HIPPOCAMPAL REPRESENTATIONS EMERGE WHEN TRAINING RECURRENT NEURAL NETWORKS ON A MEMORY DEPENDENT MAZE NAVIGATION TASK

Anonymous authors
Paper under double-blind review

ABSTRACT

Can neural networks learn goal-directed behaviour using similar strategies the brain uses to combine the relationships between the current state of the organism and the consequences of future actions? Recent work has shown that recurrent neural networks trained on goal based tasks can develop representations resembling those found in the brain, entorhinal cortex grid cells, for instance. Here we explore the evolution of the dynamics of their internal representations and compare this with experimental data. We observe that once a recurrent network is trained to learn the structure of its environment solely based on sensory prediction, an attractor based landscape forms in the network's representation, which parallels hippocampal place cells in structure and function. Next, we extend the predictive objective to include Q-learning for a reward task, where rewarding actions are dependent on delayed cue modulation. Mirroring experimental findings in hippocampus recordings in rodents performing the same task, this training paradigm causes nonlocal neural activity to sweep forward in space at decision points, anticipating the future path to a rewarded location. Moreover, prevalent choice and cue-selective neurons form in this network, again recapitulating experimental findings. Together, these results indicate that combining predictive, unsupervised learning of the structure of an environment with reinforcement learning can help understand the formation of hippocampus-like representations containing both spatial and task-relevant information.

1 INTRODUCTION

Recurrence based neural networks have been used to perform spatial navigation tasks and the subsequent study of their internal representations has yielded dynamics and structures that are strikingly biological. Metric (Cueva & Wei, 2018; Banino et al., 2018) and non-metric (Recanatesi et al., 2019) representations mimicking grid (Fyhn et al., 2004) and place cells (O'Keefe & Nadel, 1978) respectively form once the recurrent network has learned a predictive task in the context of a complex environment. Cueva et al. (2020) demonstrates not only the emergence of characteristic neural representations, but also anatomical hallmarks of head direction system cells when training a recurrent network on a simple angular velocity integration task. Biologically, non-metric representations are associated with landmark spatial memory, in which place cells within the mammalian hippocampus fire when the associated organism is present in a corresponding place field. Extrafield firing of place cells occurs when these neurons spike outside of these contiguous place field regions. Here we show that recurrent neural networks (RNNs) not only form corresponding attractor landscapes, but also produce representations with internal dynamics that closely resemble those found experimentally in the hippocampus when planning.

Results in Neuroscience such as those of Johnson & Redish (2007), show that spatial representations in mice in the CA3 region of the hippocampus frequently fire nonlocally. Griffin et al. (2007) show that a far higher proportion of hippocampal neurons in the CA1 region in rats performing an episodic task in a T-shaped maze encode the phase of the task than encode spatial information (in this case trajectory direction). Ainge et al. (2007) show CA1 place cells encode destination location at the start position of a maze. Lee et al. (2006) demonstrate that place fields of CA1 neurons gradually drift toward reward locations throughout reward training on a T-shaped maze.

In this work we show that a recurrence based neural network learning a choice-reward based task using reinforcement learning, in conjunction with an environmentally predictive task in a detailed T-shaped maze produces an internal representation with consistent extrafield firing associated with consequential decision points. In addition we find that the network's representation, once trained, follows a forward sweeping pattern as identified by Johnson & Redish (2007). We then go on to show that a higher proportion of units in the trained network show strong selectivity for the encoding or choice phase of the task than the proportion showing selectivity for spatial topology.

2 METHOD

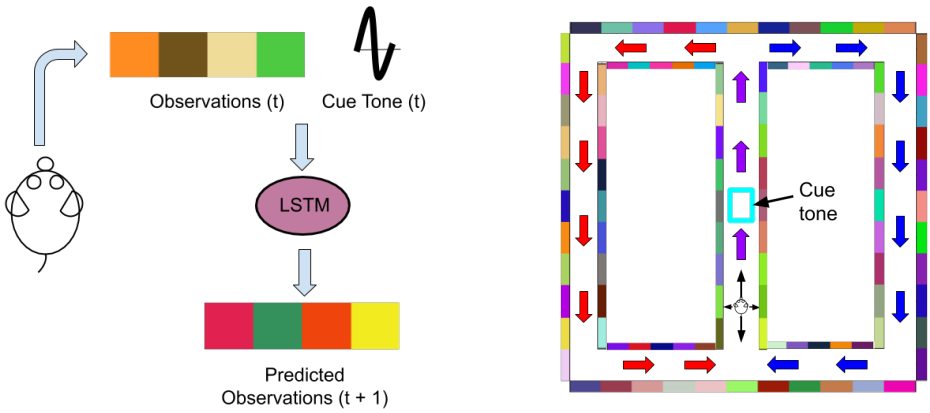

Figure 1: Left, the wall observation and cue received by the network at each timestep. Right, the entangled predictive task the LSTM network is pre-trained on in order to generate a non-metric map of the maze environment.

