SPATIAL REPRESENTATIONS THAT TRANSFER BY REMAPPING EMERGE FROM META-REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Animals and artificial agents navigating partially observable environments construct spatial representations from limited egocentric input. In artificial agents and in models of the hippocampal cognitive map alike, such representations are typically learned with predictive or auxiliary objectives decoupled from reward, then treated as a fixed basis for value learning. Spatial representations can emerge from reward alone, but how they are organised to transfer across environments is unknown. Here we train a recurrent agent with meta-reinforcement learning across grid-worlds with varying home and reward locations, using no spatial priors or auxiliary objectives, and analyse its hidden units as a neural population. With weights fixed, the agent solves novel environments in-context, returning home by a shortest path on $> 99\%$ of trajectories, even with sensory cues removed. Its hidden states form a spatial representation: units develop localised fields, position is linearly decodable, and hidden-state similarity tracks spatial distance when the agent navigates from memory but not when its goal is in view. Across unmarked home locations this representation remaps selectively, with a subset of units tracking the home while others maintain a largely stable spatial code. These remapping units determine where the agent goes: ablating them degrades home decoding far more than ablating any other units, and overwriting their state once, with the pattern for a different home, sends it to that false home. These properties resemble the hippocampus, where dedicated cells follow a remembered goal while the rest of the map is largely unchanged, and which is required for navigation from memory but not towards a visible goal. Reward maximisation alone thus yields a spatial representation that transfers across environments, so a spatial basis need neither be fixed in advance nor relearned in new environments.

1 INTRODUCTION

Animals and artificial agents navigating partially observable environments must construct spatial representations from limited egocentric input. In artificial agents such representations are typically induced with auxiliary objectives decoupled from reward, such as reconstructive, predictive, or path-integration losses (Mirowski et al., 2017; Banino et al., 2018; Uria et al., 2020), and in models of the hippocampal cognitive map the spatial representation is likewise learned by a predictive objective, as in the successor representation (Dayan, 1993; Stachenfeld et al., 2017). In both cases the representation is then treated as a fixed basis on which value is learned (Gustafson & Daw, 2011).

Spatially structured representations are known to emerge from reward alone (Wijmans et al., 2023), but it remains unclear how such codes are organised so that a single readout transfers across environments. In the hippocampus, spatial representations are reused across contexts through partial remapping, in which a subset of cells reorganises while the rest of the map is preserved (Colgin et al., 2008; Sanders et al., 2020). We ask whether a similar organisation arises when a representation is optimised for reward alone.

Here, we train a recurrent agent with meta-reinforcement learning (Duan et al., 2016; Wang et al., 2016) across a distribution of grid-worlds with varying reward locations and an unrewarded home

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

location, using no spatial priors or auxiliary objectives, and analyse the resulting hidden units with methods from systems neuroscience. Our contributions are as follows:

Spatial representations emerge from reward alone. With weights fixed, the agent solves novel environments in-context, including when sensory cues are removed, and its hidden units develop localised fields from which position, reward, and home can be linearly decoded.

The representation remaps selectively and transfers across environments. Across home locations a subpopulation of units remaps with the home while others remain largely stable, so that linear decoders trained on some home locations generalise to a held-out one.

The remapping units causally direct homing. Ablating them degrades home decoding far more than ablating other units, and overwriting their state once, as the agent collects reward, redirects it to a false home (cf. Robinson et al., 2020).

Hidden-state geometry tracks distance only when navigating from memory. Similarity between hidden states follows spatial distance when the agent returns to an unmarked home or searches for an unseen reward, but not when its goal is in view (cf. Morris et al., 1982).

2 Background

We consider the meta-reinforcement learning setting with recurrent networks (Duan et al., 2016; Wang et al., 2016), and assume a distribution $\rho_{\mathcal{M}}$ over partially observable Markov decision processes (POMDPs) $M = (\mathcal{S}, \mathcal{A}, \Omega, \mathcal{P}, \mathcal{O}, r, \gamma)$, where $\mathcal{S}$, $\mathcal{A}$, and Ω denote the state, action, and observation sets, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ a transition probability distribution, $\mathcal{O} : \mathcal{S} \times \mathcal{A} \times \Omega \to \mathbb{R}_+$ an observation function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ a reward function, and $\gamma \in [0, 1]$ a discount factor. We define a trial as a fixed number of time steps T within a sampled POMDP, $M \sim \rho_{\mathcal{M}}$, and an episode as the interval between natural boundaries within it (in our task, one excursion from the unmarked home location to a reward and back), so that a trial contains multiple episodes.

At each time step t the agent receives an observation o_t rather than the true state s_t, and passes it through a 2-layer convolutional encoder. The encoded observation is concatenated with the previous action, reward, and termination flag to form the input to a Long Short-Term Memory (LSTM) network, whose hidden state h_t parameterises both the policy $\pi_\theta(a_t \mid h_t)$ and the value function $V_\theta(h_t)$ through separate linear layers (Mnih et al., 2016). The network is trained end-to-end, such that representations are optimised jointly with behaviour (Mirowski et al., 2017). Since the observation provides only partial information about the current state, h_t must serve as a statistic of past history; it is therefore preserved across episode boundaries and reset only between trials, when a new POMDP is sampled. Accordingly, the meta-reinforcement learning objective is to maximise the expected cumulative discounted reward across a whole trial, rather than a single episode:

$$J(\theta) = \mathbb{E}_{M \sim \rho_{\mathcal{M}}} \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]. \tag{1}$$

Since each trial requires a distinct policy, the network must implement a fast learning algorithm within its recurrent dynamics, integrating information accumulated across earlier episodes to improve performance on later ones. The meta-objective therefore encourages representations that capture shared structure while remaining flexible enough to support behaviour in the current environment (Yu et al., 2020). The network weights θ are learned across trials by Proximal Policy Optimisation (PPO; Schulman et al., 2017); architecture and training details are given in Appendix A.1.

3 Related Work

Spatial reinforcement learning has largely focused on learning representations that capture the transition structure of an environment. The successor representation (Dayan, 1993) decomposes value into a transition model and a reward function, so that a shared spatial representation can be combined with different reward functions to support rapid value learning (Stachenfeld et al., 2017; Barreto et al., 2017). However, even the most straightforward approaches to learning it rely on features that are inherently spatial, such as one-hot state representations (Dayan, 1993; Stachenfeld et al., 2017) or Gaussian functions (Barreto et al., 2017; de Cothi & Barry, 2020; George et al., 2023), rather than

allowing spatial structure to emerge from learning. Moreover, since the successor representation is defined with respect to a particular transition structure, it transfers across reward functions within an environment but must be relearned across environments.

Learning spatial representations without a spatial prior has therefore been pursued through auxiliary objectives: reconstructive objectives (Kulkarni et al., 2016) that force the network to preserve state information from the input, predictive objectives (Banino et al., 2018; Whittington et al., 2020; Uria et al., 2020) that encourage it to anticipate future states, regularisation objectives (Liu & Abbeel, 2021; Whittington et al., 2023) that impose specific properties on the learned representations, or combinations thereof (Li et al., 2015; Jaderberg et al., 2017; Hafner et al., 2020). Collectively, these approaches show that spatially meaningful representations can emerge without an explicit spatial basis, but not whether an auxiliary objective is needed for them to emerge at all, nor whether the resulting representations are the most useful for spatial behaviour.

