric BTSP rule enabled the CA3 network to spontaneously generate replay-like activity during subsequent offline periods. To test this, we simulated the offline phase by delivering a uniform, non-spatial input for three seconds to W_{CA3} after training (Fig. 2e). We simulated the resulting spontaneous activity using biophysically plausible parameters and an adaptive exponential integrate-and-fire neuron model (Supplementary Materials, Section 3)[48].

During simulated replay, the CA3 neurons fired in sequence according to their place-field locations (Fig. 2f). These sequential events were best decoded as a linear traversal along the track and therefore qualify as replay events[49] (Fig. 2g). The replay activity began to emerge after a single track traversal and was robust after 5 laps of exploration (few-shot learning)[38].

Next, we showed that the asymmetric structure of W_{CA1} endows CA1 neurons with SR-like place fields (Fig. 2h) after 5 laps of learning from random initial weights. We classified a CA1 neuron as a place cell if its peak firing rate exceeded 15 Hz (75% of the model's specified maximum of 20 Hz) at any position along the track (see *Methods* for details). We found that an average of 2,148 CA1 place cells had emerged (Fig. 2i). The resulting place fields were distributed across the environment (Fig. 2j). As expected, these place fields were asymmetric: neurons began firing earlier as the animal approached the field's peak location and shut off more rapidly as it moved away (Fig. 2k, left). This was assessed by modeling the place cell activity during the "approach" and "retreat" from the center of the receptive field using the function $y = a \times \gamma^{|x|} + b$ (Fig. 2k, right), where the larger discount factor γ during approach compared to retreat reflects the expected asymmetry.

Finally, we demonstrated that the feature prediction network successfully learned to anticipate environmental features (Fig. 2l). As expected, by the end of the fifth lap, W_{FP} had become state-specific, with CA1 place cells preferentially projecting to the feature unit associated with their place field (Fig. 2m). Consequently, each background feature unit exhibited strongest prediction when the animal occupied its corresponding location (Fig. 2n), along with anticipatory activation similar to that seen in the CA1 place fields (Fig. 2o).

Salient features are rapidly encoded and updated in CA1, enhancing their prediction in the feature prediction network

Next, we demonstrated that the SR encoded within the CA1 layer in our model reflects the introduction and relocation of reward[12,16,19], one type of salient outcome, leading to rapid adaptation of place fields and outcome predictions. We introduced a food reward as a salient outcome after the five familiarization laps without reward described above (Fig. 2). The food was placed at state 4 for the next five laps (laps 6-10), and then relocated to state 8 for five additional laps (laps 11-15) (Fig. 3a). We treated the introduction of reward as affecting two key factors (Fig. 3b). First, the motivational importance of food was set to 3, compared with 1 for background features. This significantly increased the PS of the reward location, modulating the place field formation in CA1 (Fig. 3c). Second, the animal's travel speed was halved when passing through the reward state, reflecting slowing during reward consumption[12,19].

In accordance with experimental findings and existing modeling work[12,31,40] our model incorporates PS as a modulator of plateau potentials in CA1 and not in CA3. As expected, then, W_{CA3} was not meaningfully altered by reward relocation (Fig. 3d), while the predictive map encoded within W_{CA1} captured the PS information introduced by the reward. Indeed, W_{CA1} adapted rapidly to changes in reward location (Fig. 3e). After 5 laps with the reward in state 4, connections in W_{CA1} to CA1 neurons in locations near state 4 were strengthened, and after 5 laps with reward at state 8, connections W_{CA1} projecting to CA1 neurons representing this new reward location increased sharply. This adaptive reweighting arises from the modulation of BTSP in CA1 by perceived salience (Supplementary Fig. 3). As a result, the CA1 neurons were adaptively remapped to over-represent the vicinity of the reward (Fig. 3f and g), in accordance with experiments[15,16,31]. Notably, CA1 population activity at a rewarded state increased rapidly following a novel reward encounter (Fig. 3i), and remained approximately stable thereafter. This contrasts with the slower initial buildup at these and other sites when the have only non-salient background

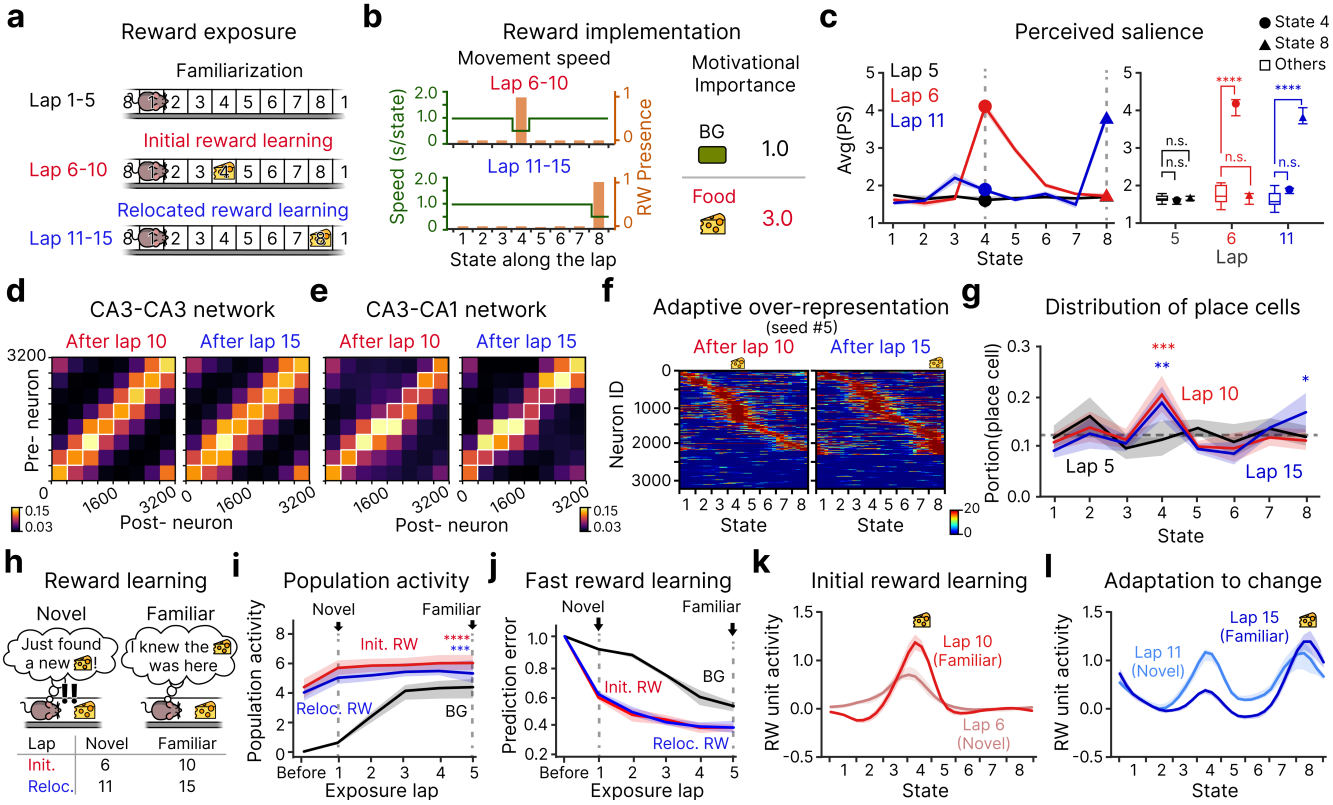


Figure 3. **a**, Simulation scheme for introducing a salient reward. **b**, Two factors modulated by reward introduction: reduced movement speed and increased motivational importance. **c**, Left: Average PS across states during laps 5, 6, and 11. The rewarded state exhibits elevated PS when the reward is present. Right: PS at states 4, 5, and all other states during laps 5, 6, and 11. Paired t-test, $N_{\text{seed}} = 10$, n.s. $P > 0.05$, ***$P < 0.005$. **d**, W_{CA3} after lap 10 (initial reward introduction) and lap 15 (reward relocation). Values are averaged across 10 random seeds. **e**, W_{CA1} after lap 10 and lap 15, averaged across 10 random seeds. **f**, CA1 place cell activity after lap 10 and lap 15. Place fields cluster around reward locations and adapt following reward introduction and relocation. **g**, Proportion of CA1 place cells representing each state after lap 5, lap 10, and lap 15. The dashed line indicates the expected proportion (1/8) for a uniform distribution. One-sample one-tailed t-tests comparing mean $< 1/8$, $N_{\text{seed}} = 10$; *$P < 0.05$, **$P < 0.01$, ***$P < 0.005$; unmarked comparisons are not significant with $P > 0.05$. **h**, Two cognitive states associated with reward perception. A reward is initially experienced as novel (left) and becomes familiar after repeated exposures at the same location (right). **i**, Development of CA1 population activity at background and reward-related states as the animal becomes familiar with environmental features. Population activity was computed as the mean firing rate across all CA1 neurons. The black line shows activity at state 4 from the initial network through lap 5, the red line shows activity at state 4 from lap 5 to lap 10 (reward present), and the blue line shows activity at state 8 from lap 10 to lap 15 (after reward relocation). Paired t-test, $N_{\text{seed}} = 10$, ****: CA1 population activity at state 4 at at lap 10 vs. CA1 population activity at lap 5, $P = 5.70 \times 10^{-7}$, ***: CA1 population activity at state 8 at at lap 15 vs. CA1 population activity at lap 5, $P = 0.004$. **j**, Prediction error for background features and the reward feature across laps. Background features require more laps to learn than the reward feature. **k**, Reward learning following initial introduction at state 4. **l**, Reward learning following relocation to state 8. All data are presented as mean ± s.d.

features.

We next considered changes in reward prediction, comparing predictions immediately following reward introduction or relocation ('novel', Fig. 3h, left) to predictions after repeated exposure ('familiar', Fig. 3h, right). Prediction error for the reward decreased much more rapidly than initial learning of background features (Fig. 3j). Furthermore, immediately after the reward was introduced at state 4, the network started to predict the reward at the reward location and to anticipate it at the preceding state (Novel; Fig. 3k). This prediction was gradually strengthened with repeated exposure to the reward at the same location (Familiar; Fig. 3k). After reward relocation to state 8, the feature-prediction layer immediately began to predict the reward at the new location (Novel; Fig. 3l). CA1 retained residual over-representation at the former reward site (Fig. 3f, right), and prediction at the old site decayed gradually with further experience (Familiar; Fig. 3l). This asymmetric adaptation reflects rapid acquisition of new salient associations[1], alongside slower decay of memories about the previous location as supported by behavioral findings[50,51].

Learning to seek rewards relevant to current motivational state

We next explored how our framework supports using learned feature predictions to obtain rewards, highlighting use of these predictions to seek rewards relevant to an animal's motivational state at the time when it is behaving. Using a setup similar to one used in an experiment demonstrating state-dependent reward seeking[52], we trained a CA3-CA1 network with 4000 simulated neurons in each area on a two-dimensional T-maze with two distinct arms divided into 10 states. During the exploration phase,

