Rapid learning of predictive maps with STDP and theta phase precession

Tom M George*✉1, William de Cothi*2, Kimberly Stachenfeld3, and Caswell Barry✉2

[1]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London, UK
[2]Research Department of Cell and Developmental Biology, University College London, London, UK
[3]DeepMind, London, UK
*denotes equal contribution
✉Correspondence to tom.george.20@ucl.ac.uk and caswell.barry@ucl.ac.uk

April 20, 2022

Abstract

The predictive map hypothesis is a promising candidate principle for hippocampal function. A favoured formalisation of this hypothesis, called the successor representation, proposes that each place cell encodes the expected state occupancy of its target location in the near future. This predictive framework is supported by behavioural as well as electrophysiological evidence and has desirable consequences for both the generalisability and efficiency of reinforcement learning algorithms. However, it is unclear how the successor representation might be learnt in the brain. Error-driven temporal difference learning, commonly used to learn successor representations in artificial agents, is not known to be implemented in hippocampal networks. Instead, we demonstrate that spike-timing dependent plasticity (STDP), a form of Hebbian learning, acting on temporally compressed trajectories known as "theta sweeps", is sufficient to rapidly learn a close approximation to the successor representation. The model is biologically plausible – it uses spiking neurons modulated by theta-band oscillations, diffuse and overlapping place cell-like state representations, and experimentally matched parameters. We show how this model maps onto known aspects of hippocampal circuitry and explains substantial variance in the true successor matrix, consequently giving rise to place cells that demonstrate experimentally observed successor representation-related phenomena including backwards expansion on a 1D track and elongation near walls in 2D. Finally, our model provides insight into the observed topographical ordering of place field sizes along the dorsal-ventral axis by showing this is necessary to prevent the detrimental mixing of larger place fields, which encode longer timescale successor representations, with more fine-grained predictions of spatial location.

1 Introduction

Knowing where you are and how to navigate in your environment is an everyday existential challenge for motile animals. In mammals, a key brain region supporting these functions is the hippocampus [1, 2], which represents self-location through the population activity of place cells – pyramidal neurons with spatially selective firing fields [3]. Place cells, in conjunction with other spatially tuned neurons [4, 5], are widely held to constitute a "cognitive map" encoding information about the relative location of remembered locations and providing a basis upon which to flexibly navigate [6, 7].

The hippocampal representation of space incorporates spike time and spike rate based encodings, with both components conveying broadly similar levels of information about self-location [8, 9]. Thus, the position of an animal in space can be accurately decoded from place cell firing rates [10] as well as from the precise time of these spikes relative to the background 8-10Hz theta oscillation in the hippocampal local field potential [9]. The latter is made possible since place cells have a tendency to spike progressively earlier in the theta cycle as the animal traverses the place field - a phenomena known as phase precession [11]. Therefore, during a single cycle of theta the activity of the place cell population smoothly sweeps from representing the past to representing the future position of the animal [12], and can simulate alternative possible futures across multiple cycles [13].

In order for a cognitive map to support planning and flexible goal-directed navigation it should incorporate information about the overall structure of space and the available routes between locations [6, 7]. Theoretical work has identified the regular firing patterns of entorhinal grid cells with the former role, providing a spatial metric sufficient to support the calculation of navigational vectors [14, 15]. In contrast, associative place cell - place cell interactions have been repeatedly highlighted as a plausible mechanism for learning the available transitions in an environment [16, 17, 18]. In the hippocampus, such associative learning has been shown to follow a spike-timing dependent plasticity (STDP) rule [19] – a form of Hebbian learning where the temporal ordering of spikes between presynaptic and postsynaptic neurons determines whether long-term potentiation or depression occurs. One of the consequences of phase precession is that correlates of behaviour, such as position in space, are compressed onto the timescale of a single theta cycle and thus coincide with the time-window of STDP $\mathcal{O}(20-50\text{ ms})$ [8]. As such, the combination of theta sweeps and STDP potentially provides an efficient mechanism to learn from an animal's experience – forming associations between cells which are separated by behavioural timescales much larger than that of STDP.

Spatial navigation can readily be understood as a reinforcement learning problem - a framework which seeks to define how an agent should act to maximise future expected reward [20]. Conventionally the value of a state is defined as the expected cumulative reward that can be obtained from that location with some temporal discount applied. Thus, the relationship between states and the rewards expected from those states are captured in a single value which can be used to direct reward-seeking behaviour. However, the computation of expected reward can be decomposed into two components – the successor representation, a predictive map capturing the expected location of the agent discounted into the future, and the expected reward associated with each state [21]. This segregation yields several advantages since information about available transitions can be learnt independently of rewards and thus changes in the locations of rewards do not require the value of all states to be re-learnt.

A growing body of empirical and theoretical evidence suggests that the hippocampal spatial code functions as a successor representations [22]. Specifically, that the activity of hippocampal place cells encodes a predictive map over the locations the animal expects to occupy in the future. Notably, this framework accounts for phenomena such as the skewing of place fields due to stereotyped trajectories [18], the reorganisation of place fields following a forced detour [23], and the behaviour of humans and rodents whilst navigating physical, virtual and conceptual spaces [24, 25]. However, the successor representation is typically conceptualised as being learnt using the temporal difference learning rule [26, 27], which uses the prediction error between expected and observed experience to improve the predictions. Whilst correlates of temporal difference learning have been observed in the striatum during reward-based learning [28], it is less clear how it could be implemented in the hippocampus to learn a predictive map. In this context, we hypothesised that the predictive and compression properties of theta sweeps, combined with STDP in the hippocampus, might be sufficient to approximately learn a successor representation.

We simulated the synaptic weights learnt due to STDP between a set of synthetic spiking place cells and compared them to the weights of a successor representation learnt with temporal difference learning. We found that the inclusion of theta sweeps with the STDP rule increased the efficiency and robustness of the learning, with the STDP weights being a close approximation to the true successor matrix. Further, we find no fine tuning of parameters is needed - biologically determined parameters are optimal to efficiently approximate a successor representation and replicate experimental results synonymous with the predictive map hypothesis, including the behaviourally biased skewing of place fields [18, 22]. Finally, we use the simulation of STDP with theta sweeps to generate insight into the observed topographical ordering of place field sizes along the dorsal-ventral hippocampal axis [29], by observing that such organisation is necessary to prevent the detrimental mixing of larger place fields, which approximate longer timescale successor representations [30], with more fine-grained predictions of future spatial location.

2 Results

We set out to investigate whether a combination of STDP and phase precession is sufficient to generate a successor representation-like matrix of synaptic weights between place cells in CA3 and downstream CA1. The model comprises of an agent exploring a maze where its position $\mathbf{x}(t)$ is encoded by the instantaneous firing of a population of N CA3 basis features, each with a spatial receptive field $f^x_j(\mathbf{x})$ given by a thresholded Gaussian of radius 1 m and 5 Hz peak firing rate. As the agent traverses the receptive field, its rate of spiking is subject to phase precession $f^\theta_j(\mathbf{x},t)$ with respect to a 10 Hz theta oscillation (see Methods and Fig. 1a) such that, in total, the instantaneous firing rate of the j^{th} basis features is given by:

$$f_j(\mathbf{x},t) = f^x_j(\mathbf{x}) f^\theta_j(\mathbf{x},t). \tag{1}$$

CA3 basis features f_j then linearly drive downstream CA1 'STDP successor features' $\tilde{\psi}_i$ (Fig. 1b)

$$\tilde{\psi}_i(\mathbf{x},t) = \sum_j \mathsf{W}_{ij} f_j(\mathbf{x},t). \tag{2}$$

Using an inhomogeneous Poisson process, the firing rates of the basis and STDP successor features are converted into spike trains which cause learning in the weight matrix W_{ij} according to an STDP rule (see Methods and Fig. 1c). The synaptic weight matrix W_{ij} (Fig. 1d) can then be directly compared to the true successor matrix M_{ij} (Fig. 1e), learnt via temporal difference (TD) learning on the CA3 basis features (the full learning rule is derived in Methods and shown in Eqn. 25). Further, the successor matrix M_{ij} can also be used to generate the true 'TD successor features':

$$\psi_i(\mathbf{x}) = \sum_j \mathsf{M}_{ij} f^x_j(\mathbf{x}), \tag{3}$$

allowing for direct comparison and analyses with the STDP successor features $\tilde{\psi}_i$, (Eqn. 2).

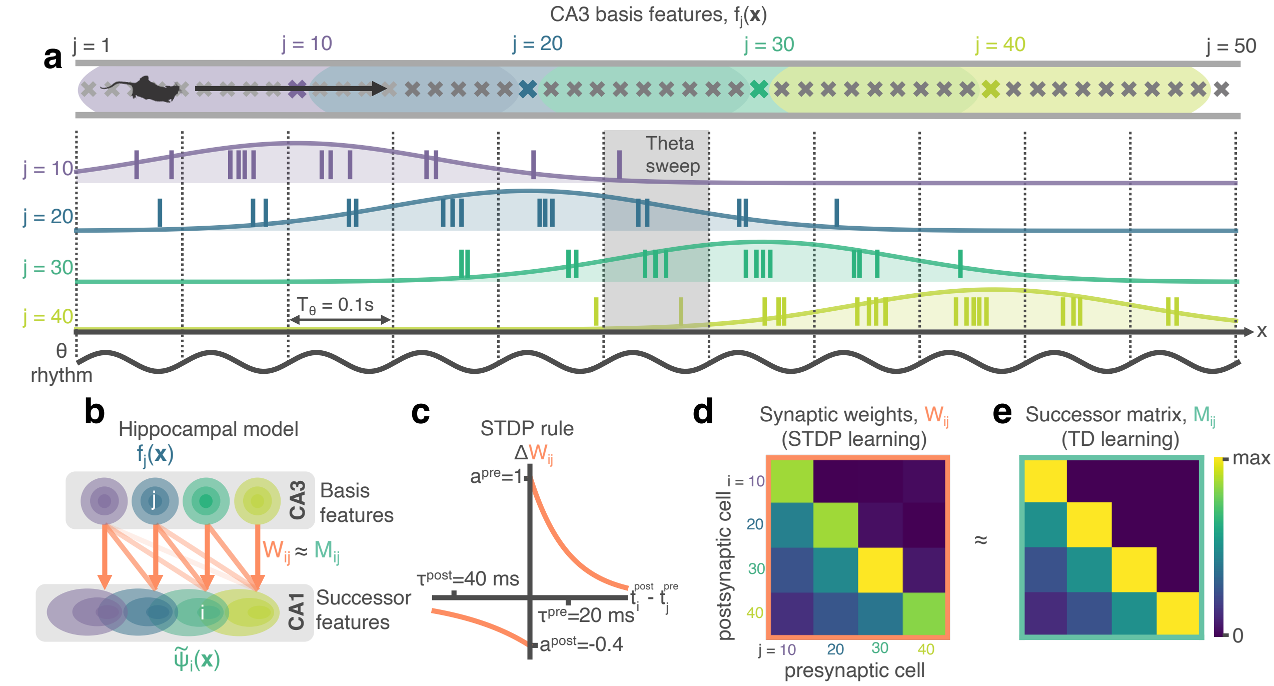


Figure 1: STDP between phase precessing place cells produces successor representation-like weight matrices. **a** Schematic of an animal running left-to-right along a track. 50 cells phase precess, generating theta sweeps (e.g. grey box) that compress spatial behaviour into theta timescales (10 Hz). **b** We simulate a population of CA3 'basis feature' place cells which linearly drive a population of CA1 'STDP successor feature' place cells through the synaptic weight matrix W_{ij}. **c** STDP learning rule; pre-before-post spike pairs $(t_i^{\text{post}} - t_j^{\text{pre}} > 0)$ result in synaptic potentiation whereas post-before-pre pairs $(t_i^{\text{post}} - t_j^{\text{pre}} < 0)$ result in depression. Depression is weaker than potentiation but with a longer time window, as observed experimentally. **d** Simplified schematic of the resulting synaptic weight matrix, W_{ij}. Each postsynaptic cell (row) fires just after, and therefore binds strongly to, presynaptic cells (columns) located to the left of it on the track. **e** Simplified schematic of the successor matrix (Eqn. 3) showing the synaptic weights after training with a temporal difference learning rule, where each CA1 cell converges to represent the successor feature of its upstream basis feature. Backwards skewing (successor features "predict" upcoming activity of their basis feature) is reflected in the asymmetry of the matrix, where more activity is in the lower triangle, similar to panel d.