Meta-reinforcement learning with recurrent networks provides an alternative (Duan et al., 2016; Wang et al., 2016; Mirowski et al., 2017; Ni et al., 2022; Beck et al., 2023). Rather than imposing a spatial prior or relying on auxiliary objectives, it optimises a single set of network parameters to support value learning across many environments and reward functions, such that any spatial structure that may emerge must arise from the meta-training objective itself. Duan et al. (2016) show that this is sufficient in practice, with end-to-end meta-training producing adaptive spatial strategies that generalise across environments and reward functions in a 3D visual navigation task, and Wijmans et al. (2023) show that agents trained by reinforcement learning alone develop map-like representations in memory that support navigation in novel environments. Spatial representations may therefore arise as a natural consequence of meta-reinforcement learning (Alver & Precup, 2021). Notably, however, Mirowski et al. (2017) report spatially structured representations under the same formulation only when the objective is augmented with auxiliary tasks, and none of these studies establish how the emergent representation is organised such that a single readout transfers across environments.

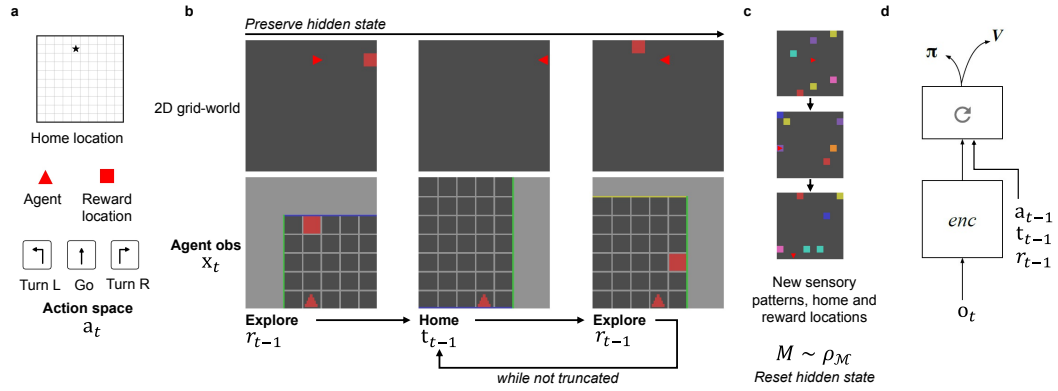


Figure 1: Task and architecture. (**a**) The agent navigates a grid-world containing an unmarked home location (star) and a visible reward, using three actions: turn left, go forward, and turn right. (**b**) Within a trial it alternates between exploring for the reward and returning to its unmarked home, whereupon a new reward is placed, seeing only a 7×7 egocentric view of the grid ahead of it (shown with sensory cues removed). (**c**) Its hidden state is preserved across these episodes and reset only at the start of each trial, when a new home location, set of reward locations, and sensory cues are sampled. (**d**) The network encodes each observation o_t and passes it, together with the previous action, termination flag, and reward, to an LSTM whose hidden state parameterises separate policy (π) and value (V) heads (schematic adapted from Mirowski et al., 2017).

Here, we take a similar approach in examining the spatial behaviour (Duan et al., 2016) and representations (Mirowski et al., 2017; Wijmans et al., 2023) that emerge under meta-reinforcement learning, but without spatial priors or auxiliary objectives, with the aim of characterising the organisation that supports transfer. To do so we draw on the analyses developed to characterise partial remapping in hippocampal recordings (Bostock et al., 1991; Anderson & Jeffery, 2003; Colgin et al., 2008; Sanders et al., 2020), in which a subset of cells reorganises while the rest of the map is preserved, and in particular the dedicated population of cells that encodes a remembered, unmarked goal and follows it when reward is relocated (Gauthier & Tank, 2018).

4 EXPERIMENT

Our experiment consists of a spatial navigation task in a procedurally generated 10×10 grid-world, in which the agent receives 7×7 egocentric observations of the layout immediately in front of it, rendered as RGB images and upsampled to 84×84. The action space is discrete and consists of three moves: turn left, turn right, and go forward (Figure 1a). A trial starts with the agent spawned at a random start location with the recurrent states set to zero, and proceeds through alternating explore and home phases (Figure 1b). In the explore phase, the agent searches for a visible $+1$ reward. In the home phase, the agent returns to its spawn location ('home'), receiving a termination flag, upon which a new reward is randomly placed. A trial spans 128 time steps and can thus include multiple episodes, each an explore trajectory followed by a return home. Each trial samples a new home location, random reward locations, and a grid-world with distinct sensory cues (Figure 1c), such that each trial produces unique trajectories to and from the home location and thus the recurrent dynamics must adapt based on experience. The hidden state is preserved across episodes within a trial and reset only between trials (Figure 1b–d), and so serves as a statistic of trial history that can be viewed as an implicit belief state. The optimal strategy is to explore efficiently, using memory to avoid revisiting locations already seen, and once the reward is found to return home by the shortest path. Since all weights are fixed at evaluation, this must be achieved by the recurrent dynamics alone.

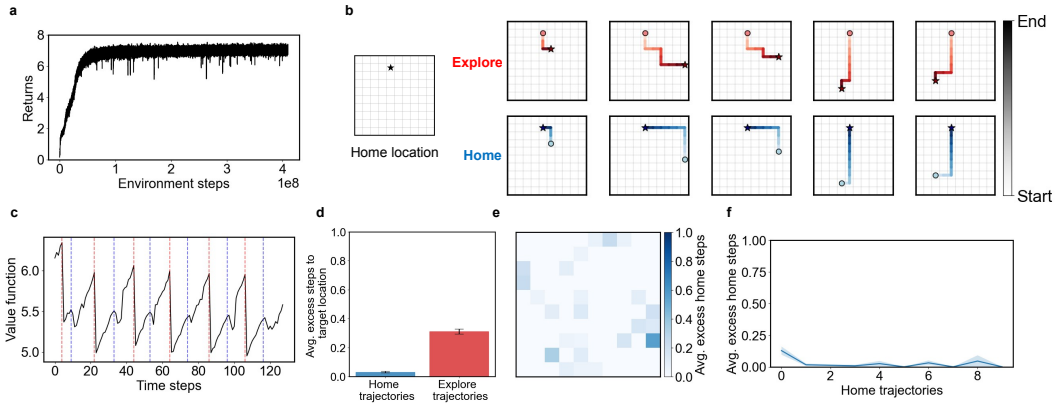


Figure 2: Meta-reinforcement learning produces efficient exploration and near-optimal homing. (**a**) Average return per trial rises over training and plateaus (mean across 5 seeds). (**b**) In a novel environment the agent explores until it finds the reward and then returns directly home, as shown by the explore (red) and home (blue) trajectories of an example trial, each running from its start (circle) to the reward or home (star) and shaded from light to dark over time. (**c**) Over the same trial the value function rises as the agent searches and again as it approaches home, with reward collection (red lines) and arrival home (blue lines) marked. (**d**) Explore trajectories, which must find a reward that is often out of view, take more forward steps beyond the shortest path than home trajectories (mean and s.e.m. over 5 trials per home location), of which (**e**) $> 99\%$ follow a shortest path to every home location, (**f**) including the very first return of a trial (mean and s.e.m. over 5 trials per home).

Returns plateau during training (Figure 2a), and when evaluated in a novel environment the agent explores effectively to locate rewards and returns directly to its spawn location (Figure 2b). The value function increases as the agent searches for possible reward locations, and also as it approaches home after collecting reward: the home carries no explicit reward itself, but returning to it triggers placement of a new reward, so it is predictive of future reward and hence value (Figure 2c). Since the new reward often lies outside the agent's observation window, exploration trajectories are by nature less direct than those back to home (Figure 2d), with $> 99\%$ of home trajectories following the shortest possible path from the preceding reward location and an average of only 0.02 excess forward steps (Figure 2e). Homing is near-optimal from the first episode: having visited the home location only once in a new environment, the agent returns along a near-shortest path, demonstrating one-shot learning (Figure 2f). Crucially, this behaviour is preserved outside the training distribution: with the sensory cues removed entirely, so that only the arena boundary and the reward remain visible (Figure 1b, $> 99\%$ of home trajectories remain optimal; see Appendix A.3.1 for generalisation performance to longer trials and other arena sizes). This implies the agent learns the underlying

spatial structure of the grid-world rather than relying on the sensory cues to navigate home. Accordingly, all subsequent analyses were conducted on a single trained network, with weights fixed and in an environment with sensory cues removed (as in Figure 1b).