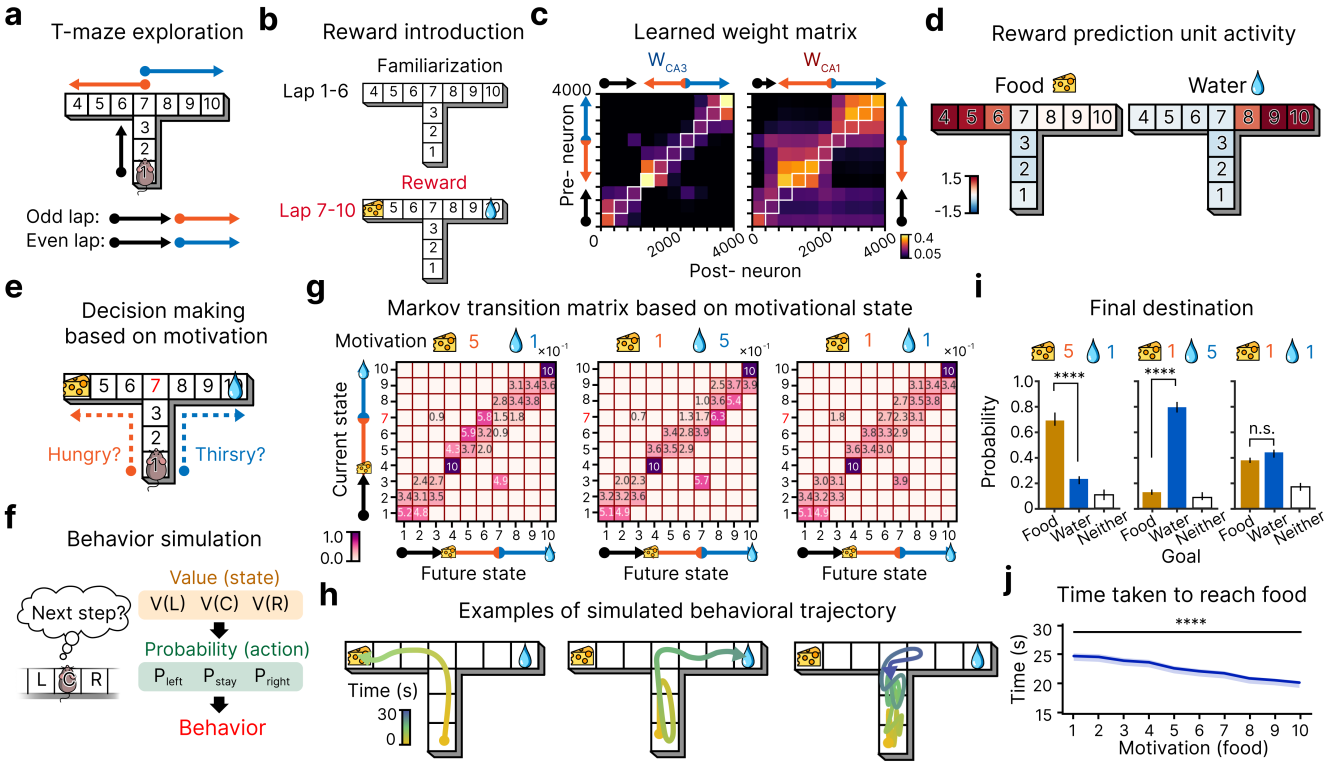


Figure 4. **a**, T-maze learning scheme. **b**, After 6 familiarization laps, two distinct rewards (food and water) are introduced at each arm for 4 laps. **c**, Resulting W_{CA3} and W_{CA1} after lap 10. **d**, Predictions for food and water across maze positions. **e**, Animal's decision to pursue either arm of the maze based on its motivational state. **f**, Behavioral decision-making simulation scheme. The choice-making point (state 7) is marked red. **g**, Markov state transition matrices under three different motivational scenarios for food and water. Transition probabilities at state 7 shows most prominently the role of motivational state and therefore marked red. **h**, Example simulated trajectories for each scenario. **i**, Goal reached within 50 seconds of movement for each scenario. Paired t-test, $N_{seed} = 10$, ****$P < 0.0005$, n.s. $P = 0.98$. Data are presented as mean±s.d. **j**, Time to reach the goal as a function of motivational strength. Mann-Kendall test, $N_{seed} = 10$, ****$P = 8.30 \times 10^{-5}$, $Z = -3.94$. Data are presented as mean±s.e.m.

we simulated alternating visits to each arm, starting from the base of the stem (state 1), with odd-numbered laps ending at the end of the left arm (state 4), and even laps ending at the end of the right arm (state 10) (Fig. 4a). After six laps (three exposures to each arm), distinct reward features (food and water) with motivational significance of 3 were introduced at the arm ends, and the animal was exposed to each reward twice over the next 4 laps (7–10) (Fig. 4b).

After lap 10, W_{CA3} encoded the structural layout of the T-maze (Fig. 4c, left), while W_{CA1} over-represented both arms, reflecting the presence of food and water (Fig. 4c, right). Correspondingly, the feature prediction layer strongly predicted each reward at its respective location (Fig. 4d).

We then simulated behavior based on the resulting learned map under different motivational needs for food and water (Fig. 4e). This simulation relied on value estimates $V(s)$ for each maze state assigned in accordance with the formal theory (Equations 2-6). As described in detail in *Methods*, these estimates were obtained by projecting the center of each state through the Gaussian projection to produce an activity pattern in CA3 and propagating this through W_{CA1} and W_{FP} to obtain feature predictions $\psi(s, f)$ and assigning the final value based on the animal's motivational state in accordance with Equation 6. These value estimates were used to assign next-state transition probabilities from each state to the set of state accessible from it using a softmax function (Fig. 4f). In each simulated behavioral episode, the animal started at the stem (state 1) and updated its location across a series of steps at a notional rate of one step per second based on the resulting Markov transition matrix. Each behavioral trial lasted up to 60 steps or until either goal state was reached.

We simulated three motivational conditions: (1) hungry (food need = 5, water need = 1), (2) thirsty (food need = 1, water need = 5), and (3) neutral (food need = 1, water need = 1). For each condition, we computed the Markov transition matrix containing the movement probabilities (Fig. 4g) as described above. The three matrices notably differed in the state transition probabilities from state 7 (row index highlighted in red) which is the choice point in the maze. When hungry, the animal was more likely to move toward the food arm (left), whereas when thirsty it preferred the water arm (right). Under neutral conditions, movements left, right, or back to the stem were roughly balanced. Accordingly, the animal predominantly reached the food when hungry, the water when thirsty, and distributed its choices more evenly under neutral conditions (Fig. 4h-i). Finally, we found that stronger motivational drives accelerated goal-directed behavior: as the motivational need for food increased from 1

to 10 (with water fixed at 1), the time required to reach the food decreased (Fig. 4j).

Offline replay propagates perceived salience information to unexperienced paths, enhancing reward seeking from novel starting locations

We next investigated how learning during offline replay episodes can enhance the network's prediction of salient outcomes and support behavior, including finding novel paths to reward, as observed in the seminal experiments of Tolman[2]. We used network weights learned through BTSP after exposure to food and water reward as described above (Fig. 4). Specifically, we hypothesized that replay activity propagates SF information across the maze, including to positions that were never visited in succession during online exploration[42] (Fig. 5a). Accordingly, we simulated an offline phase corresponding to a post-reward pause at the food location by providing place input for state 4 to CA3 cells, then observed the resulting network activity. During this phase, replay spontaneously emerged (Fig. 5b, top). Strikingly, the replay events we observed propagated not only along previously experienced trajectories (left- and right-arm replays), but also exhibited a novel "shortcut replay" that had never been traversed, connecting the two arms of the maze (Fig. 5b).

We then examined how each replay type can contribute to propagating feature predictive information from the reward location to distant states. During replay, we used a temporally symmetric-spike-time dependent learning rule[48] that adjusts synaptic weights after every spike with smaller magnitude and a narrower temporal window than BTSP (Fig. 5c).

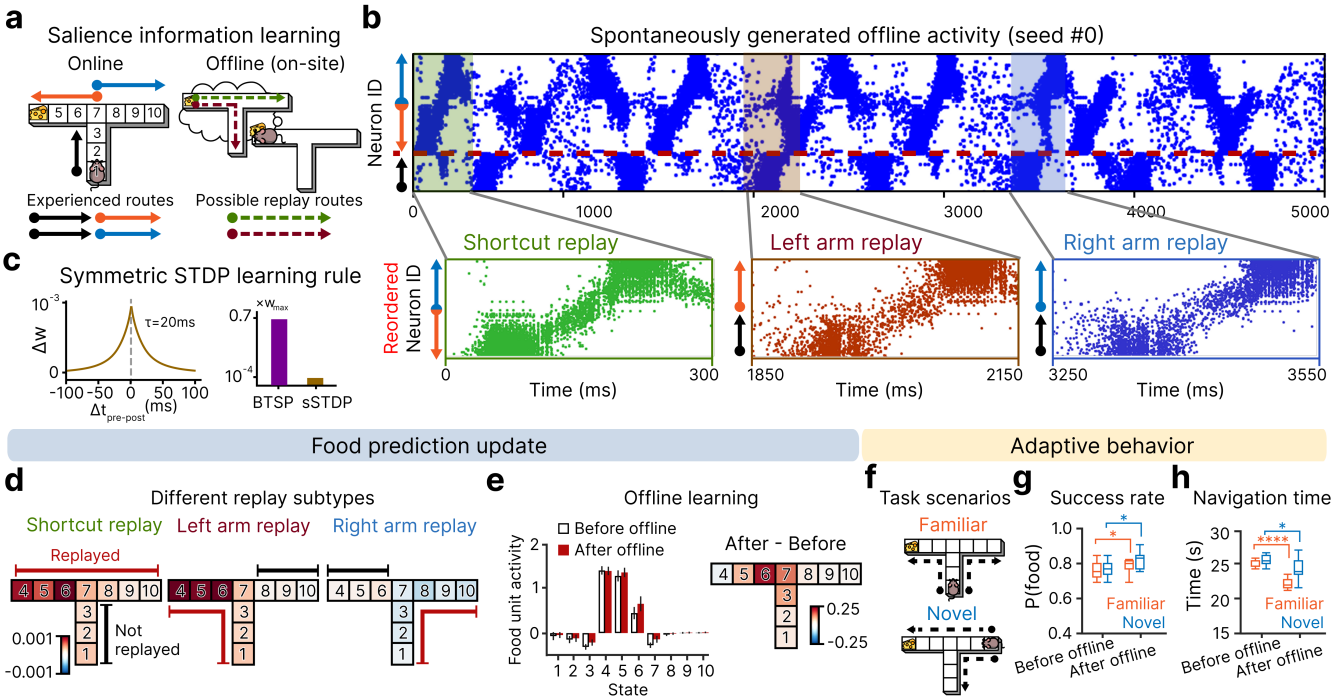


Figure 5. **a**, Propagation of PS information through online and offline learning. **b**, Spontaneously generated offline activity. Replay events are grouped into three categories based on their trajectories: shortcut, left-arm, and right-arm. For visualization, each expanded example shows only the neurons participating in the corresponding replay, reordered in the sequence in which they were activated along the replay trajectory. **c**, Left: sSTDP rule used in offline learning. Right: Largest possible weight update of BTSP and sSTDP rules. **d**, State-wise food prediction updates induced by a single replay event, shown separately for each replay subtype. Updates were computed as the difference between post-replay and pre-replay food predictions. **e**, State-wise food prediction updates induced by 5 seconds of offline learning. Left: Food unit activity at each state before (pre) and after (post) offline learning. Right: State-wise prediction updates computed as the difference between post- and pre-offline learning activity. **f**, Two behavioral simulation schemes. Left: Simulation starting from the same start position as during online exploration. Right: Simulation starting from a different start position. Both tasks aim to find food located at the same position as in the online exploration. **g**, Proportion of trials in which the animal reached the goal location within 50 seconds in familiar and novel scenarios, without and with offline learning. Paired t-test, $N_{seed} = 10$, Familiar $^*P = 0.03$; Novel $^*P = 0.02$. **h**, Time taken for the animal to reach the goal location in familiar and novel scenarios, without and with offline learning. Paired t-test, $N_{seed} = 10$, Familiar $^{****}P = 1.15 \times 10^{-5}$; Novel $^*P = 0.04$. All data are presented as mean±s.d.

To identify effects of distinct types of replay events, we determined the change in food prediction unit activity at each state, averaged over all replays of the same type observed during the 5-second replay period. Following shortcut replay events, food prediction increased across both arms of the maze. In particular, this update was most prominent at states proximal to the food location (states 4-7; Fig. 5d, left). Similarly, left-arm replay enhanced food prediction along the replayed trajectory (Fig. 5d, middle). In contrast, right-arm replay had minimal impact, as its trajectory does not include the food location and therefore does not directly strengthen synapses originating from food-related place cells (Fig. 5d, right). Finally, we updated W_{CA1} using the full spike trains recorded during the offline simulation containing multiple replay events of different subtypes. This offline

update increased food prediction in reward-adjacent states, particularly those that had not previously strongly predicted food (Fig. 5e).

We then simulated two behavioral task scenarios. In the first, the animal started at the stem (state 1) with a motivational need of 5 (Fig. 5f, top), as during the online exploration phase (Fig. 4). Each trial ended when the animal reached either end of the maze (state 4 or state 10) or failed to reach a goal within 50 steps. In the second scenario, the animal started from a novel location (state 10) (Fig. 5f, bottom). Each trial ended when the animal reached either end of the maze (state 4 or state 1) or failed to reach a goal within 50 steps.

Across 500 trials per random seed, incorporating offline learning significantly increased the probability of reaching the food goal in both scenarios (Fig. 5g). Moreover, among successful trials, offline learning reduced the time required to reach the food in both cases (Fig. 5h), indicating that replay-driven updates not only improved success rates but also enhanced behavioral efficiency.

The model learns aversive outcomes one-shot, predicts their presence through forward rollouts, and exhibits one-trial avoidance learning

We next asked whether the model could account for one trial avoidance learning, as observed in a study in which mice freely explored a linear track before encountering a footshock at the track end, after which they were promptly removed from the environment and placed in a rest box outside the linear track[1]. This differs from reward learning (Fig. 3–5) because animals rapidly move away from the shock, limiting the opportunity to associate the state and the outcome through prolonged exposure. Moreover, "on-site" replay is precluded by the animal's rapid escape and immediate removal from the environment, although off-site replay may well have occurred in the rest box. In any case, animals exhibited one trial avoidance learning when placed back in the linear track after the rest period: none of the 4 animals entered the shock zone even once during a 10 minute period in the maze, though they had done so repeatedly in a session prior to shock presentation.