2.1 The synaptic weight matrix closely approximates the TD successor matrix

We first simulated an agent with $N = 50$ evenly spaced CA3 place cell basis features on a 5 m circular track (linear track with circular boundary conditions to form a closed loop, Fig. 2a). The agent moved left-to-right at a constant velocity for 30 minutes, performing ∼58 complete traversals of the loop. The STDP weights learnt between the phase precessing basis features and their downstream STDP successor features (Fig. 2b) were markedly similar to the successor representation matrix generated using temporal difference learning applied to the same basis features under the same conditions (Fig. 2c, element-wise Pearson correlation between matrices $R^2 = 0.87$). In particular, the agent's strong left-to-right behavioural bias led to the characteristic asymmetry in the STDP weights predicted by successor representation models [22], with both matrices dominated by a wide band of positive weight shifted left of the diagonal and negative weights shifted right.

To compare the structure of the STDP weight matrix W_{ij} and TD successor matrix M_{ij}, we aligned each row on the diagonal and averaged across rows (see Methods), effectively calculating the mean distribution of learnt weights originating from each basis feature (Fig. 2d). Both models exhibited a similar distribution, with values smoothly ramping up to a peak just left of centre, before a sharp drop-off to the right caused by the left-to-right bias in the agent's behaviour. In the successor representation this is because CA3 place cells to the left of (i.e. preceding) a given basis feature are reliable predictors of that basis feature's future activity, with those immediately preceding it being the strongest predictors and thus conferring the strongest weights to its successor feature. Conversely, the CA3 place cells immediately to the right of (i.e. after) this basis feature are the furthest they could possibly be from predicting its future activity, resulting in minimal weight contributions. Indeed, we observed some of these weights even becoming negative (Fig. 2d) – necessary to approximate the sharp drop-off in predictability using the smooth Gaussian basis features. With the STDP model, the similar distribution of weights is caused by the asymmetry in the STDP learning rule combined with the consistent temporal ordering of spikes in a theta sweep. Hence, the sequence of spikes emitted by different cells within a theta cycle directly reflects the order in which their spatial fields are encountered, resulting in commensurate changes to the weight matrix. So, for example, if a postsynaptic neuron reliably precedes its presynaptic cell on the track, the corresponding weight will be reduced, potentially becoming negative.

Notably, the temporal compression afforded by theta phase precession, which brings behavioural effects into the millisecond domain of STDP, is an essential element of this process. When phase precession was removed from the STDP model, the resulting weights failed to capture the expected behavioural bias and thus did not resemble the successor matrix - evidenced by the lack of asymmetry (Fig. 2d, dashed line; ratio of mass either side of y-axis 4.54 with phase precession vs. 0.99 without) and a decrease in the explained variance of the TD successor matrix (Fig 2e, $R^2 = 0.87 \pm 0.01$ vs $R^2 = 0.63 \pm 0.02$ without phase precession). Similarly, without the precise ordering of spikes, the learnt weight matrix was less regular, having increased levels of noise, and took approximately twice as long to converge to a stable state (Fig. 2e; time to reach 75% percent of final R^2: 5 vs. 10.5 minutes without phase precession).

We also conducted a hyperparameter sweep to test if these results were robust to changes in the STDP learning rule parameters (Fig. S1). The sweep range for each parameter contained and extended beyond the "biologically plausible" values used in this paper (Fig. S1a). We found that optimised parameters (those which result in the highest final similarity between STDP and TD weight matrices, W_{ij} and M_{ij}) were very close to the biological parameters already selected for our model from a literature search (Fig. 1c & d, parameter references also listed in figure) and, when they were used, no drastic improvement was seen in the similarity between W_{ij} and M_{ij}. The only exception was firing rate for which performance monotonically improved as it increased - something the brain likely cannot achieve due to energy constraints.

Next, we examined the correspondence between our model and the successor representation in a situation without a strong behavioural bias. Thus, we reran the simulation on the linear track without the circular boundary conditions so the agent turned and continued in the opposite direction whenever it reached each end of the track (Fig. 2f). Again, the STDP and TD successor representation weight matrices where remarkably similar ($R^2 = 0.88$; Fig. 2gh) both being characterised by a wide band of positive weight centred on the diagonal (Fig. 2i) - reflecting the directionally unbiased behaviour of the agent. In this unbiased regime, theta sweeps were less important though still confer a shape, learning speed, and signal-strength advantage over the non-phase precessing model (Fig. 2j) - evidenced as an increased amount of explained variance ($R^2 = 0.88 \pm 0.01$ vs. $R^2 = 0.76 \pm 0.02$) as well as the time taken to reach 75% of this value (6.5 vs 10 minutes).

To test if the STDP model's ability to capture the successor matrix would scale up to open field spaces, we implemented a 2D model of phase precession (see Methods) where the phase of spiking is sampled according to the distance travelled through the place field along the chord currently being traversed [31]. We then simulated both the agent in an environment consisting of two interconnected 2.5 × 2.5 m square rooms (Fig. 2k) using an adapted policy modelling rodent foraging behaviour that is biased towards traversing doorways and following walls [32] (see Methods and 10 minute sample trajectory shown in Fig. 2k). After training for 2 hours of exploration, we found that the combination of STDP and phase precession was able to successfully capture the structure in the TD successor matrix (Fig. 2l-m, $R^2 = 0.74$, TD successor matrix calculated over the same 2 hour trajectory).

2.2 Theta sequenced STDP place cells show behaviourally-biased skewing, a hallmark of successor representations

We next wanted to investigate how the similarities in weights between the STDP and TD successor representation models are conveyed in the downstream CA1 successor features. One hallmark of the successor representation is that strong biases in behaviour (for example, travelling one way round a circular track) induce a reliable predictability of upcoming future locations, which in turn causes a backward skewing in the resulting successor features [22]. Such skewing, opposite to the direction of travel, has also been observed in hippocampal place cells [18]. Under strongly biased behaviour on the circular linear track, the biologically plausible STDP CA1 successor features (Eqn. 2) had a very high correlation with the TD successor features (Eqn. 3) predicted by successor theory (Fig. 3a; $R^2 = 0.98 \pm 0.01$). Both exhibited a pronounced backward skew, opposite to the direction of travel (mean TD vs. STDP successor feature skewness: $= -0.39 \pm 0.01$ vs. $= -0.24 \pm 0.07$). Furthermore, both the STDP and TD successor representation models predict that such biased behaviour should induce a backwards shift in the location of place field peaks (Fig. 3a left panel; TD vs. STDP successor feature shift in metres: -0.28 ± 0.00 vs -0.38 ± 0.03) – this phenomenon is also observed in the hippocampal place cells [18], and our model accounts for the observation that more shifting and skewing in observed in CA1 place cells than CA3 place cells[33]. As expected, when theta phase precession was removed from the model no significant skew or shift was observed in the STDP successor features. Similarly, the skew in field shape and shift in field peak were not present when the behavioural bias was removed (Fig. 3b) – in this unbiased scenario, the advantage of the STDP model with theta phase precession was modest relative to the same model without phase precession ($R^2 = 0.99 \pm 0.01$ vs. $R^2 = 0.96 \pm 0.01$).

Examining the activity of CA1 cells in the two-room open field environment, we found an increase in the eccentricity of fields close to the walls (Fig. 3c & d; average eccentricity of STDP successor features near vs. far from wall: 0.57 ± 0.06 vs. 0.33 ± 0.07). In particular, this increased eccentricity is facilitated by a shorter field width along the axis perpendicular to the wall (Fig. 3e), an effect observed experimentally in rodent place cells [34]. This increased eccentricity of cells near the wall remained when the behavioural bias to follow walls was

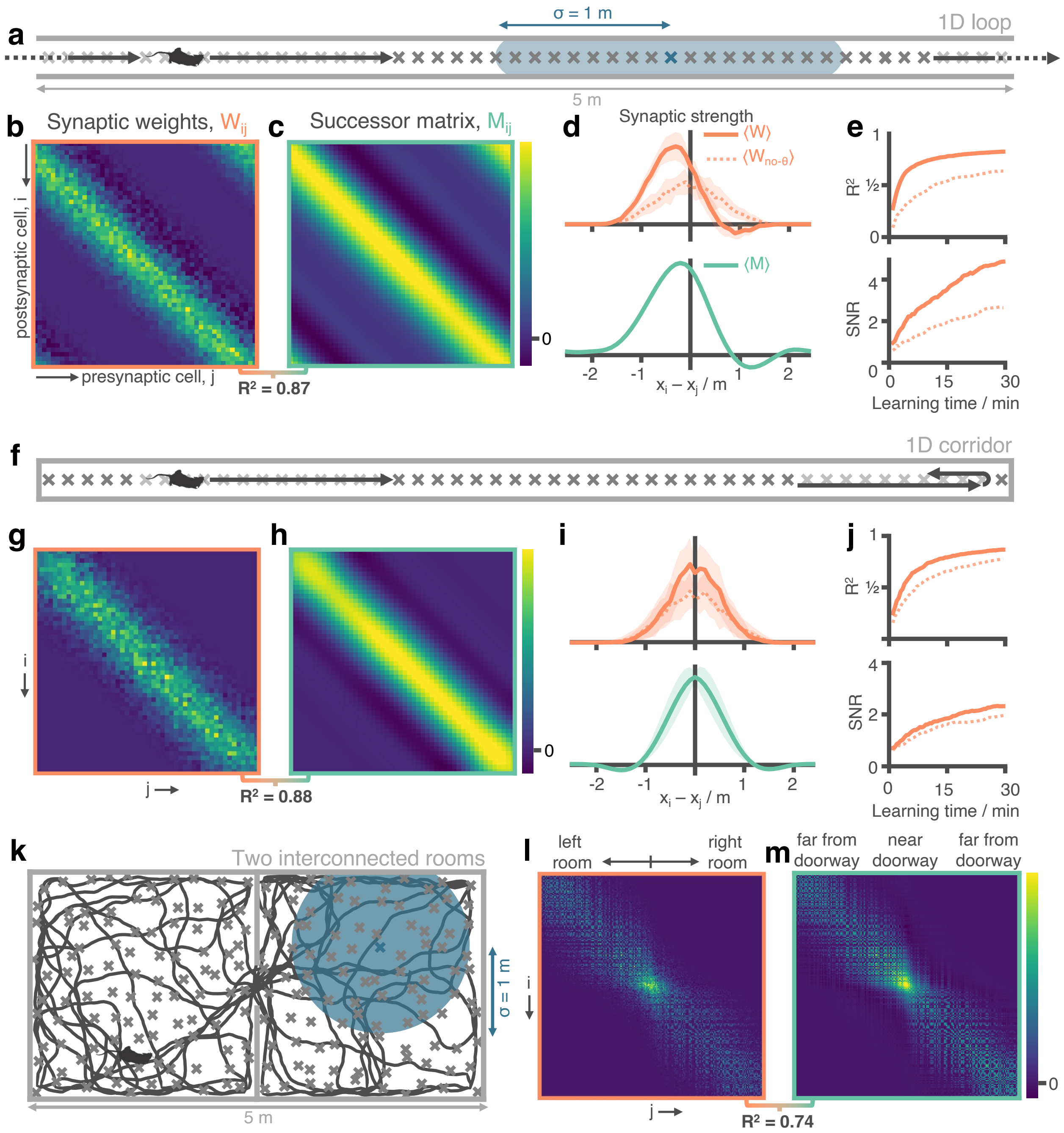


Figure 2: Successor matrices are rapidly approximated by STDP applied to spike trains of phase precessing place cells. **a** Agents traversed a 5 m circular track in one direction (left-to-right) with 50 evenly distributed CA3 spatial basis features (example thresholded Gaussian place field shown in blue, radius $\sigma = 1$ m). **b&c** After 30 minutes, the synaptic weight matrix learnt between CA3 basis features and CA1 successor features strongly resembles the equivalent successor matrix computed by temporal difference learning. Rows correspond to CA1, columns to CA3. **d** To compare the distribution of weights, matrix rows were aligned on the diagonal and averaged over rows (mean ± standard deviation shown). **e** Against training time, we plot (top) the R^2 between the synaptic weight matrix and successor matrix and (bottom) the signal-to-noise ratio of the synaptic matrix. **f-j** Same as panels a-e except the agent turns around at each end of the track. The average policy is now unbiased with respect to left and right, as can be seen in the diagonal symmetry of the matrices. **k-m** As in panels a-c except the agent explores a two dimensional maze where two rooms are joined by a doorway. The agent follows a random trajectory with momentum and is biased to traverse doorways and follow walls.