5 Analysis

5.1 Spatial tuning

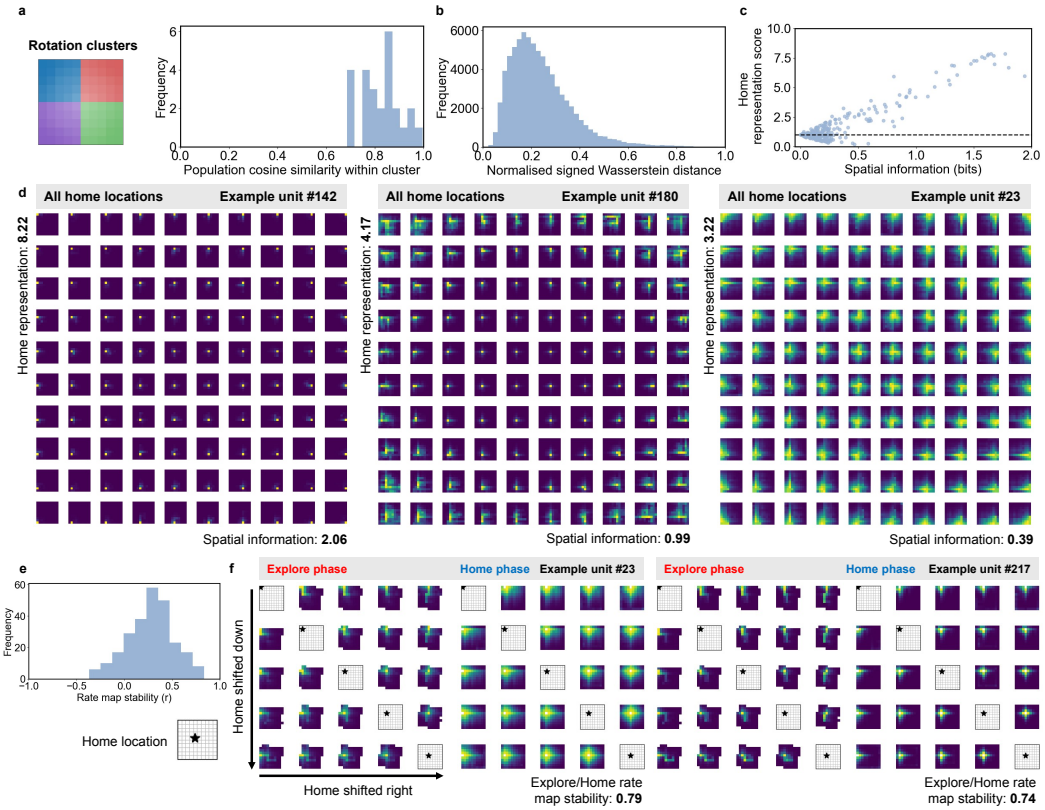


Figure 3: Spatial representations emerge from meta-reinforcement learning and remap selectively across home locations. (**a**) Population rate maps fall into four rotational clusters, one per quadrant of the home location, with high cosine similarity between the maps within each cluster (transparency scales inversely with similarity, whose distribution is shown on the right). (**b**) Within a cluster, the signed Wasserstein distance between a unit's rate maps for two home locations measures how far its activity must move to turn one map into the other; across units this distance is concentrated at low values with a long tail, so most units change little between home locations while a minority remap substantially. (**c**) Units with high spatial information concentrate their activity around the home: the home representation score, the fraction of a unit's activity within two forward steps of home relative to the area that region covers, rises with spatial information, with a score of 1 (dashed line) indicating no preference. (**d**) Three example units which remap, plotted for every home location, have fields ranging from compact (unit #142) to broad (unit #23). (**e**) Most units' rate maps are positively correlated between the explore and home phases, (**f**) as illustrated for two home-representing units across a block of home locations within one rotational cluster (columns, home shifted right; rows, home shifted down; starred cells mark the home).

We next asked whether the recurrent hidden units had developed a spatial representation, and whether that code was shared across environments. To test this, we generated population rate maps from trials with each home location, rotated them through the four cardinal orientations, and assigned each population the rotation that produced the highest cosine similarity to a reference home location. The population representation was shared across home locations up to a rotation (Figure 3a), yielding four rotational clusters, one per quadrant of the environment.

Next, to quantify how individual units remap across home locations within a cluster, we computed a signed Wasserstein distance between each unit's rate maps for pairs of home locations (Appendix A.2). This measures the average distance that activity must be moved to transform one map into the other and, by operating on deviations from a unit's mean activity, treats a moving trough in the same way as a moving peak. The distribution of these distances was concentrated at low values with a long tail (Figure 3b): most units change little between home locations, while a minority in the tail remap substantially. We refer to this pattern as selective remapping. Consistent with this, we identified a population of units with high spatial information whose activity was disproportionately concentrated around the home location (Figure 3c). To quantify this, we defined a home representation score as the proportion of a unit's total activity occurring within two forward steps of the home location, divided by the proportion of the environment that area occupies (Appendix A.2), such that scores above and below 1 indicate over- and under-representation of the home. Units with a score above 1.5 (35%) concentrated their activity around the home location, with fields ranging from compact to broad (Figure 3d). The majority of units (60%, score between 0.5 and 1.5) had low spatial information and no preference for the home location, and a small population (5%, score below 0.5) showed inverse tuning, firing throughout the environment except near the home (Figure 3c).

Rate maps were positively correlated across the explore and home phases for most units (Figure 3e, f), and the fields of some stable units resembled the agent's outbound and return trajectories rather than a single location (Figure 3f, Appendix A.3.2), suggesting that they encode transitions between home and reward, as a successor representation would (Dayan, 1993).

5.2 Decoding position, reward, and home

Spatially tuned units do not by themselves establish that the network encodes the agent's position. To test this, we trained linear decoders to predict position from hidden unit activity (Appendix A.2), evaluating them on held-out trajectories from the home locations used for training ('test') and on trials from a single home location withheld entirely ('validate'). Decoded position was accurate as soon as the agent began to move and remained so throughout the trial (Figure 4a), with comparable accuracy for the held-out home location (Figure 4b), indicating that the learned position representation generalises across home locations.

Next, to determine whether task-relevant information is similarly decodable, we trained linear decoders to predict the reward and home locations (Figure 4c–h). During exploration, the decoded posterior over the reward location narrowed progressively (Figure 4c), consistent with structured exploration in which locations already seen and found empty are ruled out, and decoding error fell sharply once the reward came into view, for trained and held-out home locations alike (Figure 4d). The home location, in contrast, was decodable throughout the trial: accurately and stably during the return trajectory (Figure 4e, f), and also while the agent was exploring for reward (Figure 4g, h). A representation of home is therefore maintained independently of the current behavioural target, as expected from the units that represent home across both phases of the trial (Figure 3e, f).

On trials with the held-out home location, home decoding error was a constant one bin (Figure 4f, h). Since the decoder has no output for the held-out location, the nearest trained home is the best prediction available, and this is what the decoder produced. That the representation of an unseen home is closest to that of its spatial neighbour, rather than to an arbitrary trained home, is consistent with home locations being encoded within a common spatial structure.