Our simulation approximated the design of this experiment. Simulated animals completed three laps of an eight-state environment without salient outcomes, followed by a fourth lap in which a foot shock occurred at state 8 (Fig. 6a). During the foot shock, we simulated movement at double the normal movement speed to align with the animal's rapid escape responses. The foot shock was modeled as highly salient by assigning it a motivational importance of 10 (Fig. 6b). As a result, PS at state 8 increased sharply relative to prior familiarization laps (Fig. 6c), inducing over-representation of that location despite the brief dwell time (Fig. 6d).

We then simulated the subsequent offline phase after the animal was removed from the environment by providing random, low-intensity input to CA3 cells for 10 seconds (Fig. 6e). As a result, spontaneous replay-like activity emerged, spanning the entire track (Fig. 6f). We then used symmetric STDP to update W_{CA3} and W_{CA1} based on the resulting 10 seconds of simulated spiking activity[48].

We found that incorporating offline learning extended the shock prediction field: replay enabled the shock to be predicted from locations farther away along the track (Fig. 6g). This change in prediction translated into altered behavior, simulated using the same method as in the T-maze simulation described above. Simulations of 600-seconds of free movement with and without inclusion of the offline replay based on the transition matrices shown in Fig. 6h revealed near-total avoidance of the shock zone even without replay. Offline learning further reduced the animal's occupancy of the half of the track containing the footshock state (states 5–8; the near-shock zone) (Fig. 6i). We quantified this effect by measuring near-shock zone occupancy, which was significantly decreased when offline learning was included (Fig. 6j, right). Although the experimental study did not directly compare avoidance behavior with and without replay, our simulations highlight the possibility that the strong avoidance observed experimentally may depend in part on replay-mediated propagation of aversive predictions, as demonstrated in our simulations.

Another important finding from the same study was that, when the mice paused as they walked toward the shock zone after reintroduction to the environment, they exhibited replay events including forward-directed replays toward the shock zone[1]. These "forward roll-outs" may allow access to information about the foot shock without physically entering the foot shock zone. To test whether our model would also access this information after the single shock exposure and off-site replay described above, we simulated a second on-site replay phase with place input from the middle of the track (state 4) (Fig. 6k). The network spontaneously generated sequential place cell activity extending toward both ends of the track (Fig. 6l, left), including forward roll-out events that included the footshock state (Fig. 6l, right). Consistent with our hypothesis, internal footshock prediction increased during these roll-out events (Fig. 6m). These results show that our framework provides a basis for capturing the finding that aversive stimuli can affect neural responses after a single exposure to an aversive stimulus, highlighting distinct learning processes during behavior and offline periods that may synergistically contribute to adaptive behavior.

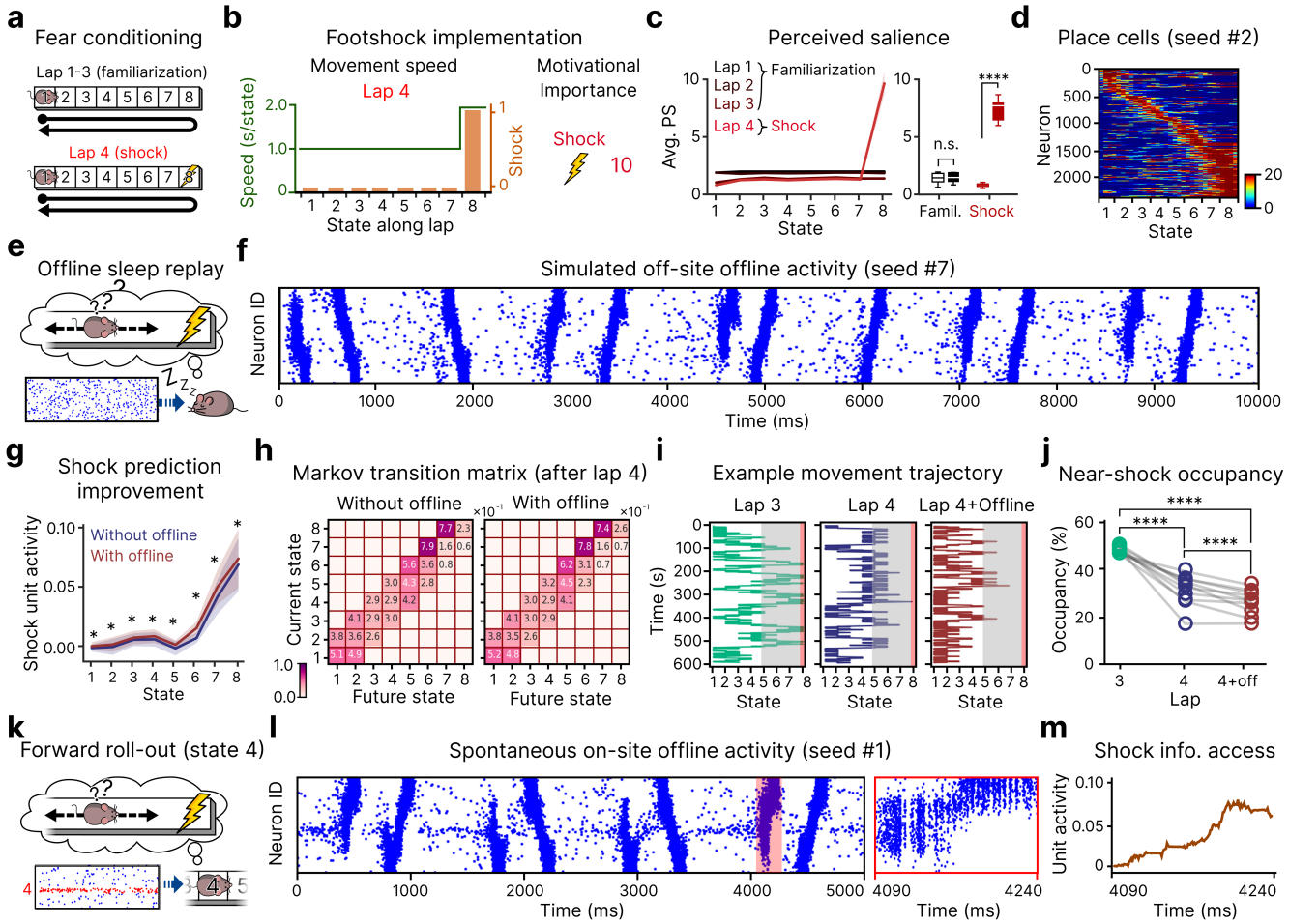


Figure 6. **a**, Aversive feature learning scheme. **b**, Two factors modulated by the introduction of aversive outcome: increased movement speed and the motivational importance of the aversive outcome. **c**, PS experienced at each state on lap 1 to 4. Independent t-test, PS of the non-salient states vs. PS of the salient state, $N_{seed} = 10$, n.s.$P = 0.49$, ****$P = 3.41 \times 10^{-64}$. **d**, CA1 place cells clustered around the punished state, over-representing it. **e**, Offline simulation during sleep. **f**, Example spontaneously generated offline activity (10 s). **g**, Elevated anticipated footshock after the offline learning phase. Paired t-test, footshock prediction at each state without offline learning vs. with offline learning, $N_{seed} = 10$, *$P < 0.05$. **h**, Markov state transition matrices before and after the offline learning. **i**, Examples of simulated movement trajectories before footshock introduction (left), after footshock introduction without offline learning (middle), and after footshock introduction with offline learning (right). The red shaded region indicates the footshock state (state 8), and the grey shaded region indicates near-shock zone (states 5–8). **j**, Occupancy of the near-shock zone. Paired t-test, $N_{seed} = 10$, ****$P < 0.001$. **k**, Forward roll-out at state 4, simulated by inputting state 4-specific input. **l**, Left: Example spontaneously generated offline activity (5 s). Right: Zoomed-in roll-out activity. **m**, Footshock prediction output from feature-prediction layer during the roll-out activity. All data are presented as mean±s.d.

Discussion

In our work, we have combined and extended the ideas and findings of many others to offer an integrative extension of the successor representation theory of the cognitive map and its role in value-based decision making, incorporating BTSP and hippocampal replay to address rapid neural and behavioral adaptation. Our work bridges abstract theory, neural representations, synaptic learning mechanisms, and neural and behavioral outcomes. We have built on the SR theory[8,9] by incorporating successor features[45]: CA3 encodes a salience-neutral representation of environmental structure, CA1 integrates this with salience information to form a PS-weighted SR, and a downstream feature prediction layer computes SFs. The value of predicted features is assessed at decision time to guide behavior to address current needs and goals. The model is learned through BTSP[12,28,40,41] during online exploration and further refined by symmetric STDP during offline replay, relying on circuit properties instantiated in previous models for replay propagation and learning[48]. It adapts swiftly when environmental contingencies change and captures a range of experimentally observed behaviors, including one-trial avoidance learning.

Our framework advances the seminal SR–based account of the cognitive map[8,9] by separately considering PS[14], feature prediction, and value assignment. These distinctions allow the model to capture the proposal that hippocampal networks over-represent salient locations, even when those locations are not inherently rewarding and do not lead to increased future visitation (Figs. 3 and 6). In addition, incorporating SFs, using them to represent environmental features including rewards,

and separately assigning value to these features enables flexible integration of time-varying information such as motivational needs[52]. Because the model predicts features rather than states *per se*, it can adapt behavior based on motivational state without relearning the underlying environmental representation (Fig. 4).

Mechanistically, our work incorporates BTSP's role in encoding distinct representations in hippocampal CA1 and CA3 areas[40]. In area CA1, our model extends the demonstration that temporally asymmetric BTSP can lead to emergence of SR-inspired place cells[28] by incorporating PS weighting to capture the over-representation of rewarding locations[12,31] and demonstrating how this can accelerate learning to predict salient environmental features. We further model how this over-representation may arise through different factors for rewarding versus aversive stimuli. In our simulations, positive outcomes were learned through moderate salience signals combined with prolonged occupancy (Fig. 3), whereas aversive outcomes were encoded through much stronger PS despite briefer dwell times (Fig. 6). Our model shows how these over-representations can accelerate learning of feature–location associations, as elevated activity at salient sites drives faster connection adjustments that support feature predictions (Fig. 3j). In CA3, on the other hand, BTSP learns a symmetric, value-neutral map of environmental structure through only a few laps of exploration, via temporally symmetric BTSP[40].

Our model also highlights how offline replay can contribute to rapid learning by integrating several existing findings. Recent work suggests that symmetric recurrent weight profiles support spontaneous replay[41,48], and that replay contributes to the consolidation of SR-like representations[21]. Integrating these results, our model shows that a value-neutral map encoded in CA3 connection weights can enable spontaneous replay after very limited experience (Fig. 2). Unlike online learning, replay sequences extend along unexplored trajectories[42,43] (Fig. 5), enabling replay-driven inference in which information about salient features is propagated to states not actually visited in sequence, improving predictions. Beyond learning, our model incorporates a role for replay during behavior: simulated offline periods during immobility within exploration generate forward roll-outs similar to those observed experimentally[1] (Fig. 6). These roll-outs transiently activate representations of non-current but salient states, enabling remote feature prediction and state-value estimation necessary for planning[49,53,54].

Notably, our model highlights how BTSP offers an alternative to backpropagation of error signals though learned connection weights to recruit upstream neurons — in our case those in CA1 — to predict down downstream observations — in our case, successor features (Magee[26] discusses related ideas, see Supplementary Materials Section 1.7). In our model, with BTSP, the signals causing CA1 neurons to change their incoming weights depend on a global PS signal, corresponding to the summed prediction error across all experienced features weighted by their intensity and motivational importance. This causes the model to produce upstream representations that support prediction without the transmission of error signals backward through connection weights or through precisely aligned return weight matrices, neither of which is thought to be biologically plausible[55]. The idea that a global signal might drive learning is not new. What is new is the large magnitude and extended temporal window of connection adjustments supported by BTSP, providing a mechanism that appears to allow an approximation of gradient-based learning to proceed far more efficiently than it could under previous proposals[56]. Indeed, this mechanism may accelerate initial learning relative to back-propagation, which requires forward weights to build up before they can carry error signals backward to drive upstream learning[57]. In sum, BTSP may endow a learning system with an implicit and highly efficient approximate error gradient propagation mechanism. This could go a long way toward explaining why biological brains learn far more efficiently than the gradient-propagation based learning systems used in today's AI systems.

In spite of the progress our model represents, many issues and questions remain. We have made many simplifying assumptions to focus on building a conceptual bridge between levels and capturing rapid adaptation. Our simulations provide qualitative demonstrations rather than quantitative matches to details to provide a proof of principle. Future work should include additional pathways in the hippocampus and further details of the physiological implementation of some of the functional circuit properties of hippocampal and other circuits, to create more fully realized simulations of details of the underlying physiological mechanisms and a closer fit to some aspects of neural activity observed during behavioral and replay episodes (See Supplementary Materials Section 1 for further discussion).