removed (Fig. 3d; average eccentricity with vs. without wall bias: 0.57 ± 0.06 vs. 0.54 ± 0.06), thus indicating it is primarily caused by the inherent bias imposed on behaviour by extended walls rather than an explicit policy bias. Note that our ellipse fitting algorithm accounts for portions of the field that have been cut off by environmental boundaries (see methods & Fig. 3c), and so this effect is not simply a product of basis features being occluded by walls.

In a similar fashion, the bias in the motion model we used - which is predisposed to move between the two rooms - resulted in a shift in STDP successor feature peaks towards the doorway (Fig. 3f & g; inwards shift in metres for STDP successor features near vs. far from doorway: 0.15 ± 0.06 vs. 0.04 ± 0.05; with doorway bias turned off: 0.05 ± 0.08 vs. 0.04 ± 0.05). At the level of individual cells this was visible as an increased propensity for fields to extend into the neighbouring room after learning (Fig. 3h). Hence, although basis features were initialised as two approximately non-overlapping populations – with only a small proportion of cells near the doorway extending into the neighbouring room – after learning many cells bind to those on the other side of the doorway, causing their place fields to diffuse through the doorway and into to the other room (Fig. 3f). This shift could partially explain why place cell activity is found to cluster around doorways [35] and rewarded locations [36] in electrophysiological experiments. Equally it is plausible that a similar effect might underlie experimental observations that neural representations in multi-compartment environments typically begin heavily fragmented by boundaries and walls but, over time, adapt to form a smooth global representations (e.g. as observed in grid cells by [37]).

2.3 Multiscale successor representations are stored along the hippocampal dorsal-ventral axis by populations of differently sized place cells

Finally we wanted to investigate whether the STDP learning rule was able form successor representation-like connections between basis features of different scales. Recent experimental work has highlighted that place fields form a multiscale representation of space, which is particularly noticeable in larger environments [34, 38], such as the one modelled here. Such multiscale spatial representations have been hypothesised to act as a substrate for learning successor features with different time horizons – large scale place fields are able to make predictions of future location across longer time horizons, whereas place cells with smaller fields are better placed to make temporally fine-grained predictions. Agents could use such a set of multiscale successor features to plan actions at different levels of temporal abstraction, or predict precisely which states they are likely to encounter soon [30]. Despite this, what is not known is whether different sized place fields will form associations when subject to STDP coordinated by phase precession and what effect this would have on the resulting successor features. Consider a small basis feature cell with a receptive field entirely encompassed by that of a larger basis cell. A consequence of theta phase precession is that the cell with the smaller field will phase precess faster through the theta cycle than the other cell - initially it will fire later in the theta cycle than the cell with a larger field, but as the animal moves towards the end of the small basis field it will fire earlier. These periods of potentiation and depression instigated by STDP will act against each other and cancel to an extent that depends on the relative placement of the two fields, their size difference, and the parameters of the learning rule. To test this, we simulated an agent, learning according to our STDP model in the circular track environment, with, simultaneously, three sets of differently sized basis features ($\sigma = 0.5$, 1.0 and 1.5 m, Fig. 4a). Such ordered variation in field size has been observed along the dorso-ventral axis of the hippocampus ([29, 39]; Fig. 4b), and has been theorised to facilitate successor representation predictions across multiple time-scales [22, 30].

When we trained the STDP model on a population of homogeneously-distributed multiscale basis features, the resulting weight matrix displayed binding across the different sizes (Fig. 4c top). This in turn leads to a population of downstream successor features with the same redundantly large scale (Fig. 4c bottom). The negative interaction between different sized fields was not sufficient to prevent binding and, as such, the place fields of small features are dominated by contributions from bindings to larger basis features. Conversely, when these multiscale basis features were ordered along the dorso-ventral axis to prevent binding between the different scales – cells of the three scales were processed separately, Fig. 4d top) – the multiscale structure is preserved in the resulting successor features (Fig. 4d bottom). We thus propose that place cell size can act as a proxy for the predictive time horizon, τ – also called the discount parameter, $\gamma = e^{-\frac{dt}{\tau}}$, in discrete Markov Decision Processes. However for this effect to be meaningful, plasticity between cells of different scales must be minimised to prevent short timescales from being overwritten by longer ones, this segregation may plausibly be achieved by the observed size ordering along the hippocampal dorsal-ventral axis.

3 Discussion

Successor representations store long-run transition statistics and allow for rapid prediction of future states [21] - they are hypothesised to play a central role in mammalian navigation strategies [22, 27]. We show that Hebbian learning between spiking neurons, resembling the place fields found in CA3 and CA1, learns an accurate approximation to the successor representation when these neurons undergo phase precession with respect to the hippocampal theta rhythm. The approximation achieved by STDP explains a large proportion of the variance

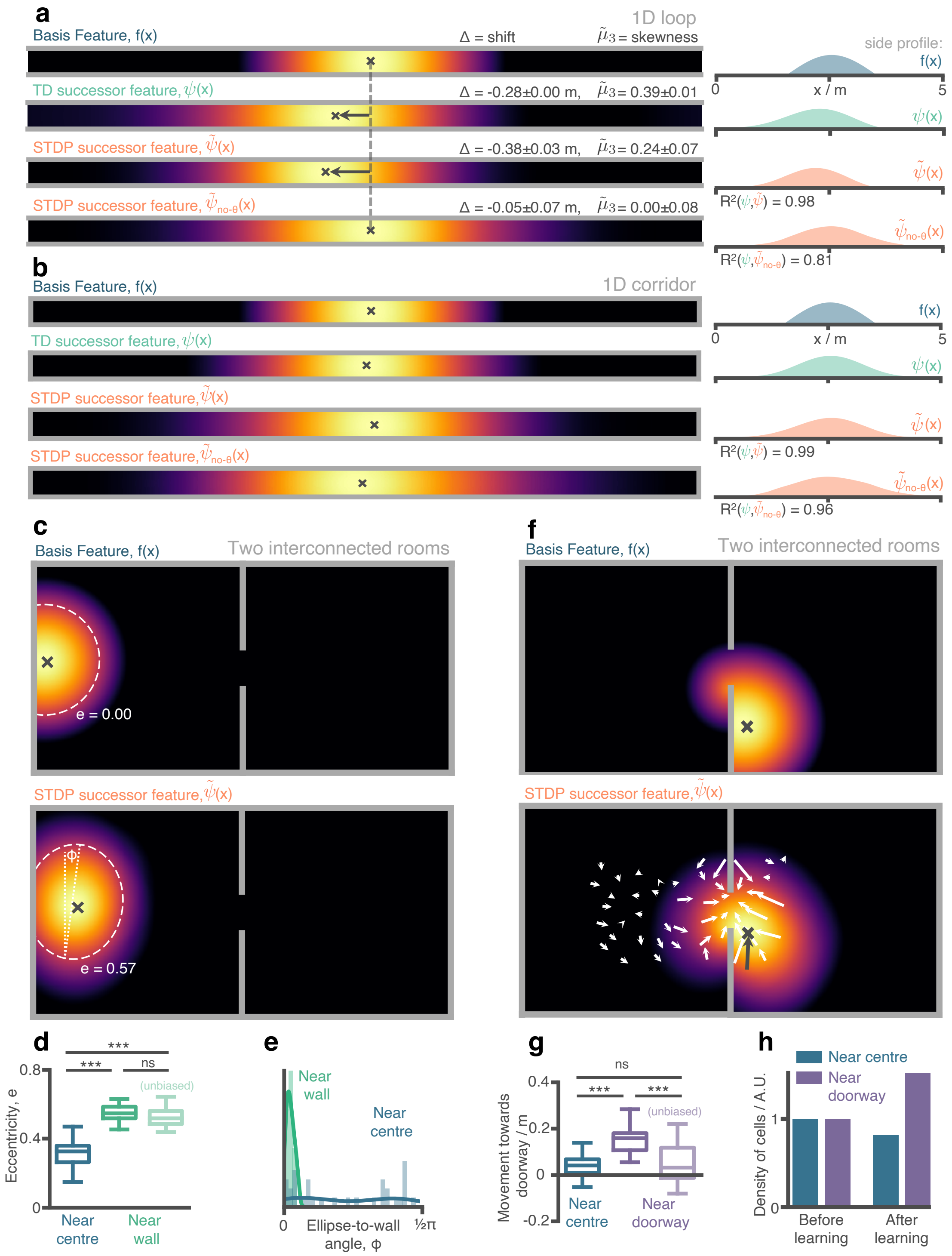


Figure 3: Place cells (aka. successor features) in our STDP model show behaviourally biased skewing resembling experimental observations and successor representation predictions. **a** In the loop maze (motion left-to-right) STDP place cells skew and shift backwards, and strongly resemble place cells obtained via temporal difference learning. This is not the case when theta phase precession is absent. **b** In the corridor maze, where travel in either direction is equally likely, place fields defuse in both directions due to the unbiased movement policy. **c** In the 2D maze, place cells (of geodesic Gaussian basis features) near the wall elongate along the wall axis (dashed line shows best fitting ellipse, angle construct show the ellipse-to-wall angle). **d** Place cells near walls have higher elliptical eccentricity than those near the centre of the environments. This increase remains even when the movement policy bias to follow walls is absent. **e** The eccentricity for fields near the walls is facilitated by an increase in the length of the place field along an axis parallel to the wall (ϕ close to zero). **f** Place cells near the doorway cluster towards it and expand through the doorway relative to their parent basis features. **g** The shift of place fields near the doorway towards the doorway is significant relative to place fields near the centre and disappears when the behavioural bias to cross doorways is absent. **h** The shift of place fields towards the doorway manifests as an increase in density of cells near the doorway after exploration.