5.3 Transfer by remapping

The preceding analyses show that a subset of units remaps with the home location while others remain comparatively stable, and that the home location can be decoded from the population. We therefore asked whether the remapping units are the ones on which this readout depends. To test this, we ranked units by the extent of their remapping between home locations, measured as the signed Wasserstein distance between their rate maps averaged across pairs of home locations (Figure 5a; Section 5.1), and ablated units cumulatively in rank order, replacing each unit's activity with its mean so that the ablation remained within the distribution of normal activity. Ablating units in order of decreasing signed Wasserstein distance increased home decoding error well beyond the range produced by ablating the same number of units in random order (Figure 5b), whereas ablating in order of increasing distance left home decoding almost unaffected, even when this extended to

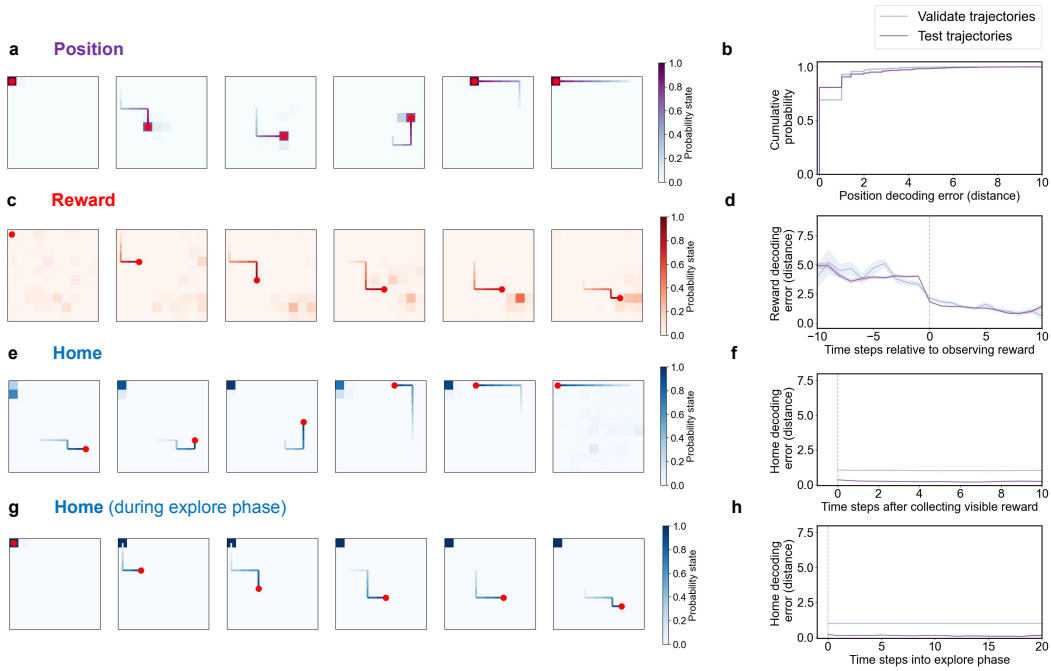


Figure 4: Position, reward, and home are linearly decodable from the hidden state, including for a held-out home location. Linear decoders were trained on hidden unit activity from all but one home location and evaluated on held-out trajectories from those locations ('test') and on trials from the withheld location ('validate'). (**a**) Along an example trajectory the decoded posterior over position is accurate from the agent's first move onwards, and (**b**) the cumulative distribution of position decoding error is similar for trained and held-out home locations. (**c**) During exploration the decoded posterior over the reward location narrows as locations are found empty, and (**d**) decoding error falls sharply once the reward comes into view (dashed line). (**e**) On the home trajectory that follows, the posterior over the home location is accurate throughout, with (**f**) an error near zero for trained home locations and a constant one bin for the held-out home, whose nearest trained neighbour the decoder returns. (**g**, **h**) The home location is decodable even while the agent is still exploring. Shaded error regions display s.e.m calculated over $n = 3000$ 'test' and $n = 300$ 'validate' trajectories.

nearly half the population. The remapping units are therefore necessary for the trained decoder to read out the home location accurately.

Next, we asked whether these units are also sufficient to determine where the agent navigates. To test this, we selected the 100 units that remapped most and, at the moment the agent collected the reward and was about to return home, injected a false home representation by overwriting their hidden and cell states (Figure 5c). The injected pattern was the mean activity of those same units at the same location, from a trial in which the home was elsewhere. To assess where the agent then navigated, we used probe trials in which neither the real nor the false home terminated the episode, leaving the agent free to move for the remaining time steps, akin to a probe trial in the Morris water maze (Morris et al., 1982). Following injection, the agent spent far more of the probe trial near the false home than near the real one (Figure 5d), and its occupancy was concentrated around the false home and, to a lesser extent, around locations rotationally symmetric to it (Figure 5e); swapping the real and false homes reproduced this result (Appendix A.3.4). This rotational symmetry mirrors the cluster structure in Figure 3a: the mean pattern conveys the position of the home within its rotational cluster more reliably than the cluster itself. When the real and false homes were instead allowed to terminate the episode, return trajectories predominantly ended at the false home (Figure 5f, g). The proportion that did so increased with the number of units overwritten, whereas overwriting in order of increasing signed Wasserstein distance, or in random order, produced less redirection (Figure 5g). The units that remap with the home location are therefore both necessary for decoding it and sufficient to redirect the agent to a location that has not served as its home in the current

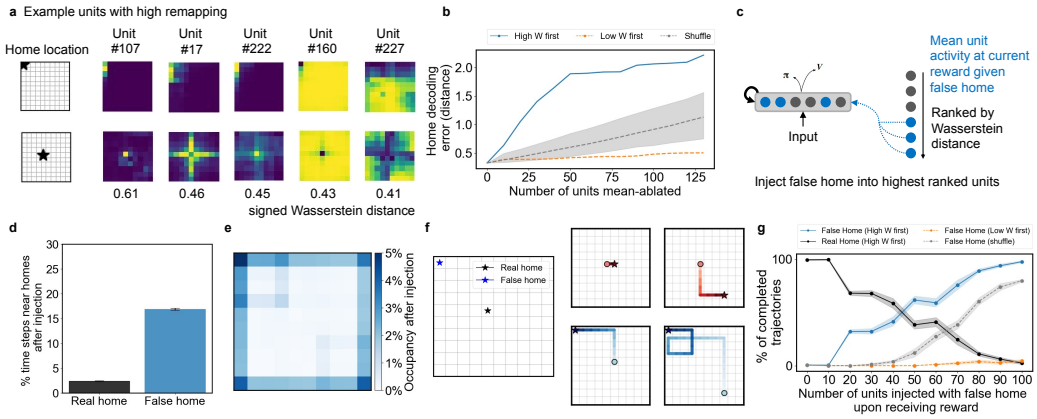


Figure 5: Remapping units are necessary for reading out the home location and sufficient to redirect the agent to a false home. (**a**) Five units that remap across home locations, ranked by the signed Wasserstein distance between their rate maps, are shown for two home locations (stars denote home). (**b**) Mean-ablating units cumulatively in this order ('High W first') raises home decoding error far beyond ablating the same number in random order ('Shuffle'; shaded, 95% interval), whereas ablating from the least remapping unit upwards ('Low W first') leaves it almost unaffected. (**c**) At the moment of reward collection, the hidden and cell states of the 100 most strongly remapping units were overwritten with their mean activity at that moment on trials with a different, false, home. (**d**) In probe trials, in which neither home terminates the episode, the agent then spends far more time near the false home than the real one (error bars, s.e.m. across 500 repeats), and (**e**) its occupancy concentrates at the false home and its rotational images. (**f**) With termination, the home phase after injection ends at the false home (explore phase red, home phase blue; real home, black star; false home, blue star), and (**g**) the proportion of completed home trajectories ending there grows with the number of remapping units overwritten, whereas overwriting the least remapping or random units produces less redirection (labels as in **b**; shading, s.e.m. across 5 repeats).

trial. Because the injected pattern was averaged over trials with a different home, these units carry a representation of home that is shared across trials and causally directs homing.