Our work so far has focused on relatively simple learning settings. Future work should address more complex settings, including those requiring sensitivity to context. There is a very rich literature exploring how neural activity and behavior generalize from one context to another or differentiates between contexts, as explored in a recent extension of the SR framework[46] and in work using a complementary computationally grounded framework that captures temporal context sensitivity[58,59]. We look forward to exploring whether our framework can be extended to capture the dynamics of representation learning and behavioral adaptation observed in these settings.

Broadening our focus, future work should address the learning efficiency gap[60] between humans and today's AI systems. As noted above, rapid online learning through BTSP and offline refinement through replay may offer alternatives to standard gradient-based updates, leading to a new understanding of how humans learn far more efficiently. It seems likely that BTSP contributes to the rapid formation of replayable mental trajectories that can be reactivated during offline periods in humans, as explored in recent investigations[61]. Insights from models employing BTSP and replay to better capture how these processes play out in human brains could contribute to the future development of more efficient learning methods in AI.

Methods

Neural network model

The hippocampal network

We modeled the hippocampal network as a two-layer spiking neural network, corresponding to the CA3 and CA1 subregions of the hippocampus, modifying and extending a simulation of exploration and replay in area CA3 from previous modeling work of Ecker et al.[48] Consistent with experimental findings, the CA3 layer included recurrent connectivity (W_{CA3}) and projected feedforward output to CA1 (W_{CA1}). Synaptic weights in both pathways were constrained between 0 and 10, initialized from a Gaussian distribution (mean = 0.01, standard deviation = 10^{-4}), and formed with a fixed connection probability of 0.1.

Both CA3 and CA1 layers contained the same number of excitatory neurons which were distributed across a small number of blocks we call 'states' tiling the simulated maze. Each state was represented by 400 neurons, with different numbers of states in the different environments we used as described below. To focus on learning processes within the hippocampus, we adopted simplifications of the inputs and plastic synapses in these brain areas. CA3 excitatory neurons were pre-specified to be place cells, with uniformly distributed place fields and spatial inputs reflecting the animal's position during online exploration, and CA1 excitatory neurons did not receive direct spatial input; instead, their place-selective responses emerged through learned W_{CA1}.

Dynamics of activity in this circuit differ between the online and offline simulation phases as described below.

Feature prediction layer

The feature prediction layer in our model aimed to provide a prediction of the physical presence of environmental features, including background, food, water, and shock. It comprised units representing the predicted presence (ψ) of each feature. These units received fully connected input from CA1 neurons via W_{pred} and allowed computation of a novelty signal (N) contributing to the modulation of the likelihood of plateau potentials in CA1 cells (See below). We interpret this network as a simplified representation of brain circuits involved in feature prediction and value computation, such as the nucleus accumbens, which forms a recurrent loop with CA1 and relays dopaminergic signals.

We now turn to our simulation of different phases of activity, again building on Ecker et al.[48] Whereas these authors used STDP during behavior to set up subsequent offline replay and did not simulate further synaptic changes during replay, we use BTSP during online behavior and STDP during replay to capture the rapid adaptation that is the focus of our work.

Online exploration

We simulated rodent's maze learning task which is commonly studied in experimental literature. On each lap, we moved a notional simulated animal along a path through the environment while training the hippocampal and feature prediction networks.

Maze exploration scenarios

We simulated three different maze learning scenarios:

- **An animal unidirectionally navigating a linear-track treadmill[12]:** Each complete rotation of the belt corresponded to a 2.4 m lap. For convenience, the track was discretized into 8 states (state 1 to 8, from left to right), each 30 cm in length. In each lap, the animal began at the leftmost state (state 1), traversed to the rightmost state (state 8), reached the end of the belt, and was returned to the starting point. Formally,

  $$\mathcal{S}_{\mathrm{lap}} = \langle 1, 2, \ldots, 8 \rangle.$$

  The next lap began immediately after the animal returned to state 1.

- **An animal navigating a T-maze[52]:** The maze consisted of a 0.9 m vertical stem (state 1 to 3, from bottom to top), connected at its midpoint to a 2.1 m horizontal arm (state 4 to 10, from left to right, with state 7 as the intersection). In each lap, the animal began at the base of the stem (state 1), moved up to the center of the horizontal arm (state 7), and then turned left in odd-numbered laps (ending at state 4) or right in even-numbered laps (ending at state 10).

  Formally,

  $$\mathcal{S}_{\mathrm{lap}N} = \begin{cases} \langle 1,2,3,7,6,5,4 \rangle, & \text{if } N \text{ is odd,} \\ \langle 1,2,3,7,8,9,10 \rangle, & \text{if } N \text{ is even.} \end{cases}$$

- **An animal round-tripping a linear maze[1]:** This scenario replicates an animal exploring a confined, 2.4 m segment of linear maze. The track was discretized into 8 states (state 1 to 8, from left to right), but unlike the linear treadmill, state 8

and state 1 are not connected. A single lap began with the animal at the left end (state 1), moving sequentially to the right end (state 8), and then returning back to the start (state 1). Formally,

$$\mathcal{S}_{\text{lap}} = \langle 1, 2, \ldots, 7, 8, 7, \ldots, 1 \rangle.$$

Environmental features

In our simulation, states were associated with distinct environmental features, f. Environmental features are accompanied with their physical intensity I_f as well as motivational importance MI_f. The presence and absence of the features at each of the designated states s of a given environment were encoded as a presence vector $P(s)$ which remained constant across all positions within that state. Although real maze exploration environments contain many features, we simplified this to capture essential elements relevant to our work: background, food, water and shock.

- **Background:** Each state was assigned a state-specific background feature, which we thought of as the set of sensory cues in that state enabling the animal to distinguish the state from other ones. Since this feature is not attached to any affective value, we set the MI of this feature to be 1. When the state only has background but not other significant environmental features, the animal moves at the baseline speed, taking $T = 1$ second.
- **Food and water:** These features serves as a positive outcome in our simulation, and has MI of 3. When the animal was passing through the state with one of these features, it slowed to half the baseline speed, simulating an extended stay during reward consumption ($T = 2$ s).
- **Shock:** This feature serves as a negative outcome, with MI of 10. Once the animal enters a state where a shock occurs, it speeds up to twice the baseline speed for the shock state and subsequent states, reflecting an escape response ($T = 0.5$ s).

While stimuli can vary in intensity, for simplicity we set I_f to a constant unit value if feature f is present, so that $I_f = P_f$.

Another component attached to environmental features is novelty, denoted as $N_f(t)$. This quantity tracks the animal's subjective surprise associated with feature f, and was defined as the difference between the predicted presence and actual presence of the feature. The novelty values were updated at each timestep based on the absolute prediction error computed by the feature prediction network, ε, described in the section below on training the feature prediction network.

$$N_f(t) = \left|\varepsilon_f(t)\right| \tag{8}$$

Computation of perceived salience

The perceived salience (PS) of an environmental feature f, denoted as PS_f, represents the animal's subjective evaluation of how noticeable or memorable f is. In our simulations, PS was modulated by three factors: motivational importance MI, intensity I, and the novelty N of the environmental feature. Specifically, the PS of feature f, PS_f, was then computed as:

$$PS_f(t) = (I_f + 1) \cdot MI_f \cdot N_f(t) \tag{9}$$

The added 1 in the expression $(I_f + 1)$ is a fixed offset that allows the absence of a feature to affect learning. The PS of a state at time t, $PS(t)$, was computed as the sum of PS across all environmental features.

Spike train generation and activity level estimation

CA3 place fields were uniformly distributed along the x- and y-axes, with each field centered at a distinct position. Place fields were circular with a radius of 30 cm.

Spiking activity in CA3 excitatory neurons arose from two sources: external spatial input from the entorhinal cortex, $\mathcal{I}_{\text{place}}$, and recurrent input from other CA3 neurons, $\mathcal{I}_{\text{recurrent}}$. i_{place} for a CA3 neuron depended on the animal's Euclidean distance from the center of the neuron's place field and followed a Gaussian profile. The maximum value of $\mathcal{I}_{\text{place}}$ was fixed at 1 and the standard deviation of the Gaussian was chosen to be the radius of the field.

Recurrent input was computed as the sum of activities from all connected presynaptic neurons, weighted by W_{CA3}. This summed input was then passed through a sigmoid nonlinearity to yield the final recurrent drive $\mathcal{I}_{\text{recurrent},i}$ for each neuron i. For a sigmoid function

$$\sigma(x; a, b) = \frac{1}{1 + \exp[-b(x - a)]}, \tag{10}$$

the recurrent input to CA3 neuron i was defined as

$$\mathcal{I}_{\text{recurrent},i} = \sigma\left(\sum_j W_{\text{CA3},ij}\, r_j;\ a = 4,\ b = 3\right), \tag{11}$$

where $\sum_j W_{\mathrm{CA3},ij}\, r_j$ is the weighted presynaptic activity, and a and b are the sigmoid offset and gain parameters, respectively.

Finally, the total firing rate of neuron i was given by:

$$FR_{\mathrm{i}} = 20 \times \mathcal{I}_{\mathrm{place,i}} + \mathcal{I}_{\mathrm{recurrent,i}}. \tag{12}$$

CA1 excitatory neurons were modeled similarly but did not receive direct spatial input. Their firing rates were determined exclusively by feedforward input from CA3, $\mathcal{I}_{\mathrm{feedforward}}$: the activity of CA3 presynaptic neurons was multiplied by W_{CA1} and passed through the same sigmoid nonlinearity (Eq. 10), with parameters $a = 4$ and $b = 2$. Formally, the feedforward drive to CA1 neuron j is:

$$\mathcal{I}_{\mathrm{feedforward},j} = \sigma\left(\sum_i W_{\mathrm{CA1},ji}\, r_i;\ a = 4,\ b = 2\right), \tag{13}$$

and the resulting firing rate is:

$$FR_j = 20 \times \mathcal{I}_{\mathrm{feedforward},j}. \tag{14}$$

After calculating FRs of all neurons in both CA3 and CA1, we generated spike trains for each neuron using inhomogeneous poisson processes[28,48]. These spike trains were then filtered to remove spikes that violated an assumed 5 ms refractory period[48].

To reduce noise in estimation of post-synaptic signals that arises from the small scale of our network, we introduced temporal smoothing: spike trains for both CA3 and CA1 were generated at every $\Delta t_{\mathrm{spike}}$ interval, which was set to 100 ms. Neuronal activity $\mathcal{A}$ was quantified as the actual average firing rate of each neuron, calculated by dividing the total number of generated spikes by $\Delta t_{\mathrm{spike}}$. Note that this differs from the ideal FR values in the equations above, as it reflects post-processing by both the inhomogeneous poisson process and the refractory period filter[41,48].

Training the hippocampal network using BTSP

Synaptic weight updates in the hippocampal network followed a BTSP rule, which depends on the interaction between presynaptic eligibility traces (ETs) and postsynaptic instructive signals (ISs)[19,31,40]. ETs were computed based on the spiking activity of presynaptic neurons, while ISs were driven by the postsynaptic neuron's dendritic plateau potentials (PPs). When a presynaptic spike occurred, the post-synaptic neuron's ET increased by a fixed amplitude, A_{ET}, and then decayed exponentially with a time constant $\tau_{\mathrm{ET}} = 1000$ ms.

ISs were evoked by dendritic plateau potentials (PPs), modeled as stochastic events occurring in both CA3 and CA1 neurons. Each neuron k was assigned a plateau probability $p_{\mathrm{plateau},k}$, updated every $\Delta t_{\mathrm{spike}}$. At each 1 ms time step, a Bernoulli trial determined whether neuron k initiated a PP.

The plateau probability $p_{\mathrm{plateau},k}$ was modulated by three factors: (1) the neuron's activity $\mathcal{A}_k$ of the current $\Delta t_{\mathrm{spike}}$ interval, (2) the perceived salience (PS) at the current time step, and (3) the mean activity of the entire layer, $\mathcal{A}_{\mathrm{population}}$, which reflects layerwise inhibition and downregulates PP probability[19,31] (Supplementary Fig. 1). Formally, the plateau probability for neuron k was:

$$p_{\mathrm{plateau},k} = \mathcal{A}_k \cdot (1 + 3 \cdot PS) \cdot (c_{\mathrm{population}} + c_{\mathrm{min}}). \tag{15}$$

As CA3 layer did not receive a PS signal, PS was set to 0 and $c_{\mathrm{min}} = 0$ for CA3 neurons. For CA1 neurons, $c_{\mathrm{min}} = 10^{-5}$.