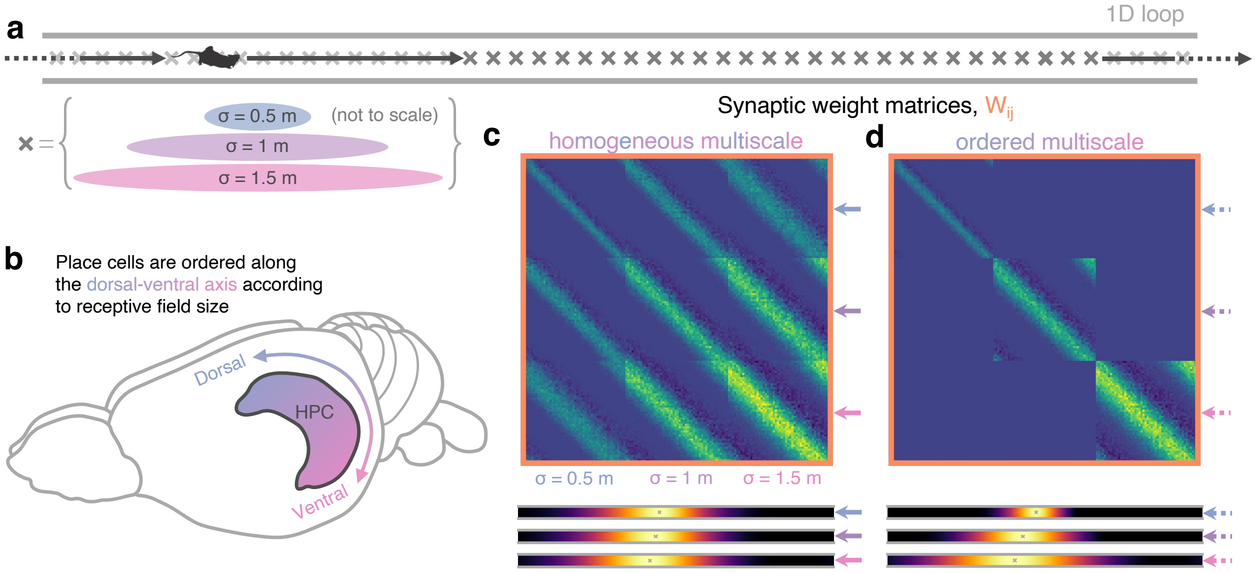


Figure 4: Multiscale successor representations are stored by place cells with multi-sized place fields but only when sizes are segregated along the dorso-ventral axis. **a** An agent explores a 1D loop maze with 150 places cells of different sizes (50 small, 50 medium, and 50 large) evenly distributed along the track. **b** In rodent hippocampus place cells are observed to be ordered along the dorso-ventral axis according to their field size. **c** When cells with different field sizes are homogeneously distributed throughout hippocampus all postsynaptic successor features can bind to all presynpatic basis features, regardless of their size (top). Short timescale successor representations are overwritten, creating three equivalent sets of redundantly large scale successor features (bottom). **d** Ordering cells creates a physical barrier; postsynaptic successor features can only bind to basis features in the same size range (off-diagonal block elements are zero) preventing cells with different size fields from binding. Now, three dissimilar sets of successor features emerge with different length scales, corresponding to successor features of different discount time horizons.

in the true successor matrix and replicates hallmarks of successor representations such as behaviourally biased place field skewing, elongation of place fields near walls, and clustering near doorways in both one and two-dimensional environments.

Theta phase precession has a dual effect not only *allowing* learning by compressing trajectories to within STDP timescales but also *accelerating* convergence to a stable representation by arranging the spikes from cells along the current trajectory to arrive in the order those cells are actually encountered. Without theta phase precession, STDP fails to learn a successor representation reflecting the current policy unless that policy is approximately unbiased. Further, by instantiating a population of place cells with multiple scales we show that topographical ordering of these place cells by size along the dorso-ventral hippocampal axis is a necessary feature to prevent small discount timescale successor representations from being overwritten by longer ones. Last, performing a grid search over STDP learning parameters, we show that those values selected by evolution are approximately optimal for learning successor representations. This finding is compatible with the idea that the necessity to rapidly learn predictive maps by STDP has been a primary factor driving the evolution of synaptic learning rules in hippocampus.

Our model is biologically plausible and extends previous work – which required successor features to recursively expand in order to make long range predictions (e.g. [40, 41] – by exploiting the existence of temporally compressed theta sweeps [11, 8], allowing place cells with distant fields to bind directly, without intermediaries. This configuration yields several advantages. First, learning with theta sweeps converges considerably faster than without them. Biologically, it is likely that successor feature learning via Hebbian learning alone would be too slow to account for the rapid stabilisation of place cells in new environments – Dong et al. observed place fields in CA1 to increase in width for approximately the first 10 laps around a 3 m track [33]. This timescale is well matched by our model with theta sweeps in which CA1 place cells reach 75% of their final extent after 5 minutes (or 9.6 laps) of exploration on a 5m track but is markedly slower without theta sweeps.

Second, as well as extending previous work to two-dimensional environments and complex movement policies our model also uses realistic population codes of overlapping Gaussian features. These naturally present a hard problem for models of spiking Hebbian learning since, in the absence of theta sweeps, the order in which features are encountered is not encoded reliably in the relative timing or order of their spikes at synaptic timescales. Theta sweeps address this by tending to sequence spikes according to the order in which their originating fields are encountered. Indeed our preliminary experiments show that when theta sweeps are absent the STDP successor features show little similarity to the true successor features. Our work is thus particularly relevant in light of a recent trend to focus on biologically plausible features for reinforcement learning [42, 27].

Our theory makes a clear prediction – if theta sweeps are used to learn a successor representation that guides navigation, then inhibiting them during exposure to a novel environment should impact subsequent navigation. After learning, however, theta sweeps could be removed without detrimental effects. Indeed, experimental work has shown that power in the theta band increases upon exposure to novel environments [43] – our work suggests this is because theta phase precession is critical for learning and updating stored predictive maps for spatial navigation.

That said, in principle, any form of sufficiently ordered and compressed trajectory would allow STDP plasticity to approximate a successor representation. Hippocampal replay is a well documented phenomena where previously experienced trajectories are rapidly recapitulated during sharp-wave ripple events [44], within which spikes show a form of phase precession relative to the ripple band oscillation (150-250Hz) [45]. Thus, our model might explain the abundance of sharp-wave ripples during early exposure to novel environments [46] – when new 'informative' trajectories, for example those which lead to reward, are experienced it is desirable to rapidly incorporate this information into the existing predictive map [47].

The distribution of place cell receptive field size in hippocampus is not homogeneous. Instead, place field size grows smoothly along the longitudinal axis (from very small in dorsal regions to very large in ventral regions). Why this is the case is not clear – our model contributes by showing that, without this ordering, large and small place cells would all bind via STDP, essentially overwriting the short timescale successor representations learnt by small place cells with long timescale successor representations. Topographically organising place cells by size potentially creates a physical barrier to binding between different sizes, preserving the multiscale successor representations. This alternative explanation, that the dorso-ventral size scale axis is needed to learn multiscale successor representations, is compatible with other theoretical accounts for the ordering. Specifically Momennejad and Howard [30] showed that exploiting multiscale successor representations downstream, in order to recover information which is 'lost' in the process of compiling state transitions into a single successor representation, typically requires calculating the derivative of the successor representation with respect to the discount parameter. This derivative calculation is significantly easier if the cells – and therefore the successor representations – are ordered smoothly along the hippocampal axis.

Work in control theory has shown that the difficult reinforcement learning problem of finding an optimal policy and value function for a given environment becomes tractable if the policy is constrained to be near a 'default policy' [48]. When applied to spatial navigation, the optimal value function resembles the value function calculated using a successor representation for the default policy. This solution allows for rapid adaptation to changes in the reward structure since the successor matrix is fixed to the default policy and need not be re-learnt even if the optimal policy changes. Building on this, recent work suggested the goal of hippocampus is not to learn the successor representation for the current policy but rather for a default diffusive policy[49].

Indeed, we found that in the absence of theta sweeps, the STDP rule learns a successor representation close to that of an unbiased policy, rather than the current policy. This is because without theta-sweeps to order spikes along the current trajectory, cells bind according to how overlapping their receptive fields are, that is, according to how close they are under a 'diffusive' policy. In this context it is interesting to note that a substantial proportion of CA3 place cells do not exhibit significant phase precession [11, 31]. We propose these place cells with weak or absent phase precession might plausibly be responsible for learning a policy-independent 'default representation', useful for rapid policy prediction when the reward structure of an environment is changed. Simultaneously, theta precessing place cells may learn a successor representation for the current (potentially biased) policy, in total giving the animal access to both an off-policy-but-near-optimal value function and an on-policy-but-suboptimal value function.

Finally we comment on the approximate nature of the successor representations learnt by our biologically plausible model. The STDP successor features described here are unlikely to converge analytically to the true TD successor features. Potentially this implies that a value function calculated according to Eqn. (29) would not be accurate and may prevent an agent from acting optimally. There are several possible resolutions to this point. First, the successor representation is unlikely to be a self contained reinforcement learning system. In reality it likely interacts with other model-based or model-free systems acting in other brain regions such as nucleus accumbens in striatum [50]. Plausibly errors in the successor features are corrected for by counteracting adjustments in the reward weights implemented by some downstream model free error based learning system. Alternatively, it is likely that value function learnt by the brain is either fundamentally approximate or uses an different, less tractable, temporal discounting scheme. Ultimately, although in principle specialised and expensive learning rules might be developed to exactly replicate TD successor features in the brain, this maybe undesirable if a simple learning rule (STDP) is adequate in most circumstances. Indeed, animals - including humans - are known to act sub-optimally [51, 25], perhaps in part because of a reliance on STDP learning rules in order to learn long-range associations.

4 Methods

4.1 General summary of the model

The model comprises of an agent exploring a maze where its position $\mathbf{x}$ at time t is encoded by the instantaneous firing of a population of N CA3 basis features, $f_j(\mathbf{x}, t)$ for $j \in \{1, .., N\}$. Each has a spatial receptive field given by a thresholded Gaussian of peak firing rate 5 Hz:

$$f_j^x\big(\mathbf{x}(t)\big) = \begin{cases} \text{Gaussian}\big(\mathbf{x}_j, \sigma\big) - c & \text{if } ||\mathbf{x}(t) - \mathbf{x}_j|| < 1\text{m} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{x}_j$ is the location of the field peak, $\sigma = 1$m is the standard deviation and c is a positive constant that keeps f_j^x continuous at the threshold.

The theta phase of the hippocampal local field potential oscillates at 10 Hz and is denoted by $\phi_\theta(t) \in [0, 2\pi]$. Phase precession suppresses the firing rate of a basis features for all but a short period within each theta cycle. This period (and subsequently the time when spikes are produced, described in more details below) precesses earlier in each theta cycle as the agent crosses the spatial receptive field. Specifically, this is implemented by simply multiplying the spatial firing rate f_j^x by a theta modulation factor which rises and falls according to a von Mises distribution in each theta cycle, peaking at a 'preferred phase', ϕ_j^*, which depends on how far through the receptive field the agent has travelled (hence the spike timings implicitly encode location);

$$f_j^\theta\big(\phi_\theta(t)\big) = \text{VonMises}\big(\phi_i^*, \kappa\big) \tag{5}$$

where $\kappa = 1$ is the concentration parameter of the Von Mises distribution. These basis features in turn drive a population of N downstream 'STDP successor features' (Eqn 2).

Firing rates of both populations ($f_j(\mathbf{x}, \phi_\theta)$ and $\tilde{\psi}_i(\mathbf{x}, \phi_\theta)$) are converted to spike trains according to an inhomogeneous Poisson process. These spikes drive learning in the synaptic weight matrix, W_{ij}, according to an STDP learning rule (details below) with W_{ij} initialised as the identity δ_{ij}. During learning the effect of changes in W_{ij} are not propagated to the successor features (CA1), their influence is only considered during post-learning recall broadly analogous to the distinct encoding and retrival phases that have been hypothesised to underpin hippcampal function [52]. In summary, if a presynaptic CA3 basis features fires immediately before a postsynaptic CA1 successor feature the binding strength between these cells is strengthened. Conversely if they fire in the opposite order, their binding strength is weakened.