We use a form of the cued-choice maze used by Johnson & Redish (2007) which has a central T structure with returning arms, shown in Figure 1. All walls of the maze are tiled with distinct RGB colours which are generated at random. An agent is initially learning to predict the next sensory stimulus given its movement. This combination of unsupervised learning and exploration has been shown previously to produce place cell-like encoding of the agent's position (Recanatesi et al., 2019). Next, rewards at four possible locations are introduced and the agent is tasked with associating a cue with the rewarding trajectory. The agent has four vision sensors, one in each cardinal direction, reading the wall RGB colours they intersect. The cue tone is played to the agent as it passes the halfway point of the central maze stem. A low frequency cue indicates that the agent will turn left at the top of the maze stem and a high frequency cue indicates a right turn. These cue tones take the form of a high or low valued scalar perturbed with normally distributed noise if at a cue point, with a zero value given at all other locations. These four RGB colours as well as the cue frequency at the current location make up the total input received by the agent.

The agent is controlled by a recurrent neural network comprised of a 380 unit Long-Short term memory (Hochreiter & Schmidhuber, 1997) (LSTM) network with a readout layer for the prediction of RGB values. We first pre-train the network by tasking it with predicting the subsequent observation of wall colours from the currently observable wall colours given its trajectory through the maze. The agent's starting location is at the bottom of the central stem of the T maze and a trajectory of left or right at the top of the central stem is chosen pseudorandomly, depicted with red and blue arrows respectively in Figure 1 and corresponding to the low (red trajectory) or high (blue trajectory) cue tone value given halfway up the stem. During pre-training the agent does not choose any of its actions and is only learning to predict the sequence of wall colors it encounters. In a given pre-training iteration, we collect all observations as the agent traverses the maze until it returns to the start location at the bottom of the central stem and finally train the LSTM on the entire collected trajectory. The network is trained with a mean-squared error loss of predicted and target wall colours (Eq. 1), with network parameters optimised using Adam (Kingma & Ba, 2015).

$$loss_{rgb} = \frac{1}{n}\sum_{i=1}^{n}(y_{rgb} - (W_{rgb}h_t + b_{rgb}))^2 \tag{1}$$

To solve this task, the network has to maintain the cue tone played in its internal memory for several time steps in order to predict subsequent wall colours from the top of the central stem. In our model, this is achieved through the network forming a non-metric representation (attractor landscape) of the maze environment, as also demonstrated by Xu & Barak (2020). Similarly, behavioural experiments typically have a comparable familiarisation phase with the environment before reward-based tasks are introduced (Johnson & Redish, 2007; Griffin et al., 2007).

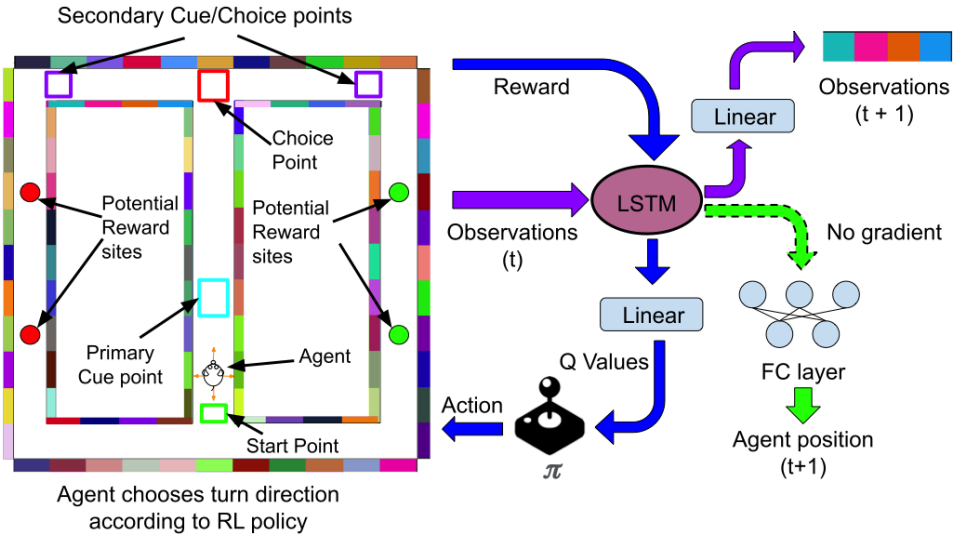

Figure 2: For the joint task to be learned by the LSTM network, we introduce secondary cue points, where the same cue tone as that played at the primary cue point will be repeated if and only if the agent has proceeded in turning in the direction corresponding to the cue tone frequency given at the primary cue location. The agent is free to choose the next action to be taken when traversing the maze at either the choice point at the top of the stem of the maze or at the secondary cue locations. There are two potential reward sites on both returning arms, with the reward sites being active if the agent is on the returning arm corresponding to the cue tone frequency.