The inhibitory scaling factor $c_{\mathrm{population}}$ was:

$$c_{\mathrm{population}} = 2\, c_{\mathrm{max}}\, \sigma\big(\mathcal{A}_{\mathrm{target}} - \mathcal{A}_{\mathrm{population}}, 0, 5\big) - c_{\mathrm{max}} \tag{16}$$

for $\mathcal{A}_{\mathrm{population}} < \mathcal{A}_{\mathrm{target}}$, and

$$c_{\mathrm{population}} = 0, \tag{17}$$

otherwise. We set $c_{\mathrm{max}} = 0.05 \times (1 + 0.5 \cdot PS)$ where PS was fixed to 0 for CA3 neurons.

When a PP occurred, the postsynaptic neuron's instructive signal increased by A_{IS} and decayed exponentially with a time constant τ_{IS}. For CA3 neurons, $A_{\mathrm{IS}} = 1.6$ and $\tau_{\mathrm{IS}} = 1000$ ms, same as τ_{ET} which allows symmetric time kernel of BTSP. On the other hand, CA1 layer's plateau potentials were modulated by P signal: $A_{\mathrm{IS}} = 0.4 \times (0.25 \cdot PS)$ and $\tau_{\mathrm{IS}} = 500$ ms for CA1 neurons, producing an asymmetric kernel[40]. BTSP-mediated synaptic updates occurred only during PP events. The BTSP update of the weights is a function of $ET \times IS$ as well as the initial value of the weight[31]. Initially weak weights undergo stronger potentiation even with smaller $ET \times IS$, whereas the weights that are already strong enough are more likely to undergo depression even with stronger ET and IS signal. Formally,

$$\Delta w_{ji} = (w_{\max} - w_{ji})\, k_{\text{pos}}\, c_{\text{pos}} - w_{ji}\, k_{\text{neg}}\, c_{\text{neg}}, \tag{18}$$

where

$$c_{\text{pos}} = \frac{\sigma(ET_i \cdot IS_j; a_{\text{pos}}, b_{\text{pos}}) - \sigma(0; a_{\text{pos}}, b_{\text{pos}})}{\sigma(1; a_{\text{pos}}, b_{\text{pos}}) - \sigma(0; a_{\text{pos}}, b_{\text{pos}})}, \tag{19}$$

$$c_{\text{neg}} = \frac{\sigma(ET_i \cdot IS_j; a_{\text{neg}}, b_{\text{neg}}) - \sigma(0; a_{\text{neg}}, b_{\text{neg}})}{\sigma(1; a_{\text{neg}}, b_{\text{neg}}) - \sigma(0; a_{\text{neg}}, b_{\text{neg}})}. \tag{20}$$

In our simulations, we set $k_{\text{pos}} = 0.085$, $k_{\text{neg}} = 0.01$, $a_{\text{pos}} = 0.8$, $b_{\text{pos}} = 6$, $a_{\text{neg}} = 0.05$, and $b_{\text{neg}} = 44.44$.

Training the feature prediction network using the delta rule

The feature prediction network was trained during the online exploration phase using the error-correcting delta rule[47]. Specifically, a synaptic weight w_{fi} connecting a CA1 neuron i and a feature unit f, was updated at each millisecond-level time step t as:

$$\Delta w_{fi}(t) = \mu \cdot \mathcal{A}_i(t) \cdot \varepsilon_f(t) \tag{21}$$

where μ is the learning rate, $\mathcal{A}_i(t)$ is the activity of neuron i and $\varepsilon_f(t)$ is the prediction error of feature f at time t. In our simulations, μ was set to 0.15×10^{-7} for the linear-track treadmill and linear maze scenarios, and 0.125×10^{-7} for the T-maze scenario. The prediction error was defined as the error between predicted presence (ψ) and the physical presence (ϕ) for that feature:

$$\varepsilon_f(t) = \psi_f(t) - \phi_f, \text{where} \tag{22}$$

$$\psi_f(t) = \sum_i \mathcal{A}_i(t) \cdot w_{ji}(t) \tag{23}$$

Finally, the novelty N associated with the feature at time step t is equal to the absolute value of the prediction error of the feature at that time step.

Offline phase

The offline simulation framework was adapted from Ecker et al.[48] Here, neurons were modeled using adaptive exponential integrate-and-fire dynamics, and synapses were modeled as conductances with biexponential kinetics (Supplementary Materials, Section 3.1 and 3.2). Hippocampal weights learned during online exploration (W_{CA3} and W_{CA1}) were carried over into the offline phase, with additional connectivity to inhibitory interneurons introduced. Spontaneous activity emerging in the CA3 and CA1 layers during this phase was then used to update W_{CA3} and W_{CA1} according to a symmetric STDP rule.

Connectivity optimization

Additional connectivity was introduced and optimized using an evolutionary algorithm, including: entorhinal cortex input to CA3 excitatory cells (EC→Ext_{CA3}); CA3 excitatory cell to CA3 inhibitory interneuron (Ext_{CA3}→Inh_{CA3}); interneuron–interneuron (Inh_{CA3}→Inh_{CA3}); and inhibitory interneuron to excitatory cell (Inh_{CA3}→Ext_{CA3}). The same connectivity types were established in CA1 (Ext_{CA1}→Inh_{CA1}, Inh_{CA1}→Inh_{CA1}, Inh_{CA1}→Ext_{CA1}).

The optimization algorithm minimized a loss function designed to match network-level physiological targets: maintaining realistic population firing rates in excitatory cells, suppressing gamma-band (30–100 Hz) oscillations, and enhancing ripple-band oscillations in excitatory cell populations (Supplementary Materials, Section 3.3).

During the offline phase, spontaneous activity was simulated in both CA3 and CA1, with low-rate entorhinal cortex input to CA3. W_{CA3} and W_{CA1} were updated using a symmetric STDP rule throughout this period.

Training the hippocampal network using symmetric STDP

During offline phase simulations with the optimized synaptic connectivity, we recorded the spiking activity of individual neurons in the CA3 and CA1 layers. These recorded spike trains were then used to update W_{CA3} and W_{CA1} according to a symmetric spike-timing–dependent plasticity (STDP) rule.

Symmetric STDP only involves potentiating the connections which connect neurons that fire together. This occurred in the second pass through of the recorded spike trains, using a time step of 1 msec. Precisely, the update rule we used is described as follows, with $A_{\text{STDP}} = 1 \times 10^{-3}$ and $\tau_{\text{STDP}} = 20$ ms.

$$\Delta w = A_{\text{STDP}} \cdot e^{|t_{post} - t_{pre}|/\tau_{\text{STDP}}} \tag{24}$$

To simulate replay-based learning after exposure to food rewards in the T-maze (Fig. 5d), we isolated spiking activity within the temporal window corresponding to a single replay event. The learning effect was quantified by comparing the network's predictions before and after the weight update induced by each replay event (Supplementary Materials, Section 3.4). Similarly, the learning effect after the full 5-second offline simulation was also simulated and analyzed using the entire spiking activity (5e).

To simulate replay-based learning after exposure to shock in a linear track (Fig 6g), weight updates were based on all of the spiking activity across the 10-second offline and off-site replay episode.

Estimating real-time feature unit activity during roll-out events

To estimate real-time feature unit activity in the feature prediction layer during offline replay (Fig. 6m), we used spike trains recorded during offline simulations. For each detected roll-out event, we computed CA1 firing rates using spikes within a 100 ms window preceding each time point, starting from the event's onset time. This procedure was repeated at 1 ms resolution throughout the duration of the roll-out event. Then, these CA1 firing rates were multiplied by the feature prediction weight matrix W_{FP} learned during online training to estimate feature unit activity at every time point.

Computation of successor feature activations

Here we describe how we determined the network's representations of successor features $\Psi_f(s)$ for a given feature f at any location s in the environment based on the knowledge in connection weights before or after exploration and/or replay. These quantities are shown directly for 32 evenly spaced positions s along the linear track used in simulations shown in Figs. 2n and 3k,l and were also used as the basis for choosing actions in our behavioral simulations shown in Figs. 4 to 6.

We first computed CA3 neuronal activity associated with position s. A position- and input-dependent firing rate for each CA3 neuron was first determined by using its specified spatial tuning function relative to the current position. This was then scaled by a factor of 0.45 to approximate the influence of recurrent input through W_{CA3}. CA1 neuronal activity was subsequently computed in the same manner as during the online exploration phase. Specifically, CA3 activity was forward-propagated through the feedforward weights W_{CA1} and passed through a sigmoid nonlinearity to obtain CA1 firing rates. Finally, feature-prediction unit activity was computed by linearly projecting CA1 activity through W_{pred}, giving estimates of predicted feature activation $\Psi_f(s)$.

Behavioral simulation

We conducted behavioral simulations using the information encoded in the network's three weight matrices W_{CA3}, W_{CA1}, and W_{FP} network to examine the effects of learning on behavior. This phase corresponds to the exploitation stage in reinforcement learning, during which the agent uses previously acquired knowledge to make choices. No weight updating occurred during these simulations. We simulated two behavioral decision-making scenarios: one involving a positive outcome in a T-maze and another involving an aversive outcome in a linear track. In both cases, we treated the behaving animal as standing at the center of one of the discrete states in the relevant environment, and as computing an estimate of $V(s)$ for the current state and for each state immediately adjacent to it in the environment. We used these $v(s)$ values to determine next state transition probabilities, and then sampled from these probabilities to transition to the next state.

To estimate $V(s)$, we first computed neuronal and feature-unit activity from the learned weights as described in the section above. The value of a given state s, denoted $V(s)$, was computed as:

$$V(s) = \sum_f \psi_f(s) \cdot v_f, \tag{25}$$

where $\psi_f(s)$ represents the predicted presence of feature f at state s, and v_f denotes the subjective value of that feature. The feature values v_f were dynamically modulated by the animal's transient motivational state. Given these computed state values, transition probabilities from each state to each other accessible state were computed using a softmax function. Specifically, given the vector of state values $\mathbf{V} = \{V_1, V_2, \ldots, V_N\}$ for the set of accessible states, transition probabilities were computed as follows:

$$P(s_{t+1} = j \mid s_t) = \frac{\exp(V_j)}{\sum_{k=1}^{N} \exp(V_k)}. \tag{26}$$

The simulated next state was determined by sampling from these probabilities.

Decision-making scenarios

- **T-maze with two goal locations and distinct rewards[52]:** Each trial began with the animal positioned at the stem of the maze (state 1) and ended when it reached either goal state (state 4 or 10). Food was located at the left end (state 4) and water at the right end (state 10), and the animal's motivational needs for these two rewards were systematically varied.

  To test the network's adaptability to different starting conditions, additional simulations were conducted in which the animal started from the right end (state 10) and traversed either toward the left end (state 4) or the stem (state 1), with food placed at state 4.

  In all cases, the animal was allowed a maximum of 50 steps per trial. Trials were terminated if the animal failed to reach a goal within this limit, and such cases were labeled as "failure to find the goal."

- **Linear track with an aversive outcome[1]**: Each trial began with the animal positioned at the left end of the track (state 1) and allowed to move freely for 600 steps. A shock stimulus was positioned at the right end of the track (state 8).

Statistical analysis

All statistical analyses were performed using custom Python code. Unless otherwise stated, data points represent independent simulation runs initialized with different random seeds (N_{seed} indicated in figure captions).

For comparisons between two conditions, we used paired t-tests as appropriate. One-sample t-tests were used when comparing a condition against a theoretical baseline.

Error bars represent mean $\pm$ standard deviation (s.d.) unless otherwise noted. Exact P values and statistical tests used are reported in the corresponding figure captions.

Acknowledgement

We thank A. Milstein (Rutgers University) and J. Magee (Howard Hughes Medical Institute) for valuable feedback on our implementation of BTSP and for their insights regarding the biological plausibility of our results. We also thank S. Jones, S. Grant, and J. Han (Stanford University) for helpful discussions during the early stages of the project. S.C. acknowledges support from the Emergent Cognitive Functions Fund, Department of Psychology, Stanford University.

References

1. Wu, C.-T., Haggerty, D., Kemere, C. & Ji, D. Hippocampal awake replay in fear memory retrieval. *Nat. Neurosci.* **20**, 571–580, DOI: 10.1038/nn.4507 (2017).
2. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208, DOI: 10.1037/h0061626 (1948).
3. Behrens, T. E. J. *et al.* What is a cognitive map? organizing knowledge for flexible behavior. *Neuron* **100**, 490–509, DOI: https://doi.org/10.1016/j.neuron.2018.10.002 (2018).
4. Rosenberg, M., Zhang, T., Perona, P. & Meister, M. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *eLife* **10**, e66175, DOI: 10.7554/eLife.66175 (2021).
5. O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Exp. Neurol.* **51**, 78–109, DOI: https://doi.org/10.1016/0014-4886(76)90055-8 (1976).
6. O'keefe, J. & Nadel, L. *The hippocampus as a cognitive map* (Oxford university press, 1978).
7. Frank, L. M., Stanley, G. B. & Brown, E. N. Hippocampal plasticity across multiple days of exposure to novel environments. *The J. Neurosci.* **24**, 7681, DOI: 10.1523/JNEUROSCI.1958-04.2004 (2004).
8. Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Comput.* **5**, 613–624 (1993).
9. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653, DOI: 10.1038/nn.4650 (2017).
10. Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25**, 707–715, DOI: 10.1016/S0896-6273(00)81072-7 (2000).
11. Gustafson, N. J. & Daw, N. D. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLOS Comput. Biol.* **7**, e1002235, DOI: 10.1371/journal.pcbi.1002235 (2011).
12. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).