For comparison, we also implement successor feature learning using a temporal difference (TD) learning rule, referred to as 'TD successor features', $\psi_i(\mathbf{x})$, to provide a ground truth against which we compare the STDP successor features. Like STDP successor features, these are constructed as a linear combination of basis features (Eqn. 3).

Temporal difference learning updates M_{ij} as follows

$$\mathsf{M}_{ij} \leftarrow \mathsf{M}_{ij} + \eta \delta_{ij}^{\text{TD}} \tag{6}$$

where δ_{ij}^{TD} is the temporal difference error, which we derive below. In reinforcement learning the temporal difference error is used to learn discounted value functions (successor features can be considered a special type of value function). It works by comparing an unbiased sample of the true value function to the currently held estimate. The difference between these is known as the temporal difference error and is used to update the value estimate until, eventually, it converges on (or close to) the true value function.

4.2 Definition of TD successor features and TD successor matrix

To test our hypothesis that STDP is a good approximation to TD learning we simultaneously computed the true *TD successor features* defined as the total expected future firing of a basis feature:

$$\psi_i(\mathbf{x}) = \mathbb{E}\bigg[\int_t^\infty \frac{1}{\tau} e^{-\frac{t'-t}{\tau}} f_i^x\big(\mathbf{x}(t')\big) dt' \;\bigg|\; \mathbf{x}(t) = \mathbf{x}\bigg]. \tag{7}$$

τ is the temporal discounting time-horizon (related to γ, the discount factor used in reinforcement learning on temporally discretised MDPs, $\gamma = e^{-\frac{dt}{\tau}}$) and the expectation is over trajectories initiated at position $\mathbf{x}$. This formula explains the one-to-one correspondence between CA3 cells and CA1 cells in our hippocampal model (Fig. 1b): each CA1 cell, indexed i, learns to approximate the TD successor feature for its target basis feature, also indexed i. We set the discount timescale to $\tau = 4$ s to match relevant behavioural timescales for an animal exploring a small maze environment where behavioural decisions, such as whether to turn left or right, need to be made with respect to optimising future rewards occurring on the order of seconds.

It is typical in computational SR models to discretise time [41, 22]. Instead, we choose a continuous time formulation [53] which calls for using an integral form of the successor feature, (Eqn. 7), as opposed to the more conventional sum over discrete time steps (Eqn. (26)). In section 4.5 we present arguments for why this is a sensible way to model reinforcement learning in biological agents and gives improvements in terms of learning efficiency.

A second deviation from the 'typical' SR set-up is that we do not discretise space (as done by Stachenfeld et al. [22]) and a recent Hebbian SR model by Bono et al. [41]). Instead location is encoded by a vector of continuous basis features, $\vec{f}(\mathbf{x})$, which vary smoothly as the agent moves around the maze. By analogy we learn successor *features* [27], the expected future discounted firing rate of a basis feature, (Eqn. 7).

In our hippocampal model this change to learning successor features of Gaussian place cells is not only more biologically plausible than using one-hot state vectors but is crucial for our synaptic learning rule. Since STDP works on a rapid timescale, features must be overlapping in space in order for the cells encoding these feature to ever fire simultaneously, and thus form associations. To our knowledge this is the first SR model where spatially diffuse basis features are *required* by the learning rule, potentially helping to explain why the hippocampus uses heavily overlapping population codes.

Our TD successor matrix, M_{ij}, should not be confused with the successor *representation* as defined in Stachenfeld et al. [22] and denoted $M(\mathsf{s}_i, \mathsf{s}_j)$, though they are analogous. M_{ij} can be thought of as a spatially continuous analogue to $M(\mathsf{s}_i, \mathsf{s}_j)$ which, we show in the section 4.5, are equal (strictly, $M(\mathsf{s}, \mathsf{s}') = \mathsf{M}_{ij}^{\mathsf{T}}$) in the limit of a discrete one-hot place cell code.

4.3 Phase precession model details

In our hippocampal model CA3 place cells, referred to as basis features and indexed by j and have thresholded Gaussian receptive fields. The threshold radius is $\sigma = 1$ m and peak firing rate is $F = 5$ Hz. Mathematically, this is written as

$$f_j^x(\mathbf{x}(t)) = \frac{F}{1 - e^{-\frac{1}{2}}} \left[e^{-\frac{\|\mathbf{x}(t) - \mathbf{x}_j\|^2}{2\sigma^2}} - e^{-\frac{1}{2}} \right]_+, \tag{8}$$

where $[f(x)]_+ = max\big(0, f(x)\big)$, $\mathbf{x}_j$ is the centre of the receptive field and $\mathbf{x}(t)$ is the current location of the agent.

Phase precession is implemented by multiplying the spatial firing rate, $f_j^x(\mathbf{x})$, by a phase precession factor

$$f_j^\theta(\phi_\theta(t)) = 2\pi f_{\mathrm{VM}} \Big(\phi_\theta(t) \Big| \phi_j^*(\mathbf{x}), \kappa \Big). \tag{9}$$

where $f_{\mathrm{VM}}(x|\mu, \kappa)$ denotes the circular Von Mises distribution on $x \in (0, 2\pi]$ with mean $\mu = \phi_j^*(\mathbf{x})$ and spread parameter $\kappa = 1$. This factor is large only when the current theta phase,

$$\phi_\theta(t) = 2\pi\nu_\theta t \pmod{2\pi}, \tag{10}$$

which oscillates at $\nu_\theta = 10$ Hz, is close to the cell's 'preferred' theta phase,

$$\phi_j^*(\mathbf{x}(t)) = \pi + \beta\pi d_j(\mathbf{x}(t)). \tag{11}$$

$d_j(\mathbf{x}(t)) \in [-1, 1]$ tracks how far through the cell's spatial receptive field, as measured in units of σ, the agent has travelled:

$$d_j(\mathbf{x}(t)) = \frac{(\mathbf{x}(t) - \mathbf{x}_j) \cdot \frac{\dot{\mathbf{x}}(t)}{\|\dot{\mathbf{x}}(t)\|}}{\sigma}. \tag{12}$$

In instances where the agent travels directly across the centre of a cell (as is the case in 1D environments) then $(\mathbf{x}(t) - \mathbf{x}_j)$ and its normalised velocity (a vector of length 1, pointing in the direction of travel) $\frac{\dot{\mathbf{x}}(t)}{\|\dot{\mathbf{x}}(t)\|}$ are parallel such that $d_j(\mathbf{x})$ progresses smoothly in time from it's minimum, -1, to it's maximum, 1. In general, however, this extends to any arbitrary curved path an agent might take across the cell and matches the model used in [31]. We fit β and κ to biological data in Fig. 5a of Jeewajee et al. (2014) [31] ($\beta = 0.5$, $\kappa = 1$). The factor of 2π normalises this term, although the instantaneous firing may briefly rise above the spatial firing rate $f_j^x(\mathbf{x})$, the average firing rate over the entire theta cycle is still given by the spatial factor $f_j^x(\mathbf{x})$. In total, the instantaneous firing rate of the basis feature is given by the product of the spatial and phase precession factors (Eqn. 1).

Note that the firing rate of a cell depends explicitly on its location through the spatial receptive field (its "rate code") and implicitly on location through the phase precession factor (its "spike-time code") where location dependence is hidden inside the calculation of the preferred theta phase. Notably, the effect of phase precession is only visible on rapid "sub-theta" timescales. Its effect disappears when averaging over any timescale, T_{av} substantially longer than theta timescale of $T_\theta = 0.1$ s:

$$\frac{1}{T_{av}} \int_t^{t+T_{av}} f_j(\mathbf{x}(t), \phi_\theta(t)) dt \approx \frac{1}{T_{av}} \int_t^{t+T_{av}} f_j^x(\mathbf{x}(t)) dt \qquad \text{for} \qquad T_{av} >> T_\theta \tag{13}$$

This is important since it implies that the effect of phase precession is only important for synaptic processes with very short integration timescales, for example, STDP.

4.4 Synaptic learning via STDP

STDP is a discrete learning rule. In summary, if a presynaptic neuron j fires before a postsynaptic neuron i their binding strength (represented by the synaptic strength W_{ij}) is potentiated, conversely if the postsynaptic neuron fires before the presynaptic then weight is depressed. This is implemented as follows.

First, we convert the firing rates to spike trains. We sample, for each neuron, from an inhomogeneous spike train with rate parameter $f_j(\mathbf{x}, t)$ (for presynaptic basis features) or $\tilde{\psi}_i(\mathbf{x}, t)$ for postsynaptic successor features. This is done over the period $[0, T]$ across which the animal is exploring.

$$\big(f_j(\mathbf{x}, t), [0, T]\big) \overset{Poisson}{\longmapsto} \{t_j^{\text{pre}}\} \quad , \quad \big(\tilde{\psi}_i(\mathbf{x}, t), [0, T]\big) \overset{Poisson}{\longmapsto} \{t_i^{\text{post}}\} \tag{14}$$

Asymmetric Hebbian STDP is implemented online using a trace learning rule. Each presynaptic spike from CA3 cell, indexed j, increments an otherwise decaying memory trace, $T_j^{\text{pre}}(t)$, according to:

$$\tau^{\text{pre}} \frac{dT_j^{\text{pre}}(t)}{dt} = -T_j^{\text{pre}}(t) + \sum_{t' \sim \{t_j^{\text{pre}}\}} \delta(t - t'). \tag{15}$$

And likewise for CA1 postsynaptic spikes

$$\tau^{\text{post}} \frac{dT_i^{\text{post}}(t)}{dt} = -T_i^{\text{post}}(t) + \sum_{t' \sim \{t_i^{\text{post}}\}} \delta(t - t'). \tag{16}$$

We matched the STDP plasticity window decay times to experimental data: $\tau^{\text{pre}} = 20$ ms and $\tau^{\text{post}} = 40$ ms. Finally, the synaptic weight from neuron j to neuron i, denoted W_{ij}, updates whenever either neuron fires according to

$$\mathsf{W}_{ij} \leftarrow \mathsf{W}_{ij} + \eta \Big[\underbrace{a^{\text{pre}} \sum_{t_i \sim \{t_i^{\text{post}}\}} \delta(t - t_i) T_j^{\text{pre}}(t)}_{\text{“pre-before-post”}} + \underbrace{a^{\text{post}} \sum_{t_j \sim \{t_j^{\text{pre}}\}} \delta(t - t_j) T_i^{\text{post}}(t)}_{\text{“post-before-pre”}} \Big], \tag{17}$$

where η is the learning rate (here set to 0.01) and a^{pre} and a^{post} give the relative amounts of pre-before-post potentiation and post-before-pre depression, set to match experimental data from [19] as 1 and -0.4 respectively.