5.4 Hidden-state similarity tracks spatial distance

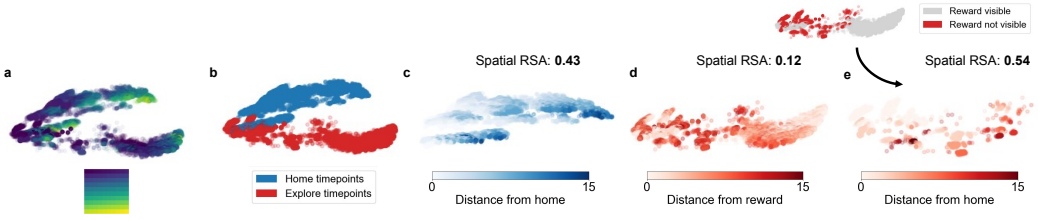


Figure 6: Hidden-state similarity tracks spatial distance when the agent navigates from memory but not when its goal is in view. (**a**) Population activity from an example trial is projected onto its first two principal components, with each point representing one time step, coloured by the agent's position. (**b**) The points form two clusters corresponding to the home and explore phases. (**c**) Colouring the home-phase points by distance from home reveals a smooth gradient, as does (**d**) colouring the explore-phase points by distance from reward. (**e**) Restricting the explore phase to time steps at which the reward was not visible reveals a gradient of distance from home as well (inset, reward visible in grey, not visible in red). Spatial RSA is given above **c**–**e**: similarity tracks distance during homing (0.43) and during search for an unseen reward (0.54), but only weakly across the whole explore phase (0.12).

We next asked whether the population as a whole represents a spatial geometry, such that hidden state activities are more similar at nearby grid-world locations. To visualise this, we projected the population activity from an example trial onto its first two principal components (Figure 6). The pro-

jection revealed two clusters, each of which mapped the entire environment separately (Figure 6a) and which corresponded to the explore and home phases of the task (Figure 6b). To quantify this geometry, we used spatial representational similarity analysis (spatial RSA): for every pair of time points, the cosine distance between the two population activity vectors was compared with the Euclidean distance between the agent's positions at those times, using the Spearman correlation across all pairs (Appendix A.2). Within the home cluster, colouring by distance from home revealed a smooth gradient (Figure 6c), and spatial RSA confirmed that hidden-state distances during the home phase reflected spatial distances (spatial RSA = 0.43).

Colouring the explore cluster by distance from reward likewise revealed a smooth gradient (Figure 6d), yet spatial RSA during the explore phase was markedly weaker (0.12). This difference was associated with reward visibility: restricting the analysis to explore time steps at which the reward was not visible restored the correlation (spatial RSA = 0.54; Figure 6e), and the hidden state also showed a continuous gradient of distance from home. Approaching a visible reward can be guided by the observation alone, whereas searching for one that is not yet visible requires the agent to track its position relative to the regions it has already explored and to home. Hidden-state similarity therefore tracks spatial distance when the agent navigates from memory, whether returning to an unmarked home or searching for an unseen reward, but not when its goal is in view.

6 Discussion

The agent presented here learns spatial representations as a by-product of maximising reward across a distribution of environments, without spatial priors or auxiliary objectives. Across home locations the learned representation remaps selectively, with a subset of units following the home, while the others retain a largely stable spatial code, so that linear decoders trained on some home locations generalise to a held-out one. The remapping units are necessary for reading out the home and sufficient to redirect the agent to a false one, and hidden-state geometry tracks spatial distance only when the agent navigates from memory.

This work builds upon previous meta-reinforcement learning approaches to spatial navigation, which have focused on exploring towards, and returning to, a single learned reward location (Duan et al., 2016; Mirowski et al., 2017; Zintgraf et al., 2020); by contrast, our agent achieves near-optimal homing in a task requiring adaptive exploration and exploitation along a continuous trajectory. Spatially structured representations have been reported under meta-reinforcement learning only with auxiliary objectives (Mirowski et al., 2017), or without an account of how they support transfer (Wijmans et al., 2023). Here the meta-reinforcement learning objective alone is sufficient, and the resulting spatial basis is organised for transfer. In this respect the hidden state can be viewed as performing task inference (Humplik et al., 2019): the remapping units carry the part of the representation that identifies the current environment, while the stable population carries the part shared across environments.

Similar results have been observed in the hippocampus, although we make no direct comparison to neural data. Place cells remap partially across contexts (Bostock et al., 1991; Anderson & Jeffery, 2003; Colgin et al., 2008) and reorganise when goals are moved (Dupret et al., 2010), a small population of cells encodes a remembered goal and follows it when reward is relocated (Gauthier & Tank, 2018), targeted activation of place cells is sufficient to drive memory-guided navigation (Robinson et al., 2020), and the hippocampus is required for navigation to a hidden but not a visible goal (Morris et al., 1982). There are also differences: the population of units with a home representation score above 1.5 (35%) is far larger than the 4% of spatially tuned cells reported by Gauthier & Tank (2018), although the criteria differ, and the home is never itself rewarded but is predictive of subsequent reward ('value') through task structure.

Limitations. Our environment is a partially observable 2D grid-world, we have not compared against an agent with a fixed spatial basis, and our conclusions rest on a single architecture and trained network. Future work should extend the approach to richer, higher-dimensional environments (Duan et al., 2016; Mirowski et al., 2017; Hafner et al., 2020; Uria et al., 2020), to continuous control (Lillicrap et al., 2016), and to procedurally generated mazes (Duan et al., 2016), where transition statistics vary through structure rather than behaviour. Even so, reward maximisation alone produced a spatial basis that transfers across environments, which need neither be fixed in advance nor relearned in each new one.

AI USE STATEMENT

We used large language models only to improve the clarity and quality of the written English and to assist with debugging code. We did not use them to generate scientific content, to design experiments, to analyse data, or to make any other intellectual contribution. We have reviewed all AI-assisted text and code, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We will provide the source code for producing the primary results of the paper in the supplementary materials. The model architecture and training procedure are described in Appendix A.1, the PPO hyperparameters are listed in Table A1, and all analyses are defined in Appendix A.2. Modelling was performed on a machine with a single CPU with 24 cores and 32 threads (Intel i9-14900K), a single GPU (NVIDIA RTX 5000 Ada Generation), and 32 GB of RAM.

ETHICS STATEMENT

This work uses only simulated agents and involves no human or animal subjects or personal data.

REFERENCES

Marco P. Abrate, Laurenz Muessig, Joshua P. Bassett, Hui Min Tan, Francesca Cacucci, Thomas J. Wills, and Caswell Barry. From movement to cognitive maps: Recurrent neural networks reveal how locomotor development shapes hippocampal spatial coding. In *International Conference on Learning Representations*, 2026.

Safa Alver and Doina Precup. What is going on inside recurrent meta reinforcement learning agents? *arXiv preprint arXiv:2104.14644*, 2021.

Michael I. Anderson and Kathryn J. Jeffery. Heterogeneous modulation of place cell firing by changes in context. *Journal of Neuroscience*, 23(26):8827–8835, 2003.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jacob Beck, Risto Vuorio, Zheng Xiong, and Shimon Whiteson. Recurrent hypernetworks are surprisingly strong in Meta-RL. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Elizabeth Bostock, Robert U. Muller, and John L. Kubie. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, 1(2):193–205, 1991.