13. Krishnan, S., Heer, C., Cherian, C. & Sheffield, M. E. Reward expectation extinction restructures and degrades ca1 spatial maps through loss of a dopaminergic reward proximity signal. *Nat. Commun.* **13**, DOI: 10.1038/s41467-022-34465-5 (2022).

14. Kutlu, M. G. *et al.* Dopamine release in the nucleus accumbens core signals perceived saliency. *Curr. Biol.* **31**, 4748–4761.e8, DOI: 10.1016/j.cub.2021.08.052 (2021). Doi: 10.1016/j.cub.2021.08.052.

15. Hollup, S. A., Molden, S., Donnett, J. G., Moser, M.-B. & Moser, E. I. Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *The J. Neurosci.* **21**, 1635, DOI: 10.1523/JNEUROSCI.21-05-01635.2001 (2001).

16. Gauthier, J. L. & Tank, D. W. A dedicated population for reward coding in the hippocampus. *Neuron* **99**, 179–193.e7, DOI: https://doi.org/10.1016/j.neuron.2018.06.008 (2018).

17. Mamad, O. *et al.* Place field assembly distribution encodes preferred locations. *PLOS Biol.* **15**, e2002365– (2017).

18. Nikbakht, N. *et al.* Efficient encoding of aversive location by CA3 long-range projections. *Cell Rep.* **43**, 113957 (2024).

19. Grienberger, C. & Magee, J. C. Entorhinal cortex directs learning-related changes in ca1 representations. *Nature* **611**, 554–562, DOI: 10.1038/s41586-022-05378-6 (2022).

20. Marr, D. Vision: A computational investigation into the human representation and processing of visual information (1982).

21. Bono, J., Zannone, S., Pedrosa, V. & Clopath, C. Learning predictive cognitive maps with spiking neurons during behavior and replays. *Elife* **12** (2023).

22. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018), 2nd edn.

23. Shouval, H. Z., Bear, M. F. & Cooper, L. N. A unified model of nmda receptor-dependent bidirectional synaptic plasticity. *Proc. Natl. Acad. Sci.* **99**, 10831–10836, DOI: 10.1073/pnas.152343099 (2002). https://www.pnas.org/doi/pdf/10.1073/pnas.152343099.

24. Priestley, J. B., Bowler, J. C., Rolotti, S. V., Fusi, S. & Losonczy, A. Signatures of rapid plasticity in hippocampal CA1 representations during novel experiences. *Neuron* **110**, 1978–1992.e6 (2022).

25. Madar, A. D. *et al.* Behavioral timescale synaptic plasticity: A burst in the field of learning and memory. *J. Neurosci.* **45**, e1332252025 (2025).

26. Magee, J. C. Behavioral timescale synaptic plasticity: properties, elements and functions. *Nat. Neurosci.* (2026).

27. Madar, A. D., Jiang, A., Dong, C. & Sheffield, M. E. J. Synaptic plasticity rules driving representational shifting in the hippocampus. *Nat. Neurosci.* **28**, 848–860 (2025).

28. Galloni, A. R. *et al.* Neuromorphic one-shot learning utilizing a phase-transition material. *Proc. Natl. Acad. Sci.* **121**, e2318362121, DOI: 10.1073/pnas.2318362121 (2024). https://www.pnas.org/doi/pdf/10.1073/pnas.2318362121.

29. Bittner, K. C. *et al.* Conjunctive input processing drives feature selectivity in hippocampal ca1 neurons. *Nat. Neurosci.* **18**, 1133–1142, DOI: 10.1038/nn.4062 (2015).

30. Balind, S. R. *et al.* Diverse synaptic and dendritic mechanisms of complex spike burst generation in hippocampal ca3 pyramidal cells. *Nat. Commun.* **10**, 1859, DOI: 10.1038/s41467-019-09767-w (2019).

31. Milstein, A. D. *et al.* Bidirectional synaptic plasticity rapidly modifies hippocampal representations. *eLife* **10**, e73046, DOI: 10.7554/eLife.73046 (2021).

32. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679, DOI: 10.1126/science.8036517 (1994). Doi: 10.1126/science.8036517.

33. Skaggs, W. E. & McNaughton, B. L. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870–1873 (1996).

34. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223, DOI: 10.1038/nn.2384 (2009).

35. McNaughton, B. L. Cortical hierarchies, sleep, and the extraction of knowledge from memory, DOI: 10.1016/j.artint.2009.11.013 (2010).

36. Goto, A. & Hayashi, Y. Offline neuronal activity and synaptic plasticity during sleep and memory consolidation. *Neurosci. Res.* **189**, 29–36 (2023).

37. Robinson, H. L. *et al.* Large sharp-wave ripples promote hippocampo-cortical memory reactivation and consolidation during sleep. *Neuron* (2025).

38. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).

39. Gillespie, A. K. *et al.* Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. *Neuron* **109**, 3149–3163.e6, DOI: https://doi.org/10.1016/j.neuron.2021.07.029 (2021).

40. Li, Y., Briguglio, J. J., Romani, S. & Magee, J. C. Mechanisms of memory-supporting neuronal dynamics in hippocampal area ca3. *Cell* **187**, 6804–6819.e21, DOI: 10.1016/j.cell.2024.09.041 (2024). Doi: 10.1016/j.cell.2024.09.041.

41. Milstein, A. D., Tran, S., Ng, G. & Soltesz, I. Offline memory replay in recurrent neuronal networks emerges from constraints on online dynamics. *The J. Physiol.* **601**, 3241–3264, DOI: https://doi.org/10.1113/JP283216 (2023).

42. Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. Hippocampal replay is not a simple function of experience. *Neuron* **65**, 695–705, DOI: https://doi.org/10.1016/j.neuron.2010.01.034 (2010).

43. Bakermans, J. J. W., Warren, J., Whittington, J. C. R. & Behrens, T. E. J. Constructing future behavior in the hippocampal formation through composition and replay. *Nat. Neurosci.* **28**, 1061–1072, DOI: 10.1038/s41593-025-01908-3 (2025).

44. Gershman, S. J. *et al.* Explaining dopamine through prediction errors and beyond. *Nat. Neurosci.* **27**, 1645–1655, DOI: 10.1038/s41593-024-01705-4 (2024).

45. Barreto, A. *et al.* Successor features for transfer in reinforcement learning. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

46. Geerts, J. P., Gershman, S. J., Burgess, N. & Stachenfeld, K. L. A probabilistic successor representation for context-dependent learning. *Psychol. Rev.* **131**, 578 (2024).

47. Widrow, B. & Hoff, J., Marcian E. Adaptive switching circuits. In *IRE WESCON Conference Record, part 4*, 96–104 (1960).

48. Ecker, A. *et al.* Hippocampal sharp wave-ripples and the associated sequence replay emerge from structured synaptic interactions in a network model of area ca3. *eLife* **11**, DOI: 10.7554/eLife.71850 (2022).

49. Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63**, 497–507, DOI: 10.1016/j.neuron.2009.07.027 (2009). Doi: 10.1016/j.neuron.2009.07.027.

50. Gallistel, C. R. & Gibbon, J. Time, rate, and conditioning. *Psychol. Rev.* **107**, 289–344 (2000).

51. Bouton, M. E. Context and behavioral processes in extinction. *Learn. Mem.* **11**, 485–494 (2004).

52. Carey, A. A., Tanaka, Y. & van der Meer, M. A. A. Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nat. Neurosci.* **22**, 1450–1459, DOI: 10.1038/s41593-019-0464-6 (2019).

53. Johnson, A. & Redish, A. D. 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 (2007).

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

55. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).

56. Mazzoni, P., Andersen, R. A. & Jordan, M. I. A more biologically plausible learning rule for neural networks. *Proc. Natl. Acad. Sci.* **88**, 4433–4437 (1991).

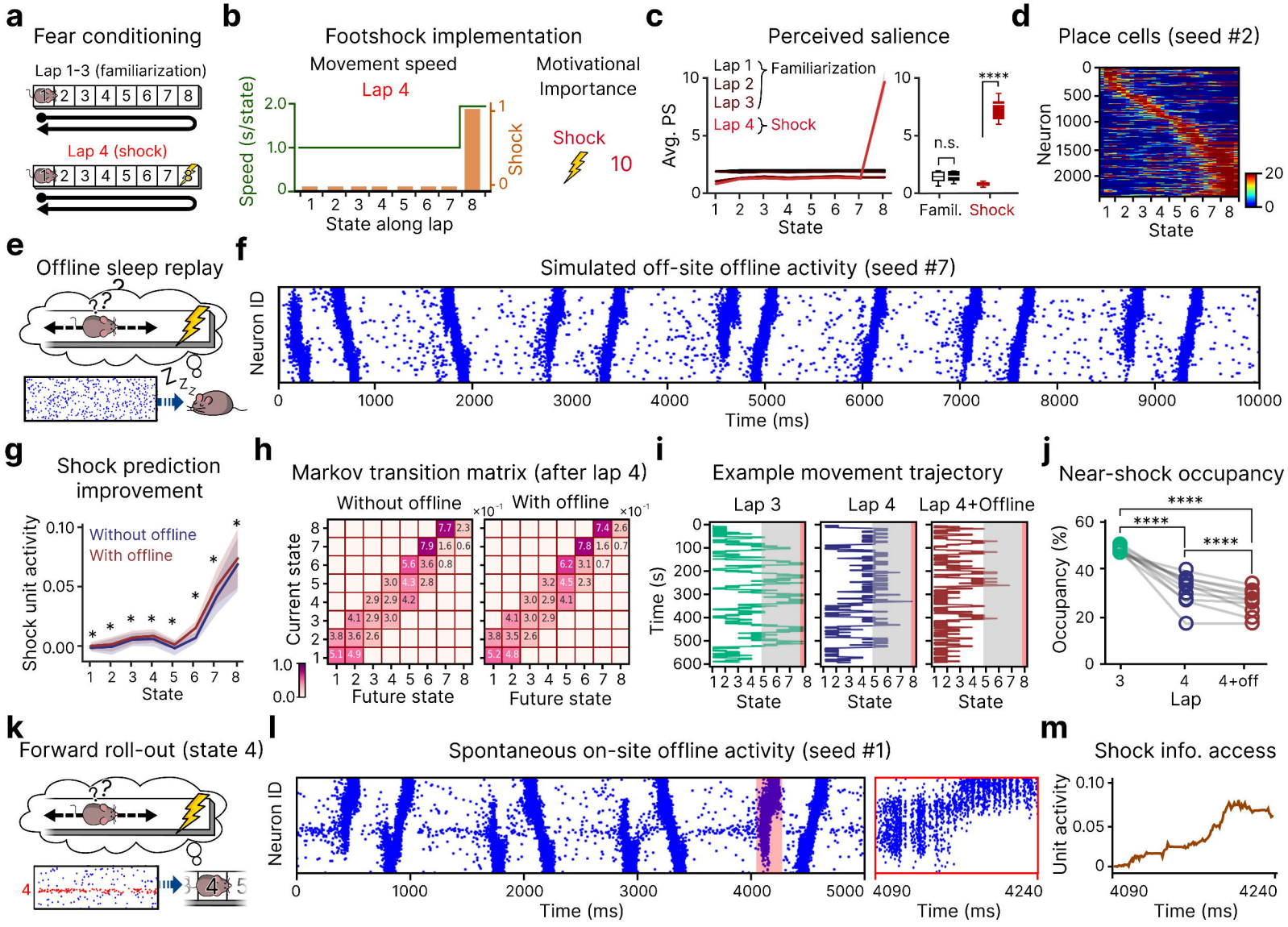
57. Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci.* **116**, 11537–11546 (2019).