4.5 Comparison to temporal difference learning solution

Temporal difference learning The temporal difference (TD) update rule is used to learning the true successor matrix (Eqn. 7) and successor features (Eqn. 3). The standard TD(0) learning rule for a linear value function, $\psi_i(\mathbf{x})$, which basis feature weights M_{ij} is [20]:

$$\mathsf{M}_{ij} \leftarrow \mathsf{M}_{ij} + \eta \delta_i f_j^x(\mathbf{x}) \tag{18}$$

where δ_i is the observed TD-error for the i^{th} successor feature and η is the learning rate. Note that we are only considering the spatial component of the firing rate, $f_j^x(\mathbf{x})$, not the phase modulation component, $f_j^\theta(\mathbf{x})$, which (as shown) would average away over any timescale significantly longer than the theta timescale (100 ms). For now we will drop the superscript and write $f_j^x(\mathbf{x}) = f_j(\mathbf{x})$

To find the TD-error we must derive a temporally continuous analogue of the Bellman equation. Following [53] we take the derivative of Eqn. (7) which gives a consistency equation on the successor feature as follows:

$$\frac{d}{dt}\psi_i\big(\mathbf{x}(t)\big) = \frac{d}{dt}\int_t^\infty \frac{1}{\tau} e^{-\frac{t'-t}{\tau}} f_i\big(\mathbf{x}(t')\big) dt' \tag{19}$$

$$= \frac{1}{\tau}\Big(\psi_i\big(\mathbf{x}(t)\big) - f_i\big(\mathbf{x}(t)\big)\Big) \tag{20}$$

This gives a continuous TD-error of the form

$$\delta_i(t) = \frac{d}{dt}\psi_i\big(\mathbf{x}(t)\big) + \frac{1}{\tau}\Big(f_i\big(\mathbf{x}(t)\big) - \psi_i\big(\mathbf{x}(t)\big)\Big) \tag{21}$$

which can be rediscretised and rewritten by Taylor expanding the derivative ($\psi_i(t) = \frac{\psi_i(t) - \psi_i(t-dt)}{dt}$) to give

$$\delta_i(t) = \frac{1}{dt}\Big(\frac{dt}{\tau} f_i\big(\mathbf{x}(t)\big) + (1 - \frac{dt}{\tau})\psi_i\big(\mathbf{x}(t)\big) - \psi_i\big(\mathbf{x}(t - dt)\big)\Big). \tag{22}$$

This looks like a conventional TD-error term (typically something like $\delta_t = \mathsf{R}_t + \gamma \mathsf{V}_t - \mathsf{V}_{t-1}$) except that we can choose dt (the timestep between learning updates) freely. Finally expanding $\psi_i(\mathbf{x}(t))$ using (Eqn. 3) and substituting this back into Eqn. 18 gives the update rule:

$$\mathsf{M}_{ij} \leftarrow \mathsf{M}_{ij} + \frac{\eta}{dt}\left[\frac{dt}{\tau} f_i\big(\mathbf{x}(t)\big) + \sum_k \mathsf{M}_{ik}\Big[\big(1 - \frac{dt}{\tau}\big) f_k\big(\mathbf{x}(t)\big) - f_k\big(\mathbf{x}(t-dt)\big)\Big]\right] f_j\big(\mathbf{x}(t)\big). \tag{23}$$

Or, as a matrix update equation:

$$\mathsf{M} \leftarrow \mathsf{M} + \frac{\eta}{dt}\left[\frac{dt}{\tau}\mathbf{f}\big(\mathbf{x}(t)\big) + \mathsf{M}\Big[\big(1 - \frac{dt}{\tau}\big)\mathbf{f}\big(\mathbf{x}(t)\big) - \mathbf{f}\big(\mathbf{x}(t-dt)\big)\Big]\right]\mathbf{f}^{\mathsf{T}}\big(\mathbf{x}(t)\big). \tag{24}$$

As mentioned, this rule doesn't stipulate a fixed time step between updates. Unlike traditional TD updates rules on discrete MDPs, dt can take *any* positive value. The ability to adaptively vary dt has potentially underexplored applications for efficient learning: when information density is high (e.g. when exploring new or complex environments, or during a compressed replay event [54]) it may be desirable to learn regularly by setting dt small. Conversely when the information density is low (for example in well known or simple environments) or learning is undesirable (for example the agent is aware that a change to the environment is transient and should not be committed to memory), dt can be increased to slow learning and save energy. In practise, we set our agent to perform a learning update approximately every 1 cm along it's trajectory ($dt \approx 0.1$ s).

We add a small amount of $\mathsf{L2}$ regularisation by adding the term $-2\eta\lambda\mathsf{M}$ to the right hand side of Eqn. (25). This breaks the degeneracy in M_{ij} caused by having a set of basis features which is overly rich to construct the successor features and can be interpreted, roughly, as a mild energy constraint favouring smaller synaptic connectomes. In total the full update rule from our TD successor matrix is given by

$$\mathsf{M} \leftarrow \mathsf{M} + \frac{\eta}{dt}\left[\frac{dt}{\tau}\mathbf{f}\big(\mathbf{x}(t)\big) + \mathsf{M}\Big[\big(1 - \frac{dt}{\tau}\big)\mathbf{f}\big(\mathbf{x}(t)\big) - \mathbf{f}\big(\mathbf{x}(t-dt)\big)\Big]\right]\mathbf{f}^{\mathsf{T}}\big(\mathbf{x}(t)\big) - 2\eta\lambda\mathsf{M}. \tag{25}$$

Successor features in continuous time and space Typically, as in Stachenfeld et al. [22], the successor representation, $M(\mathsf{s}_i, \mathsf{s}_j)$, encodes the expected discounted future occupancy of state s_j along a trajectory initiated in state s_i:

$$M(\mathsf{s}_i, \mathsf{s}_j) = \mathbb{E}\left[\sum_{t=0} \gamma^t \delta(\mathsf{s}_t = \mathsf{s}_j) \;\middle|\; \mathsf{s}_0 = \mathsf{s}_i\right] \tag{26}$$

There are two forms of discretisation here. Firstly, time is discretised: it increases by a fixed increment, +1, to transition the state from $\mathsf{s}_t \to \mathsf{s}_{t+1}$. Secondly, assuming this is a spatial exploration task, space is discretised: the agent can be in exactly one state on any given time.

We loosen both these constraints reinstating time and space as continuous quantities. Since, for space, we cannot hope to enumerate an infinite number of locations, we represent the state by a population vector of diffuse, overlapping spatially localised place cells. Thus it is no longer meaningful to ask what the expected future occupancy of a single location will be. The closest analogue, since the place cells are spatially localised, is to ask how much we expect place cell, i, centred at $\mathbf{x}_i$, to fire in the near (discounted) future. This continuous time constraint alters the sum over time into an integral over time. Further, the role of γ which discounts state occupancy many time steps into the future, is replaced by τ which discounts firing a long time into the future. Thus the extension of the successor representation, $M(\mathsf{s}_i, \mathsf{s}_j)$, to continuous time and space is given by the successor *feature*,

$$\psi_i(\mathbf{x}) = \mathbb{E}\left[\int_t^{\infty} \frac{1}{\tau} e^{-\frac{t'-t}{\tau}} f_i\big(\mathbf{x}(t')\big) dt' \;\middle|\; \mathbf{x}(t) = \mathbf{x}\right]. \tag{27}$$

Why have we chosen to do this? Temporally it makes little sense to discretise time in a continuous exploration task: γ, the reinforcement learning discount factor, describes how many timesteps into the future the predictive encoding accounts for and so undesirably ties the predictive encoding to the otherwise arbitrary size of the simulation timestep, dt. In the continuous definition, τ intuitively describes how long into the future the predictive encoding discounts over and is independent of dt. This definition allows for online flexibility in the size of dt, as shown in Eqn. (25). This relieves the agent of a burden imposed by discretisatation; namely that it must learn with a fixed time step, +1, all the time. Now the agent potentially has the ability to choose the fidelity over which to learn and this may come with significant benefits in terms of energy efficiency, as described above. Further, using the discretised form implicitly ties the definition of the successor representation (or any similarly defined value value function) to the time step used in their simulation.

When space *is* discretised, the successor representation is a matrix encoding predictive relationships between these discrete locations. TD successor features, defined above, are the natural extension of the successor representation in a continuous space where location is encoded by a population of overlapping basis features, rather than exclusive one-hot states. The TD successor matrix, M_{ij}, can most easily be viewed as set of driving weights: M_{ij} is large if basis feature $f_j(\mathbf{x})$ contributes strongly to successor feature $\psi_i(\mathbf{x})$. They are closely related (for example, in the effectively discrete case of non-overlapping basis features, it can be shown that the TD successor matrix then corresponds directly to the transpose of the successor representation, $\mathsf{M}^{\mathsf{T}}_{ij} = \mathsf{M}(\mathsf{s}_i, \mathsf{s}_i)$, see below for proof) but we believe the continuous case has more applications in terms of biological plausibility; electrophysiological studies show hippocampus encodes position using a population vector of overlapping place cells, rather than one-hot states. Furthermore the continuous case maps neatly onto known neural circuity, as in our case with CA3 place cells as basis features, CA1 place cells as successor features, and the successor matrix as the synaptic weights between them. In our case, the choice not to discretise space and use a more biologically compatible basis set of large overlapping place cells is necessary - were our basis features to not overlap they would not be able to reliably form associations using STDP since often only one cell would ever fire in a given theta cycle.

For completeness (though this is not something studied in this report) this continuous successor feature form also allows for rapid estimation of the value function in a neurally plausible way. Whereas for the discrete case value can be calculated as:

$$V(\mathsf{s}_i) = \sum_j M(\mathsf{s}_i, \mathsf{s}_j) R(\mathsf{s}_j) \tag{28}$$

where $R(\mathsf{s}_j)$ is the per-time-step reward to be found at state s_j, for continuous successor feature setting:

$$\mathsf{V}(\mathbf{x}) = \sum_j \psi_j(\mathbf{x}) \mathsf{R}_j \tag{29}$$

where R_j is a vector of weights satisfying $\sum_j \mathsf{R}_j f_j(\mathbf{x}) = R(x)$ where $R(x)$ is the reward-rate found at location $\mathbf{x}$. Eqn. (29) can be confirmed by substituting into it Eqn. 27. R_j (like $R(\mathsf{s}_j)$) must be learned independent to, and as well as, the successor features, a process which is not the focus of this study although correlates have been observed in the hippocampus [55]. $\mathsf{V}(\mathbf{x})$ is the temporally continuous value associated with trajectories initialised at $\mathbf{x}$:

$$\mathsf{V}(\mathbf{x}) = \mathbb{E}\left[\int_t^\infty \frac{1}{\tau} e^{-\frac{t'-t}{\tau}} R\big(\mathbf{x}(t')\big) dt' \;\Big|\; \mathbf{x}(t) = \mathbf{x} \right]. \tag{30}$$

Equivalence of the TD successor matrix to the successor representation Here we show the equivalence between $M(\mathsf{s}_i, \mathsf{s}_j)$ and M_{ij}. First we can rediscretise time by setting dt' to be constant and defining $\gamma = 1 - \frac{dt'}{\tau}$ and $\mathbf{x}_n = \mathbf{x}(n \cdot dt')$. The integral in Eqn. (27) becomes a sum,

$$\psi_i(\mathbf{x}) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f_i(\mathbf{x}_t) \;\Big|\; \mathbf{x}_0 = \mathbf{x} \right]. \tag{31}$$

Next we rediscretise space by supposing that CA3 place cells in our model have strictly non-overlapping receptive fields which tile the environment. For each place cell, i, there is continuous area, $\mathcal{A}_i$, such that for any location within this area place cell i fires at a constant rate whilst all others are silent. When $\mathbf{x} \in \mathcal{A}_i$ we denote this state $\mathsf{s}(\mathbf{x}) = \mathsf{s}_i$ (since all locations in this area have identical population vectors).

$$f_i(\mathbf{x}) = \delta(\mathbf{x} \in \mathcal{A}_i) = \delta\big(\mathsf{s}(\mathbf{x}) = \mathsf{s}_i\big) \tag{32}$$

Let the initial state be $\mathsf{s}(\mathbf{x}) = \mathsf{s}_j$ (i.e. $\mathbf{x} \in \mathcal{A}_j$). Putting this into Eqn. (31) and equating to Eqn. (3), the definition of our TD successor matrix, gives

$$\psi_i(\mathbf{x}) = \sum_k \mathsf{M}_{ik} \delta(\mathsf{s}_j = \mathsf{s}_k) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \delta(\mathsf{s}_t = \mathsf{s}_i) \;\Big|\; \mathsf{s}_0 = \mathsf{s}_j \right], \tag{33}$$

confirming that

$$\mathsf{M}^{\mathsf{T}}_{ij} \propto M(\mathsf{s}_i, \mathsf{s}_j). \tag{34}$$

4.6 Simulation and analysis details

Maze details In the 1D open loop maze (Fig. 2a-e) the policy was to always move around the maze in one direction (left to right, as shown) at a constant velocity of 16 cm s^{-1} along the centre of the track. Although figures display this maze as a long corridor it is topologically identical to a loop; place cells close to the left or right sides have receptive fields extending into the right or left of the corridor respectively. 50 Gaussian basis features of radius 1 m, as described above, are placed with their centres uniformly spread along the track. Agents explored for a total time of 30 minutes.