Laura Lee Colgin, Edvard I. Moser, and May-Britt Moser. Understanding memory through hippocampal remapping. *Trends in Neurosciences*, 31(9):469–477, 2008.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

William de Cothi and Caswell Barry. Neurobiological successor features for spatial navigation. *Hippocampus*, 30(12):1347–1355, 2020.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

David Dupret, Joseph O'Neill, Barty Pleydell-Bouverie, and Jozsef Csicsvari. The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nature Neuroscience*, 13(8):995–1002, 2010.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Jeffrey L. Gauthier and David W. Tank. A dedicated population for reward coding in the hippocampus. *Neuron*, 99(1):179–193.e7, 2018.

Tom M. George, William de Cothi, Kimberly L. Stachenfeld, and Caswell Barry. Rapid learning of predictive maps with STDP and theta phase precession. *eLife*, 12:e80663, 2023.

Nicholas J. Gustafson and Nathaniel D. Daw. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Computational Biology*, 7(10):e1002235, 2011.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A. Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Tejas D. Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman. Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*, 2016.

Xiujun Li, Lihong Li, Jianfeng Gao, Xiaodong He, Jianshu Chen, Li Deng, and Ji He. Recurrent reinforcement learning: A hybrid approach. *arXiv preprint arXiv:1509.03044*, 2015.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Hao Liu and Pieter Abbeel. APS: Active pretraining with successor features. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6736–6747. PMLR, 2021.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J. Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2017.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

R. G. M. Morris, P. Garrud, J. N. P. Rawlins, and J. O'Keefe. Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683, 1982.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16691–16723. PMLR, 2022.

Nick T. M. Robinson, Lucie A. L. Descamps, Lloyd E. Russell, Moritz O. Buchholz, Brendan A. Bicknell, Georgy K. Antonov, Joanna Y. N. Lau, Rebecca Nutbrown, Christoph Schmidt-Hieber, and Michael Häusser. Targeted activation of hippocampal place cells drives memory-guided spatial behavior. *Cell*, 183(6):1586–1599, 2020.

Honi Sanders, Matthew A. Wilson, and Samuel J. Gershman. Hippocampal remapping as hidden state inference. *eLife*, 9:e51140, 2020.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

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

William Skaggs, Bruce McNaughton, and Katalin Gothard. An Information-Theoretic Approach to Deciphering the Hippocampal Code. In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1993.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017.

Ben Uria, Borja Ibarz, Andrea Banino, Vinicius Zambaldi, Dharshan Kumaran, Demis Hassabis, Caswell Barry, and Charles Blundell. A model of egocentric to allocentric understanding in mammalian brains. *bioRxiv*, 2020.

Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matthew Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020.

James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy E. J. Behrens. Disentanglement with biological constraints: A theory of functional cell types. In *International Conference on Learning Representations*, 2023.

Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S. Morcos, and Dhruv Batra. Emergence of maps in the memories of blind navigation agents. In *International Conference on Learning Representations*, 2023.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 2020.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020.

A Appendix

A.1 Model architecture and training

We adopt a model architecture similar to Duan et al. (2016); Wang et al. (2016); Mirowski et al. (2017). Specifically, we encode the RGB observation with a 2-layer convolutional network and represent the policy as a Long Short-Term Memory (LSTM) network. The first convolutional layer has a kernel of size 8×8 and a stride of 4×4, and 16 feature maps. The second convolutional layer has a kernel of size 4×4 and a stride of 2×2, and 32 feature maps. The extracted features are then passed to a fully connected layer with 256 hidden units which feeds into the LSTM. The recurrent component consists of 256 hidden units. At each time step, the LSTM also receives the previous action a_{t-1}, expressed as a one-hot vector of dimension 3, along with the previous termination flag t_{t-1} and previous reward r_{t-1}. The policy and value function share all intermediate representations, with each being derived from the LSTM through separate linear layers (Mnih et al., 2016).

We limit the activations of the LSTM to positive values by replacing the $\tanh$ output activation function with a sigmoid, such that the hidden state activations h_t are computed as $h_t = \sigma(c_t) \odot o_t$, where σ is the element-wise sigmoid function, c_t is the cell state, o_t is the output gate, and $\odot$ is the Hadamard product. This modification constrains the hidden state activations to non-negative values, simplifying the analysis and interpretation of their corresponding activations (Uria et al., 2020; Abrate et al., 2026).

The network weights θ are learned across trials by Proximal Policy Optimisation (PPO; Schulman et al., 2017), which maximises a clipped surrogate objective that discourages the updated policy from deviating too far from the previous policy $\pi_{\theta_{\text{old}}}$:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\Big[\min\Big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t\Big)\Big], \qquad \rho_t(\theta) = \frac{\pi_\theta(a_t \mid h_t)}{\pi_{\theta_{\text{old}}}(a_t \mid h_t)}, \quad (2)$$

where $\hat{A}_t$ is the generalised advantage estimate (Schulman et al., 2016) and ϵ the clip ratio (Table A1).

In each parallel environment process, agents step independently within their environment, collecting a trajectory of observations, rewards, and actions that differ from those experienced by agents in other processes. We use a flag-based system to coordinate buffer collection once all agents have completed the required number of steps. Collected trajectories are partitioned into fixed-length sequences, which are shuffled across epochs to form PPO minibatches. We use 32 workers and optimise the shared parameters with the AdamW optimiser. Gradients are computed over non-overlapping sequences of length 128 from each trial, using truncated backpropagation through time (BPTT), where the hidden state at the start of each sequence is preserved for unrolling. Episodes end by natural termination rather than teleportation, which preserves a continuous trajectory and avoids discontinuous observations that would leak trial structure. The experiment is run for approximately 4.0×10^8 environment steps performed over 1.0×10^5 trials of 128 time steps each.

Table A1: Parameters for PPO

Parameter	Value
Hidden units N_h	256
Discount factor γ	0.99
Generalised Advantage Estimate λ	0.97
Learning rate α	0.0001
Entropy coefficient β	0.01
Clip ratio ϵ	0.2
Value function coefficient	0.25
Number of epochs	2
Number of minibatches	8
Max gradient norm	0.5

A.2 SPATIAL ANALYSES

Rotation cosine similarity. To quantify structural rotations, we compute cosine similarity between rate maps from each home location and a set of four 90-degree rotated versions of a reference rate map derived from a single home location. Each rate map was vectorised, and similarity was computed as the normalised dot product between vectors. For each target home location, we select the rotation that maximised cosine similarity with the reference, yielding both the best-matching rotation and the corresponding maximum similarity score.

Wasserstein distance. We compute a signed-difference Wasserstein distance for each unit and each pair of home locations. We mean-centre both rate maps, split each into positive and negative components relative to the mean, and normalise each to unit mass. We then calculate the Wasserstein distance between these components using a normalised Euclidean ground-distance matrix (Flamary et al., 2021) and scale it by the transported mass to give a signed measure. This captures both the magnitude and spatial displacement of rate-map changes.

Spatial information. We calculate the spatial information of a rate map corresponding to unit k as $SI_k = \sum_i \frac{O_i}{\|O\|_1} \frac{R_i}{\widehat{R}} \log_2 (\frac{R_i}{\widehat{R}})$, where $O \in \mathbb{R}^{B_p \times B_p}$ is the occupancy matrix, R_i is the mean activation in state i, and $\widehat{R}$ is the overall mean activation across all states (Skaggs et al., 1993).

Home representation. We quantify home location representation for individual units by calculating the proportion of a unit's total activity occurring within two forward steps of the home location, divided by the proportion of the environment contained within this area. A score greater than 1 indicates higher activity near the home location, whereas a score of 1 indicates no preferential representation.