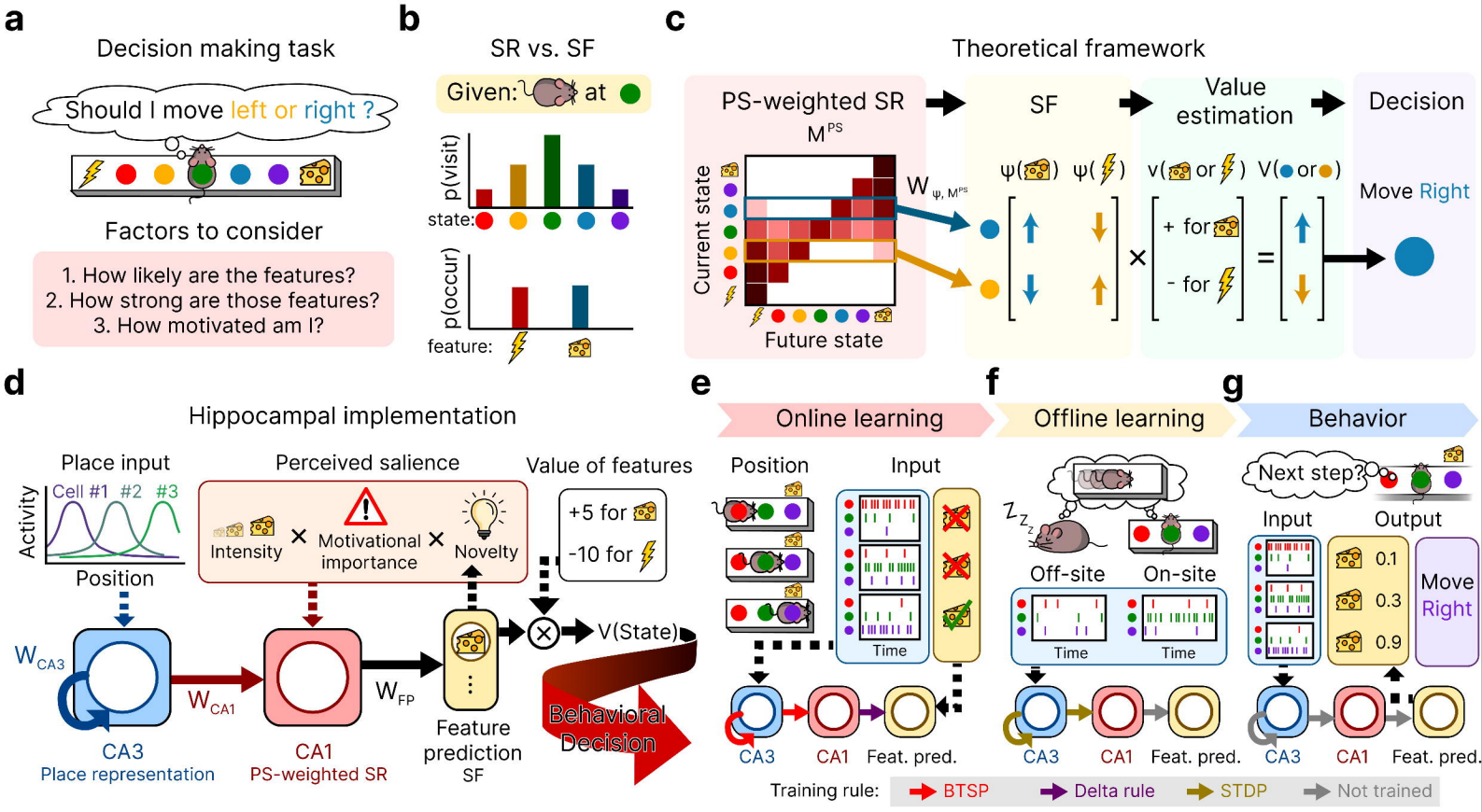
58. George, D. *et al.* Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. communications* **12**, 2392 (2021).

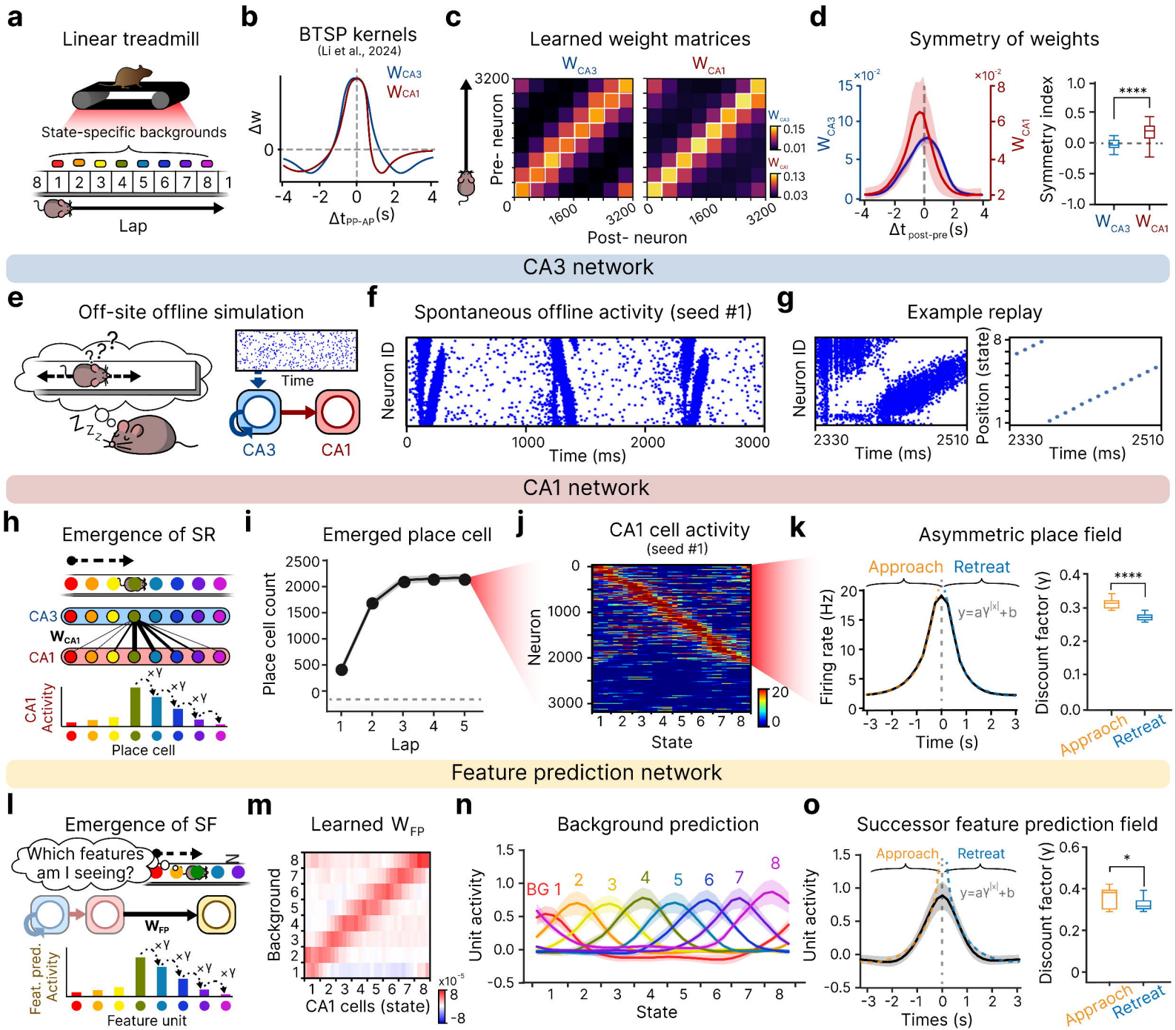
59. Raju, R. V. *et al.* Space is a latent sequence: A theory of the hippocampus. *Sci. Adv.* **10**, eadm8470 (2024).

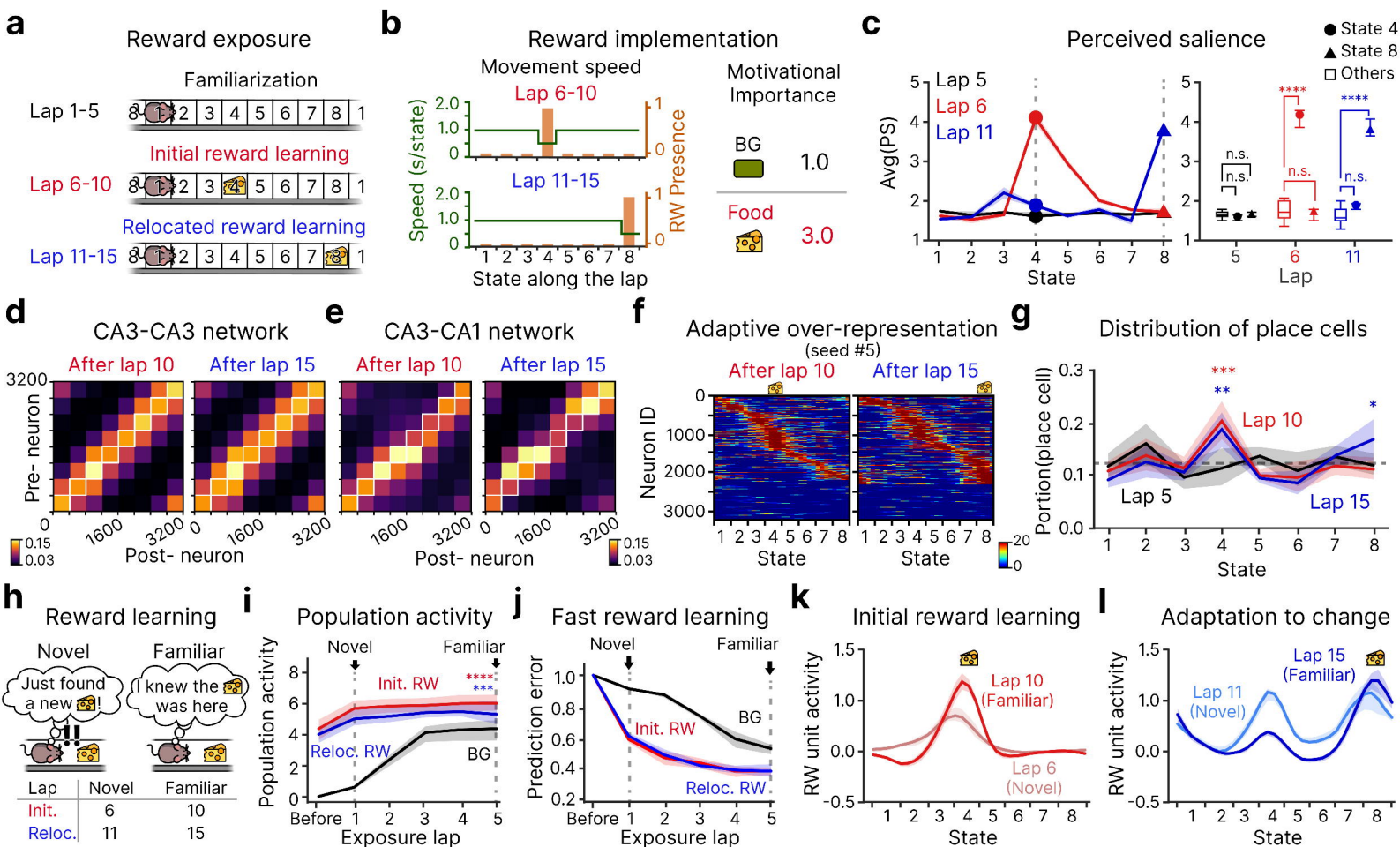
60. Frank, M. C. Bridging the data gap between children and large language models. *Trends Cogn. Sci.* **27**, 990–992 (2023).