In the 1D corridor maze, Fig. 2f-j, the situation is only changed in one way: the left and right hand edges of the maze are closed by walls. When the agent reaches the wall it turns around and starts walking the other way until it collides with the other wall. Agents explored for a total time of 30 minutes.

In the 2D two room maze, 200 basis feature are positioned in a grid across the two rooms (100 per room) then their location jittered slightly (Fig. 2k). The cells are geodesic Gaussians. This means that the $\|\mathbf{x}(t)-\mathbf{x}_i\|^2$ term in Eqn. 8 measures the distance from the agent location the the centre of cell i along the shortest walk which complies with the wall geometry. This explains the bleeding of the basis feature through the door in Fig. 3d. Agents explored for a total time of 120 minutes.

The movement policy of the agent is a random walk with momentum. The agent moves forward with the speed at each discrete time step drawn from a Rayleigh distribution centred at 16 cm s^{-1}. At each time step the agent rotates a small amount; the rotational speed is drawn from a normal distribution centred at zero with standard deviation 3π rad s^{-1} (π rad s^{-1} for the 1D mazes). Whenever the agent gets close to a wall (within 10 cm) the direction of motion is changed parallel to the wall, thus biasing towards trajectories which "follow" the boundaries, as observed in real rats. This model was designed to match closely the behaviour of freely exploring rats and was adapted from the model initially presented in Raudies and Hasselmo, 2012 [32]. We add one additional behavioural bias: in the 2D two room maze, whenever the agent passes within 1 metre of the centre point of the doorway connecting the two rooms it's rotational velocity is biased to turn it towards the door centre. This has the effect of encouraging room-to-room transitions, as is observed in freely moving rats [37].

Analyses of the STDP and TD successor matrices For the 1D mazes there exists a translational symmetry relating the $N = 50$ uniformly distributed basis features and their corresponding rows in the STDP/TD weight matrices. This symmetry is exact for the 1D loop maze (all cells around a circle are rotated versions of one another) and approximate for the corridor maze (broken only for cells near to the left or right bounding wall). The result is that the rows much the information in the linear track weight matrices matrices Fig. 2b,c,g,h can be viewed more easily by collapsing this matrix over the rows centred on the diagonal entry (plotted in Fig. 2d and i). This is done using a circular permutation of each matrix row by a count, n_i, equal to how many times we must shift cell i to the right in order for it's centre to lie at the middle of the track, $x_i = 2.5$m,

$$\mathsf{W}^{\text{aligned}}_{ij} = \mathsf{W}_{i,(j+n_i \pmod{50})}. \tag{35}$$

This is the 'row aligned matrix'. Averaging over its rows removes little information thanks to the symmetry,

$$\hat{\mathsf{W}}_j = \frac{1}{N}\sum_{i=1}^{N} \mathsf{W}^{\text{aligned}}_{ij}. \tag{36}$$

References

[1] W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions, Journal of neurology, neurosurgery, and psychiatry 20 (1) (1957) 11–21. doi:10.1136/jnnp.20.1.11.
URL https://www.ncbi.nlm.nih.gov/pubmed/13406589https://www.ncbi.nlm.nih.gov/pmc/articles/PMC497229/

[2] R. G. M. Morris, P. Garrud, J. N. P. Rawlins, J. O'Keefe, Place navigation impaired in rats with hippocampal lesions, Nature 297 (5868) (1982) 681–683, number: 5868 Publisher: Nature Publishing Group. doi:10.1038/297681a0.
URL https://www.nature.com/articles/297681a0

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

[4] J. S. Taube, R. U. Muller, J. B. Ranck, Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis., The Journal of neuroscience : the official journal of the Society for Neuroscience 10 (2) (1990) 420–35, iSBN: 0270-6474 (Print). doi:10.1212/01.wnl.0000299117.48935.2e.
URL http://www.cogsci.ucsd.edu/sereno/201/readings/08.04-HeadDirCells.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/2303851

[5] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, E. I. Moser, Microstructure of a spatial map in the entorhinal cortex, Nature 436 (7052) (2005) 801–806, iSBN: 1476-4687 (Electronic)$\backslash$r0028-0836 (Linking) _eprint: /dx.doi.org/10.1038/nature01964. doi:10.1038/nature03721.
URL https://www.nature.com/nature/journal/v436/n7052/pdf/nature03721.pdfhttp://www.nature.com/doifinder/10.1038/nature03721

[6] E. C. Tolman, Cognitive maps in rats and men., Psychological Review 55 (4) (1948) 189–208. doi:10.1037/h0061626.
URL https://doi.org/10.1037/h0061626

[7] J. O'Keefe, L. Nadel, The hippocampus as a cognitive map, Oxford: Clarendon Press, 1978.

[8] W. E. Skaggs, B. L. McNaughton, M. A. Wilson, C. A. Barnes, Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences, Hippocampus 6 (2) (1996) 149–172. doi:10.1002/(sici)1098-1063(1996)6:2<149::aid-hipo6>3.0.co;2-k.
URL https://doi.org/10.1002/(sici)1098-1063(1996)6:2<149::aid-hipo6>3.0.co;2-k

[9] J. Huxter, N. Burgess, J. O'Keefe, Independent rate and temporal coding in hippocampal pyramidal cells, Nature 425 (6960) (2003) 828–832. doi:10.1038/nature02058.
URL https://doi.org/10.1038/nature02058

[10] M. A. Wilson, B. L. McNaughton, Dynamics of the hippocampal ensemble code for space, Science (New York, N.Y.) 261 (5124) (1993) 1055–1058. doi:10.1126/science.8351520.

[11] J. O'Keefe, M. L. Recce, Phase relationship between hippocampal place units and the EEG theta rhythm, Hippocampus 3 (3) (1993) 317–330. doi:10.1002/hipo.450030307.
URL https://doi.org/10.1002/hipo.450030307

[12] A. P. Maurer, S. L. Cowen, S. N. Burke, C. A. Barnes, B. L. McNaughton, Organization of hippocampal cell assemblies based on theta phase precession, Hippocampus 16 (9) (2006) 785–794. doi:10.1002/hipo.20202.

[13] A. Johnson, A. D. Redish, Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point, Journal of Neuroscience 27 (45) (2007) 12176–12189, publisher: Society for Neuroscience Section: Articles. doi:10.1523/JNEUROSCI.3761-07.2007.
URL https://www.jneurosci.org/content/27/45/12176

[14] D. Bush, C. Barry, D. Manson, N. B. Correspondence, N. Burgess, Using Grid Cells for Navigation, Neuron 87 (2015) 507–520. doi:10.1016/j.neuron.2015.07.006.
URL http://dx.doi.org/10.1016/j.neuron.2015.07.006

[15] A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, G. Wayne, H. Soyer, F. Viola, B. Zhang, R. Goroshin, N. Rabinowitz, R. Pascanu, C. Beattie, S. Petersen, A. Sadik, S. Gaffney, H. King, K. Kavukcuoglu, D. Hassabis, R. Hadsell, D. Kumaran, Vector-based navigation using grid-like representations in artificial agents, Nature 557 (7705) (2018) 429–433. doi:10.1038/s41586-018-0102-6.
URL http://www.nature.com/articles/s41586-018-0102-6

[16] R. U. Muller, J. L. Kubie, R. Saypoff, The hippocampus as a cognitive graph (abridged version), Hippocampus 1 (3) (1991) 243–246. doi:10.1002/hipo.450010306.
URL https://doi.org/10.1002/hipo.450010306

[17] K. I. Blum, L. F. Abbott, A Model of Spatial Map Formation in the Hippocampus of the Rat, Neural Computation 8 (1) (1996) 85–93, publisher: MIT Press. doi:10.1162/neco.1996.8.1.85.
URL https://www.mitpressjournals.org/doi/10.1162/neco.1996.8.1.85

[18] M. Mehta, M.C. Quirk, M. Wilson, Experience-Dependent Asymmetric Shape of Hippocampal Receptive Fields, Neuron 25 (2000) 707–715.

[19] G.-q. Bi, M.-m. Poo, Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type, Journal of Neuroscience 18 (24) (1998) 10464–10472, publisher: Society for Neuroscience Section: ARTICLE. doi:10.1523/JNEUROSCI.18-24-10464.1998.
URL https://www.jneurosci.org/content/18/24/10464

[20] R. S. Sutton, A. G. Barto, Introduction to Reinforcement Learning, 1st Edition, MIT Press, Cambridge, MA, USA, 1998.

[21] P. Dayan, Improving generalization for temporal difference learning: The successor representation, Neural Computation 5 (4) (1993) 613–624. doi:10.1162/neco.1993.5.4.613.
URL https://doi.org/10.1162/neco.1993.5.4.613

[22] K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map, Nature Neuroscience 20 (11) (2017) 1643–1653. doi:10.1038/nn.4650.
URL http://www.nature.com/articles/nn.4650

[23] A. Alvernhe, E. Save, B. Poucet, Local remapping of place cell firing in the Tolman detour task, European Journal of Neuroscience 33 (9) (2011) 1696–1705, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2011.07653.x. doi:https://doi.org/10.1111/j.1460-9568.2011.07653.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2011.07653.x

[24] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, S. J. Gershman, The successor representation in human reinforcement learning, Nature Human Behaviour 1 (9) (2017) 680–692. doi:10.1038/s41562-017-0180-8.
URL http://www.nature.com/articles/s41562-017-0180-8

[25] W. de Cothi, N. Nyberg, E.-M. Griesbauer, C. Ghanamé, F. Zisch, J. Lefort, L. Fletcher, C. Newton, S. Renaudineau, D. Bendor, R. Grieves, E. Duvelle, C. Barry, H. J. Spiers, Predictive Maps in Rats and Humans for Spatial Navigation, preprint, Animal Behavior and Cognition (Sep. 2020). doi:10.1101/2020.09.26.314815.
URL http://biorxiv.org/lookup/doi/10.1101/2020.09.26.314815

[26] E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, Predictive representations can link model-based reinforcement learning to model-free mechanisms, PLoS Computational Biology 13 (9). doi:10.1371/journal.pcbi.1005768.
URL http://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1005768&type=printable

[27] W. de Cothi, C. Barry, Neurobiological successor features for spatial navigation, Hippocampus 30 (12) (2020) 1347–1355, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hipo.23246. doi:https://doi.org/10.1002/hipo.23246.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.23246

[28] W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward, Science 275 (5306) (1997) 1593–1599. doi:10.1126/science.275.5306.1593.
URL https://doi.org/10.1126/science.275.5306.1593

[29] K. B. Kjelstrup, T. Solstad, V. H. Brun, T. Hafting, S. Leutgeb, M. P. Witter, E. I. Moser, M.-B. Moser, Finite Scale of Spatial Representation in the Hippocampus, Science 321 (5885) (2008) 140–143, publisher: American Association for the Advancement of Science Section: Report. doi:10.1126/science.1157086.
URL https://science.sciencemag.org/content/321/5885/140

[30] I. Momennejad, M. W. Howard, Predicting the Future with Multi-scale Successor Representations, preprint, Neuroscience (Oct. 2018). doi:10.1101/449470.
URL http://biorxiv.org/lookup/doi/10.1101/449470

[31] A. Jeewajee, C. Barry, V. Douchamps, D. Manson, C. Lever, N. Burgess, Theta phase precession of grid and place cell firing in open environments, Philosophical Transactions of the Royal Society B: Biological Sciences 369 (1635) (2014) 20120532. doi:10.1098/rstb.2012.0532.
URL https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0532

[32] F. Raudies, M. E. Hasselmo, Modeling Boundary Vector Cell Firing Given Optic Flow as a Cue, PLoS Computational Biology 8 (6) (2012) e1002553. doi:10.1371/journal.pcbi.1002553.
URL https://dx.plos.org/10.1371/journal.pcbi.1002553

[33] C. Dong, A. D. Madar, M. E. J. Sheffield, Distinct place cell dynamics in CA1 and CA3 encode experience in new environments, Nature Communications 12 (1) (2021) 2977, number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41467-021-23260-3.
URL https://www.nature.com/articles/s41467-021-23260-3

[34] S. Tanni, W. de Cothi, C. Barry, State transitions in the statistically stable place cell population are determined by rate of perceptual change, arxivdoi:10.1101/2021.06.16.448638.
URL https://doi.org/10.1101/2021.06.16.448638

[35] H. J. Spiers, R. M. A. Hayman, A. Jovalekic, E. Marozzi, K. J. Jeffery, Place field repetition and purely local remapping in a multicompartment environment, Cerebral Cortex 25 (1) (2015) 10–25, iSBN: 1047-3211. doi:10.1093/cercor/bht198.