Decoding. We train a series of linear decoders to predict reward location, home location, and current position from timeseries hidden activations. Specifically, we train separate linear decoders to predict: (i) the current target location (reward location during explore, home location during home), (ii) home location, and (iii) current position, each by minimising the cross-entropy loss between decoder logits and ground-truth labels. To evaluate spatial generalisation to unseen environments, we withhold trials from a single home location entirely as a validation set. For the remaining home locations, we construct training and test sets using an 80-20 split. Figure 4 labels these the 'validate' and 'test' trajectories, respectively. We use the Adam optimiser with learning rate 0.001 for 5.0×10^4 weight updates with minibatches of 512 randomly sampled time steps from the training set.

Principal component analysis. We apply principal component analysis (PCA) with two components to the hidden unit activations from the timeseries data. Trial time steps were then split by phase, and activity from each phase was separately projected onto this shared PCA basis. This allowed us to relate the resulting low-dimensional trajectories to task variables recorded at each time step, including distance to home, distance to reward, and reward visibility.

Spatial representational similarity analysis. Spatial RSA quantifies the correspondence between spatial and hidden unit distances. We calculate spatial distance as the Euclidean distance between the grid coordinates of each pair of positions $D_p = |p_t - p_{t'}|_2$ and calculate the hidden unit distance as the cosine distance between the corresponding hidden unit activity vectors at each pair of time points $D_A = 1 - \frac{A_t \bullet A_{t'}}{|A_t|_2 \, |A_{t'}|_2}$. We then calculate the Spearman rank correlation between the spatial and hidden unit distance matrices.

A.3 ADDITIONAL RESULTS

A.3.1 GENERALISATION TO ENVIRONMENTS OUTSIDE THE TRAINING DISTRIBUTION

In the main text we state that spatial behaviour generalises to environments outside the training distribution. Here, we show two further examples: (i) when the truncation limit is increased (Figure A1) and (ii) when the environment size is changed (Figure A2). The agent continues to perform coherent trajectories well beyond the original truncation limit, which indicates that the recurrent dynamics has learned a representation of task structure that can be iterated indefinitely. Indeed, we show successful explore and home trajectories in environments of varying sizes (Figure A2a–c), though optimal spatial behaviour is degraded with increasing environment size (Figure A2d, e). Taken together, these results suggest that the network has not simply memorised trajectories specific

to the training distribution, but has instead acquired a more general spatial strategy for exploration and homing that remains partially effective outside the conditions it was trained under.

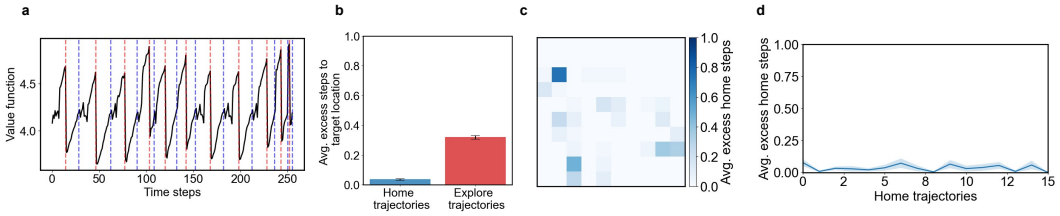


Figure A1: Hidden state can be unrolled past the original truncation limit. Example trial of 256 time steps, twice the truncation limit used in training. (**a**) The value function continues to operate past the original truncation limit, rising as the agent searches and again as it approaches home, with reward collection (red lines) and arrival home (blue lines) marked. (**b**) Explore trajectories still take more forward steps beyond the shortest path than home trajectories (mean and s.e.m. over 5 trials per home location), of which (**c**) $> 99\%$ still follow a shortest path to every home location, (**d**) including the latter home trajectories that occur past the original truncation limit (mean and s.e.m. over 5 trials per home)

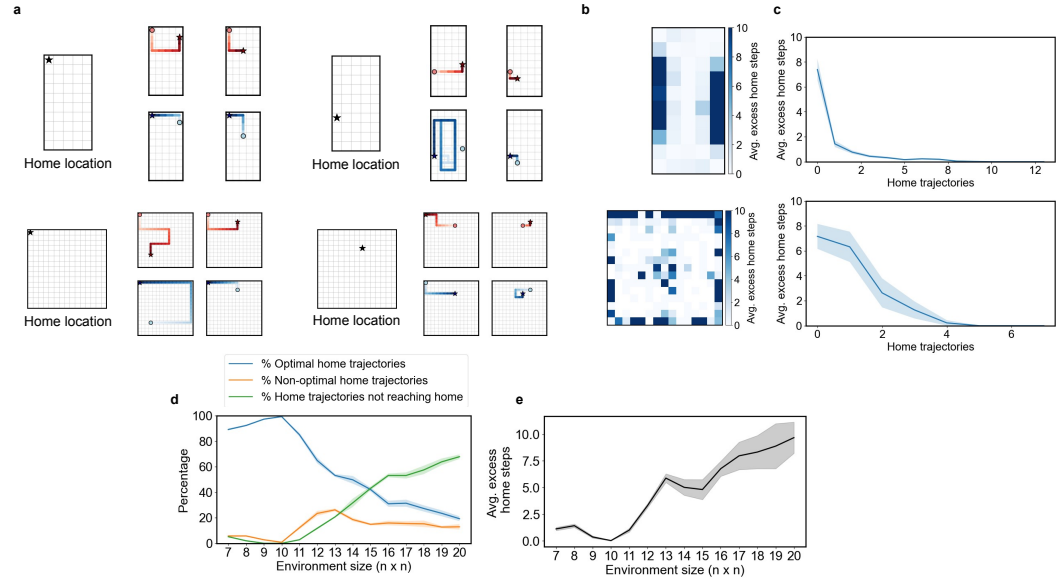


Figure A2: Generalisation to environments outside the training distribution. (**a**) The agent still performs structured exploration and homing behaviour in a 5×10 and a 15×15 environment, though (**b**) trajectories back home are less optimal. (**c**) However, in these out-of-distribution environments, the agent shows within-context learning of the home location, with trajectories home becoming more optimal with more experience. (**d**) As the environment size increases, more home trajectories do not make it home within the truncation limit, (**e**) rather than becoming less optimal, hinting at a complete degradation in performance past a certain environment size. Shaded error regions display s.e.m calculated over 5 trials per home.

A.3.2 Additional stable unit visualisations

Here we show an example unit with an inverse home representation that also seems to resemble common trajectories both away from and towards the home location during the explore and home phases, respectively (Figure A3a, b). Additionally, we show an example of another stable unit that remains spatially consistent across both phases (Figure A3c). Although both units represent the home location to some extent, their spatial fields point to different computations being carried out throughout the trial.