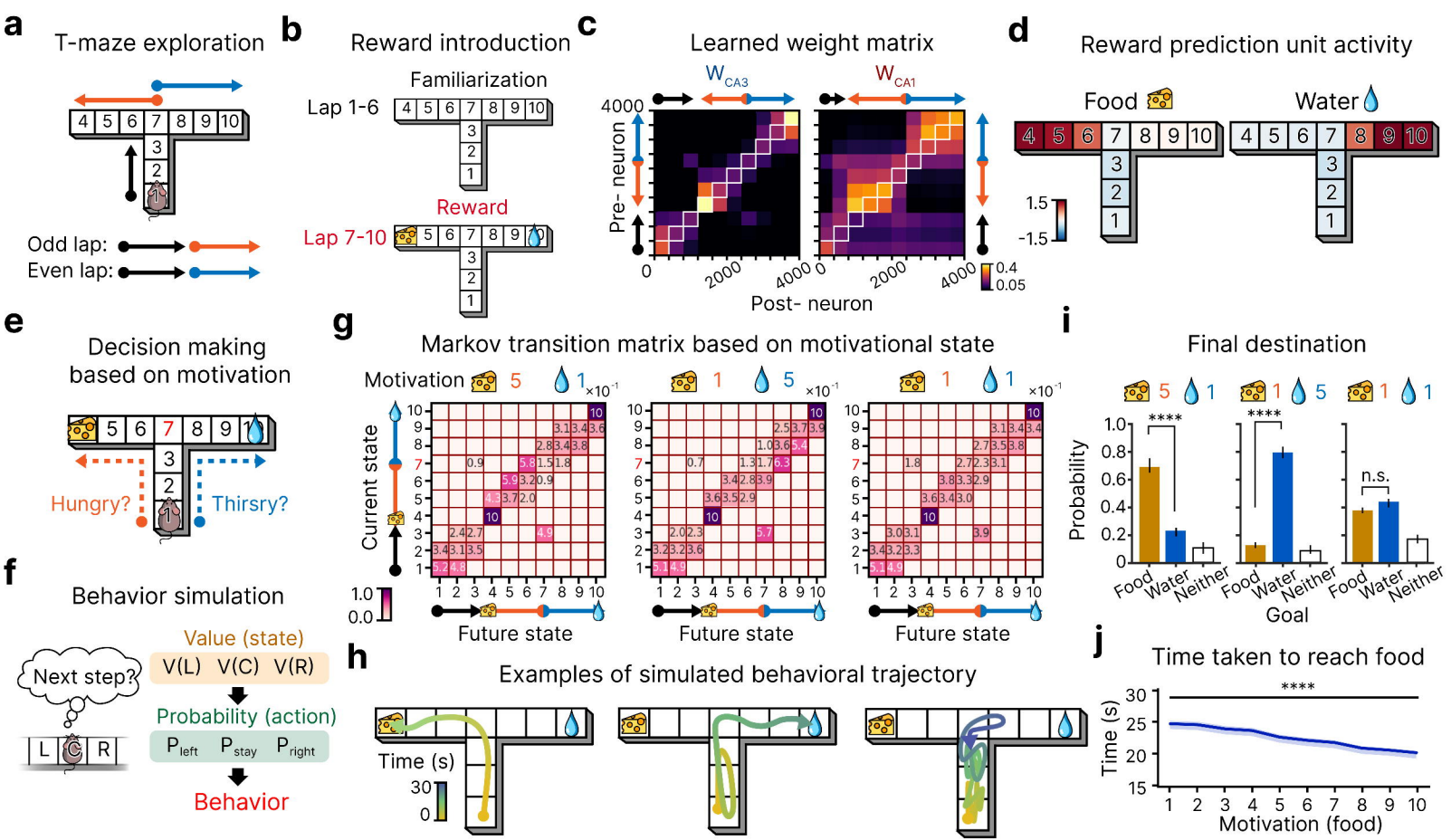
61. Su, H. *et al.* Neural dynamics of spontaneous memory recall and future thinking in the continuous flow of thoughts. *Nat. communications* **16**, 6433 (2025).

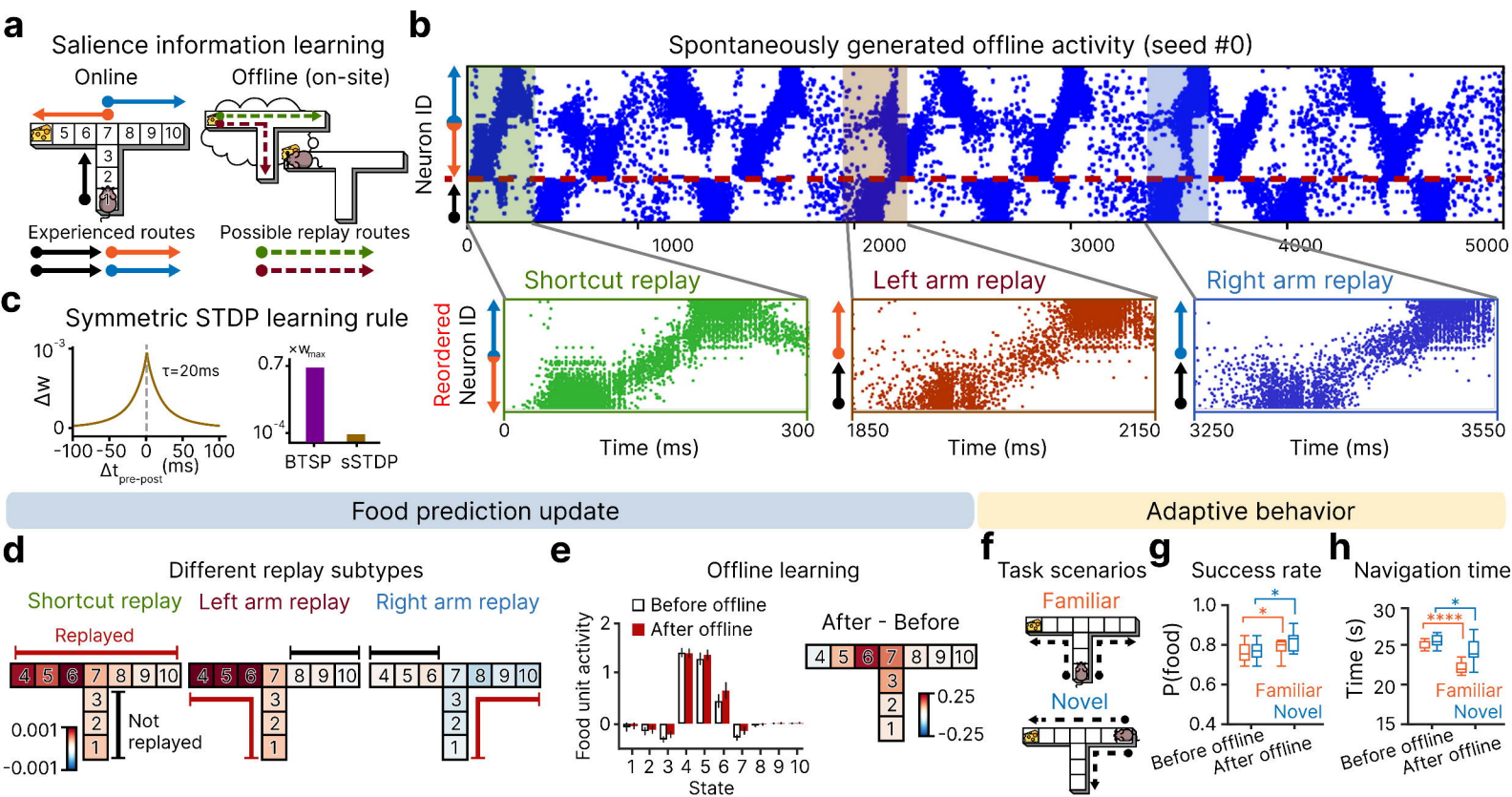

a
Fear conditioning
Lap 1-3 (familiarization)
Lap 4 (shock)
b
Footshock implementation
Movement speed
Lap 4
Speed (s/state)
Shock
State along lap
Motivational Importance
Shock
10
c
Perceived salience
Lap 1
Lap 2
Lap 3
Familiarization
Lap 4
Shock
Avg. PS
State
n.s.
****
Famil.
Shock
d
Place cells (seed #2)
Neuron
State
e
Offline sleep replay
f
Simulated off-site offline activity (seed #7)
Neuron ID
Time (ms)
g
Shock prediction improvement
Without offline
With offline
Shock unit activity
State
h
Markov transition matrix (after lap 4)
Without offline
With offline
Current state
Future state
i
Example movement trajectory
Lap 3
Lap 4
Lap 4+Offline
Time (s)
State
j
Near-shock occupancy
Occupancy (%)
Lap
3
4
4+off
k
Forward roll-out (state 4)
l
Spontaneous on-site offline activity (seed #1)
Neuron ID
Time (ms)
m
Shock info. access
Unit activity
Time (ms)


a
Decision making task
Should I move left or right ?
Factors to consider
1. How likely are the features?
2. How strong are those features?
3. How motivated am I?
b
SR vs. SF
Given: at
p(visit)
state:
p(occur)
feature:
c
Theoretical framework
PS-weighted SR
M^{PS}
Current state
Future state
$W_{\psi, M^{PS}}$
SF
ψ( ) ψ( )
Value estimation
v(or) V(or)
+ for
- for
Decision
Move Right
d
Hippocampal implementation
Place input
Activity
Cell #1 #2 #3
Position
Perceived salience
Intensity × Motivational importance × Novelty
Value of features
+5 for
-10 for
W_{CA3}
W_{CA1}
W_{FP}
V(State)
CA3
Place representation
CA1
PS-weighted SR
Feature prediction
SF
Behavioral Decision
e
Online learning
Position
Input
Time
CA3
CA1
Feat. pred.
f
Offline learning
Off-site
On-site
Time
Time
CA3
CA1
Feat. pred.
g
Behavior
Next step?
Input
Output
0.1
0.3
0.9
Move Right
CA3
CA1
Feat. pred.
Training rule:
BTSP
Delta rule
STDP
Not trained


a
Linear treadmill
State-specific backgrounds
8 1 2 3 4 5 6 7 8 1
Lap
b
BTSP kernels
(Li et al., 2024)
W_{CA3}
W_{CA1}
Δw
Δt_{PP-AP}(s)
c
Learned weight matrices
W_{CA3}
W_{CA1}
Pre- neuron
Post- neuron
d
Symmetry of weights
W_{CA3}
W_{CA1}
$\Delta t_{post-pre}$(s)
Symmetry index
****
CA3 network
e
Off-site offline simulation
Time
CA3
CA1
f
Spontaneous offline activity (seed #1)
Neuron ID
Time (ms)
g
Example replay
Neuron ID
Time (ms)
Position (state)
Time (ms)
CA1 network
h
Emergence of SR
CA3
W_{CA1}
CA1
CA1 Activity
Place cell
i
Emerged place cell
Place cell count
Lap
j
CA1 cell activity
(seed #1)
Neuron
State
k
Asymmetric place field
Approach
Retreat
$y=a\gamma^{|x|}+b$
Firing rate (Hz)
Time (s)
Discount factor (γ)
****
Appraoch
Retreat
Feature prediction network
l
Emergence of SF
Which features am I seeing?
W_{FP}
Feat. pred. Activity
Feature unit
m
Learned W_{FP}
Background
CA1 cells (state)
n
Background prediction
BG 1
Unit activity
State
o
Successor feature prediction field
Approach
Retreat
$y=a\gamma^{|x|}+b$
Unit activity
Times (s)
Discount factor (γ)
*
Appraoch
Retreat


a
Reward exposure
Familiarization
Lap 1-5
Initial reward learning
Lap 6-10
Relocated reward learning
Lap 11-15
b
Reward implementation
Movement speed
Lap 6-10
Lap 11-15
Speed (s/state)
RW Presence
State along the lap
Motivational Importance
BG
1.0
Food
3.0
c
Perceived salience
State 4
State 8
Others
Lap 5
Lap 6
Lap 11
Avg(PS)
State
Lap
n.s.
****
d
CA3-CA3 network
After lap 10
After lap 15
Pre- neuron
Post- neuron
e
CA3-CA1 network
f
Adaptive over-representation
(seed #5)
Neuron ID
State
g
Distribution of place cells
Portion(place cell)
Lap 10
Lap 5
Lap 15
***
**
*
h
Reward learning
Novel
Familiar
Just found a new !
I knew the was here
Lap | Novel | Familiar
Init. | 6 | 10
Reloc. | 11 | 15
i
Population activity
Novel
Familiar
Init. RW
Reloc. RW
BG
Before
Exposure lap
j
Fast reward learning
Prediction error
k
Initial reward learning
RW unit activity
Lap 10 (Familiar)
Lap 6 (Novel)
l
Adaptation to change
Lap 11 (Novel)
Lap 15 (Familiar)


a
T-maze exploration
Odd lap:
Even lap:
b
Reward introduction
Familiarization
Lap 1-6
Reward
Lap 7-10
c
Learned weight matrix
W_{CA3}
W_{CA1}
Pre- neuron
Post- neuron
d
Reward prediction unit activity
Food
Water
e
Decision making based on motivation
Hungry?
Thirsry?
f
Behavior simulation
Next step?
Value (state)
V(L) V(C) V(R)
Probability (action)
P_{left} P_{stay} P_{right}
Behavior
g
Markov transition matrix based on motivational state
Motivation
Current state
Future state
h
Examples of simulated behavioral trajectory
Time (s)
i
Final destination
Probability
Food
Water
Neither
Goal
n.s.
j
Time taken to reach food
Time (s)
Motivation (food)


a
Salience information learning
Online
Offline (on-site)
Experienced routes
Possible replay routes
b
Spontaneously generated offline activity (seed #0)
Neuron ID
Shortcut replay
Left arm replay
Right arm replay
Reordered Neuron ID
Time (ms)
c
Symmetric STDP learning rule
τ=20ms
Δw
BTSP
sSTDP
Food prediction update
Adaptive behavior
d
Different replay subtypes
Shortcut replay
Left arm replay
Right arm replay
Replayed
Not replayed
e
Offline learning
Before offline
After offline
Food unit activity
State
After - Before
f
Task scenarios
Familiar
Novel
g
Success rate
P(food)
Before offline
After offline
h
Navigation time
Time (s)