[36] D. Dupret, J. O'Neill, B. Pleydell-Bouverie, J. Csicsvari, The reorganization and reactivation of hippocampal maps predict spatial memory performance, Nature Neuroscience 13 (8) (2010) 995–1002. doi:10.1038/nn.2599.
URL https://doi.org/10.1038/nn.2599

[37] F. Carpenter, D. Manson, K. Jeffery, N. Burgess, C. Barry, Grid Cells Form a Global Representation of Connected Environments, Current Biology 25 (9) (2015) 1176–1182. doi:10.1016/j.cub.2015.02.037.
URL https://linkinghub.elsevier.com/retrieve/pii/S0960982215002067

[38] T. Eliav, S. R. Maimon, J. Aljadeff, M. Tsodyks, G. Ginosar, L. Las, N. Ulanovsky, Multiscale representation of very large environments in the hippocampus of flying bats, Science 372 (6545) (2021) eabg4020. doi:10.1126/science.abg4020.
URL https://www.sciencemag.org/lookup/doi/10.1126/science.abg4020

[39] B. A. Strange, M. P. Witter, E. S. Lein, E. I. Moser, Functional organization of the hippocampal longitudinal axis, Nature Reviews Neuroscience 15 (10) (2014) 655–669. doi:10.1038/nrn3785.
URL https://doi.org/10.1038/nrn3785

[40] J. Brea, A. T. Gaál, R. Urbanczik, W. Senn, Prospective Coding by Spiking Neurons, PLOS Computational Biology 12 (6) (2016) e1005003. doi:10.1371/journal.pcbi.1005003.
URL https://dx.plos.org/10.1371/journal.pcbi.1005003

[41] J. Bono, S. Zannone, V. Pedrosa, C. Clopath, Learning predictive cognitive maps with spiking neurons during behaviour and replays, arxivdoi:10.1101/2021.08.16.456545.
URL https://www.biorxiv.org/content/10.1101/2021.08.16.456545v2

[42] N. J. Gustafson, N. D. Daw, Grid cells, place cells, and geodesic generalization for spatial reinforcement learning, PLoS Computational Biology 7 (10) (2011) e1002235. doi:10.1371/journal.pcbi.1002235.
URL https://doi.org/10.1371/journal.pcbi.1002235

[43] J. F. Cavanagh, C. M. Figueroa, M. X. Cohen, M. J. Frank, Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation., Cerebral cortex 22 11 (2012) 2575–86.

[44] M. A. Wilson, B. L. McNaughton, Reactivation of hippocampal ensemble memories during sleep, Science 265 (5172) (1994) 676–679. doi:10.1126/science.8036517.
URL https://doi.org/10.1126/science.8036517

[45] D. Bush, H. F. Ólafsdóttir, C. Barry, N. Burgess, Ripple band phase precession of place cell firing during replay, Current Biology 32 (1) (2022) 64–73.

[46] S. Cheng, L. M. Frank, New experiences enhance coordinated neural activity in the hippocampus, Neuron 57 (2) (2008) 303–313. doi:10.1016/j.neuron.2007.11.035.
URL https://doi.org/10.1016/j.neuron.2007.11.035

[47] M. G. Mattar, N. D. Daw, Prioritized memory access explains planning and hippocampal replay, Nature Neuroscience 21 (11) (2018) 1609–1617. doi:10.1038/s41593-018-0232-z.
URL https://doi.org/10.1038/s41593-018-0232-z

[48] E. Todorov, Efficient computation of optimal actions, Proceedings of the National Academy of Sciences 106 (28) (2009) 11478–11483. doi:10.1073/pnas.0710743106.
URL https://doi.org/10.1073/pnas.0710743106

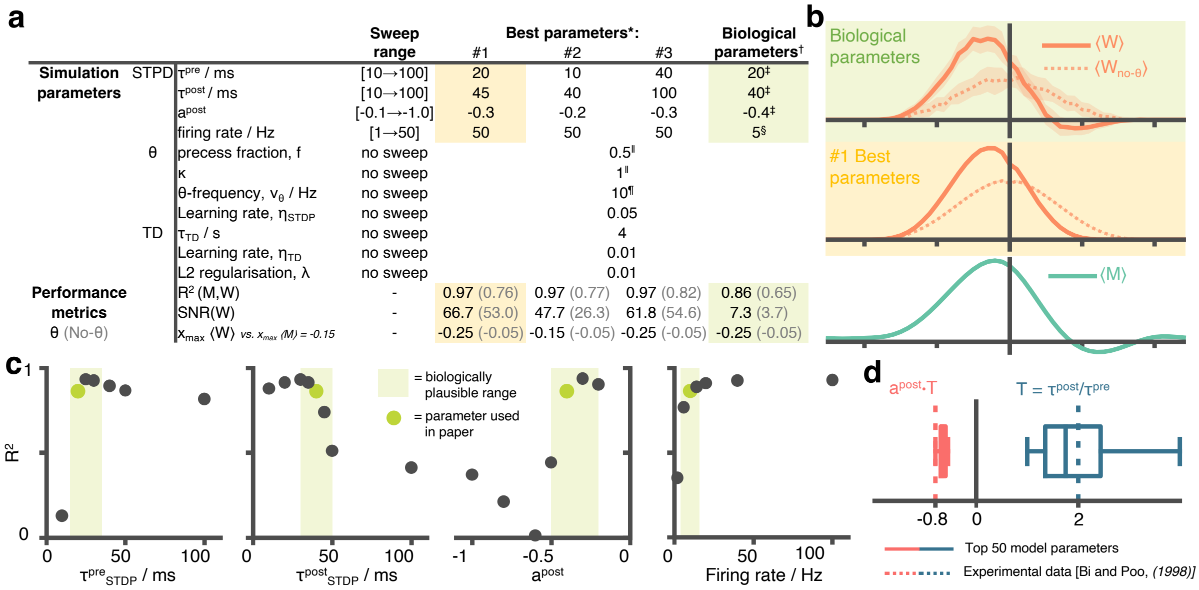
[49] P. Piray, N. D. Daw, Linear reinforcement learning in planning, grid fields, and cognitive control, Nature Communications 12 (1) (2021) 4942. doi:10.1038/s41467-021-25123-3.
URL https://www.nature.com/articles/s41467-021-25123-3

[50] J. E. Lisman, A. A. Grace, The hippocampal-VTA loop: Controlling the entry of information into long-term memory, Neuron 46 (5) (2005) 703–713. doi:10.1016/j.neuron.2005.05.002.
URL https://doi.org/10.1016/j.neuron.2005.05.002

[51] T. R. Zentall, When animals misbehave: Analogs of human biases and suboptimal choice, Behavioural Processes 112 (2015) 3–13. doi:10.1016/j.beproc.2014.08.001.
URL https://doi.org/10.1016/j.beproc.2014.08.001

[52] M. E. Hasselmo, C. Bodelón, B. P. Wyble, A Proposed Function for Hippocampal Theta Rhythm: Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning, Neural Computation 14 (4) (2002) 793–817. doi:10.1162/089976602317318965.
URL https://direct.mit.edu/neco/article/14/4/793-817/6585

[53] K. Doya, Reinforcement learning in continuous time and space, Neural Computation 12 (1) (2000) 219–245. doi:10.1162/089976600300015961.
URL https://doi.org/10.1162/089976600300015961



			Sweep range	Best parameters*: #1	#2	#3	Biological parameters†
Simulation parameters	STPD	τ^{pre} / ms	[10→100]	20	10	40	20‡
		τ^{post} / ms	[10→100]	45	40	100	40‡
		a^{post}	[-0.1→-1.0]	-0.3	-0.2	-0.3	-0.4‡
		firing rate / Hz	[1→50]	50	50	50	5§
	θ	precess fraction, f	no sweep		0.5‖		
		κ	no sweep		1‖		
		θ-frequency, ν_θ / Hz	no sweep		10¶		
		Learning rate, η_{STDP}	no sweep		0.05		
	TD	τ_{TD} / s	no sweep		4		
		Learning rate, η_{TD}	no sweep		0.01		
		L2 regularisation, λ	no sweep		0.01		
Performance metrics		R^2 (M,W)	-	0.97 (0.76)	0.97 (0.77)	0.97 (0.82)	0.86 (0.65)
θ (No-θ)		SNR(W)	-	66.7 (53.0)	47.7 (26.3)	61.8 (54.6)	7.3 (3.7)
		x_{max} ⟨W⟩ *vs.* x_{max} *⟨M⟩ = -0.15*	-	-0.25 (-0.05)	-0.15 (-0.05)	-0.25 (-0.05)	-0.25 (-0.05)

*as determined by R^2 (M,W_θ)
† as used in paper, determined from literature...
‡ Bush et al. (2010)
§ Skaggs et al. (1996)
‖ Fitted to Fig. 5a from Jeewajee and Barry (2014)
¶ Jeewajee and Barry (2014)

Supplementary Figure 1: A hyperparameter sweep over STDP parameters shows that biological parameters are suffice, and are near-optimal for approximating the successor features **a** A table showing all parameters used in this paper and the ranges over which the hyperparameter sweep was performed. For each parameter setting we estimate performance metrics to judge whether the STDP parameters do well at learning the successor features. **b** Visually inspecting the row aligned STDP weight matrices we see the optimal parameters do not significantly out perform the biologically chosen ones. Although the optimal parameter setting results in a slightly higher R^2, they fail to capture the right-of-centre negative weights present in the true TD successor matrix, unlike the biological ones. **c** Slices through the parameter sweep hypercube. For each plot, parameter values of the other three variables are fixed to the green values (i.e. are the ones used in this paper). **d** The top 50 performing parameter combination are stored and box plots for the conjugate parameter $T = \frac{\tau^{\text{pre}}}{\tau^{\text{post}}}$, the ratio of time windows for potentiation and depression, and $-a^{\text{post}} \cdot T$, effectively the ratio of the areas under the curve left and right of the y-axis on the STDP plot Fig. 1b. In both cases the 'best parameters' include the true parameter values, measures expimentally by Bi and Poo (1998) [19]

[54] W. E. Skaggs, B. L. McNaughton, Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience, Science 271 (5257) (1996) 1870–1873. doi:10.1126/science.271.5257.1870.
URL https://doi.org/10.1126/science.271.5257.1870

[55] J. L. Gauthier, D. W. Tank, A Dedicated Population for Reward Coding in the Hippocampus, Neuron 99 (1) (2018) 179–193.e7. doi:10.1016/j.neuron.2018.06.008.
URL https://www.sciencedirect.com/science/article/pii/S0896627318304768