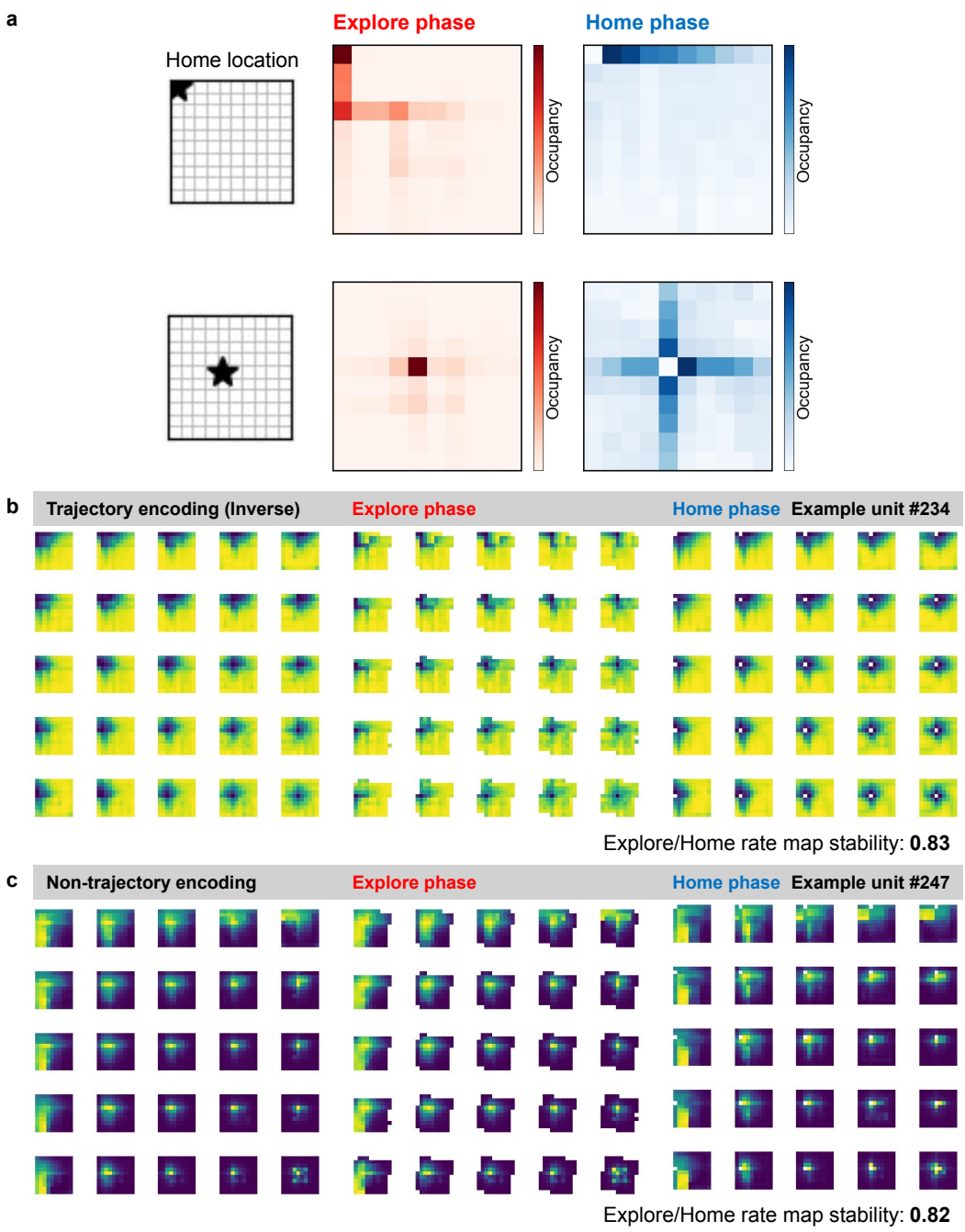


Figure A3: Stable rate maps resemble trajectories or remain consistent across trial phase. (**a**) Occupancy maps reveal structured exploration trajectories and optimal paths home. (**b**) Example unit with inverse activations resembling trajectories in both the explore and home phase. (**c**) Example unit with consistent activation patterns across the explore and home phase.

A.3.3 Within-representation decoding

Decoding of position, reward, and home location is more accurate when activations are sampled from a fixed home location (Figure A4), since remapping units need not generalise across home locations in this experiment. We note that for decoding home location the target is constant, hence the perfect decoding.

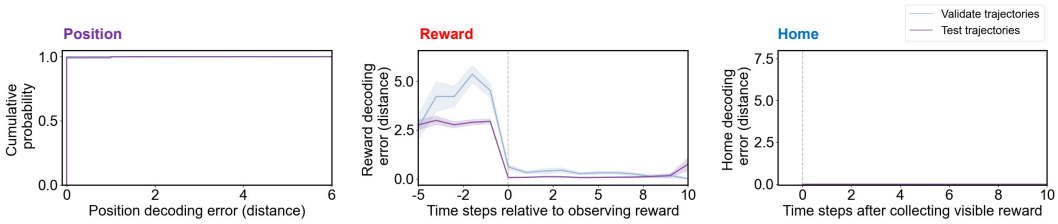


Figure A4: Decoding position, reward, and home is more accurate with a fixed home location. Decoders were trained and evaluated on trajectories from a single home location, with 'test' and 'validate' curves as in Figure 4. (**a**) Cumulative distribution of position decoding error. (**b**) Reward decoding error before and after the reward comes into view (dashed line). (**c**) Home decoding error after reward collection, which is zero because the target is constant. Shading, s.e.m. across 3000 test and 300 validate trajectories.

A.3.4 Alternate real/false home experiment

We perform the same experiment as in Section 5.3 but with the real and false homes swapped. We replicate the same main results: the agent spends far more time near the false home than near the real one (Figure A5a), its occupancy is concentrated around the false home and locations rotationally symmetric to it (Figure A5b), and the proportion of home trajectories ending at the false home increases with the number of units overwritten in order of decreasing signed Wasserstein distance (Figure A5d).

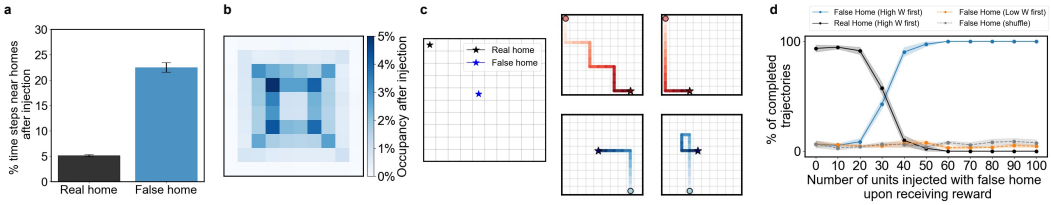


Figure A5: Alternate real/false home experiment. **a** Percentage of time steps spent near the real and false home after the 100 hidden units that remap most were injected with a false home representation, as in Figure 5d but with reversed real/false home (error bars, s.e.m. across 500 repeats). **b** Agent occupancy after injection, as in Figure 5e, agent spends most time in the false home and rotationally symmetric locations. **c** Example trajectories: explore phase from the real home (red), followed by the home phase after injection, terminating at the false home (blue). **d** Percentage of completed home trajectories reaching the real or false home as a function of the number of units overwritten, in order of decreasing signed Wasserstein distance, compared with random order (shading, s.e.m. across 5 repeats).

A.3.5 Spatial representational similarity analysis

Figure A6 shows the spatial RSA scores reported in Section 5.4, averaged across trials from all home locations, for the home phase, the explore phase, and explore-phase time steps at which the reward was not visible.

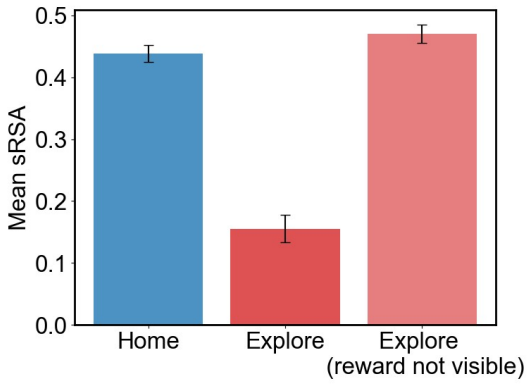


Figure A6: Spatial representational similarity analysis (RSA) score for the home phase, the explore phase, and explore-phase time steps at which the reward was not visible. Error bars, s.e.m. across all 100 home locations.