Once the LSTM has formed an internal representation of the maze, the agent is tasked with navigating towards potential reward sites, whose activeness is modulated by the cue signal - a low frequency cue indicates active reward sites on the left return arm and a high frequency cue indicates active reward sites on the right return arm - the cue tone and corresponding side of active reward sites are together chosen randomly at each iteration with a secondary cue given if the agent has turned correctly. In this phase there are three choice points wherein the agent is able to choose its next action and is constrained to follow the forward maze direction elsewhere. These choice points are at the top of the maze stem and at the two secondary choice points as shown in Figure 2, with initially random movement at these points during reward training. We introduce a new readout layer for the LSTM network which predicts state-action values associated with the four cardinal directions in relation to the agent's current position and direction. At each timestep, this ensemble receives the agent's environment observation and the agent follows an epsilon-greedy policy (starting with fully random movement at choice points and a decaying epsilon thereafter) for choosing optimal actions at each of the three choice points.

$$Q(s, a) = W_Q h_t + b_Q \tag{2}$$

The recurrent network controlling the agent is trained on a weighted combined loss (Eq. 3) of a reinforcement learning (RL) task loss and the previously described predictive wall colour loss (Eq. 1). The first component of this loss is the difference between predicted and observed state-action values which are represented by Q-values (Eq.2) (Watkins & Dayan, 1992), which are a prediction of future global reward. We use double-Q learning (Van Hasselt et al., 2016) to train the agent on the task, updating the target Q value predictor (Q' - a LSTM with same number of units) every 15 training iterations, with a discount factor (γ) of 0.8. The second loss component is the sensory prediction task which we used to pre-train the network ($\lambda = 0.5$).

$$loss_{combined} = |Q(s, a) - (r + \gamma \cdot Q'(s', \arg\max_{a'} Q(s', a'))| + \lambda \cdot loss_{rgb} \tag{3}$$

In contrast to much of the previous work on spatial representations in recurrent networks, we do not give the network any indication of the agent's location or movement. This makes the task considerably more difficult due to the unpredictable movement possible at choice points during the reward task. The network is coerced into storing the current movement direction of the agent in its representation, in addition to storing the cue frequency. As such, a network of Gated Recurrent Units (Cho et al., 2014) (GRUs) or vanilla RNN units was unable to perform well in either the pre-training or joint RL task due to these prevalent long term dependencies.

To analyse the representations formed by the network, we train a further single layered fully connected network with a softmax layer (shown in green in Figure 2) to predict the agent's next location using the activity of the LSTM. There is no backpropagation of gradients between this predictor and the LSTM network, and the predictor is trained at the end of reward training. The plots in Figures 4 and 5 are distributions indicating the probability of agent location inferred from LSTM activity by the predictor. This is used in place of the decoding algorithm used by Johnson & Redish (2007) to predict the neuronally inferred maze location of rats when performing a cue based task.

3 RESULTS

The agent learns the sensory prediction task to a high degree of recall and after around a thousand training iterations, the agent was able to achieve perfect performance on the reward task. We trained the reinforcement learning portion of the network in an epsilon greedy manner, with a steadily decaying epsilon to ensure that the agent would choose the rewarding path consistently once actions were chosen at choice points completely by the network. Notably, the agent did not turn at either of the secondary choice points once training had completed - only at the primary choice point. Therefore the agent takes the most direct route to the currently active reward sites, thus the forward sweeping representation which emerges (see below; Figure 5), cannot be a signature of training.

We attempted to run the reinforcement learning task alone in a maze with no wall colours or environment statistics except the cue. In this scenario the network is not able to learn the task due to a lack of self-localisation. In addition, the reward based reinforcement learning task was attempted alone, with a loss function that did not include the wall colour prediction error. Here we find that the reward task is also not learnable as the network quickly forgets the attractor landscape of the maze formed during pre-training, which we maintain through the combined loss (Eq. 3). Generally, when the network loses its ability to self-localise the agent, state-action values are no longer reliable indicators of future reward potential as the current environmental state is not clearly discernible.

3.1 EXTRAFIELD PLACE CELL FIRING

First, we investigate the representation learned by the the network during these two stages of training. Pre-training causes the formation of discrete attractors that resemble place cells in the hippocampus. Individual units in the network generally have well isolated place fields, which together cover the whole maze and therefore allow reliable decoding of agent location. In addition to an increase in activity in a particular unit when the agent moves across its respective place field, we also observe substantial extrafield firing of these units. This activity occurs mainly at the primary cue location and at the first choice point after pre-training. After training on the reward task, in addition to the prior spatial correlates, the network also has units with extrafield activity at the secondary choice points (Fig. 3E).

In experiments, rodents seem to pause at high consequence decision points (Johnson & Redish, 2007) with alternating head movement behaviour signifying vicarious trial and error (VTE) (Muenzinger, 1938; Hu & Amsel, 1995). We simulate this using our reward trained model by running the agent from the start position at the bottom of the maze stem, then pausing it at the top of the stem. The network representation seems to sample both return arms, with surprisingly high activity in the displayed (Fig. 3A-D) LSTM units when the agent is present at the maze choice point, for which these units do not usually have corresponding activity (activity above 30% of unit peak average from left and right maze run). We define nonlocal firing as unit activity above 60% of peak averaged unit activity when running the agent along the central stem (bottom row Fig. 3A-D) to observe only the most poignant extrafield behaviour.

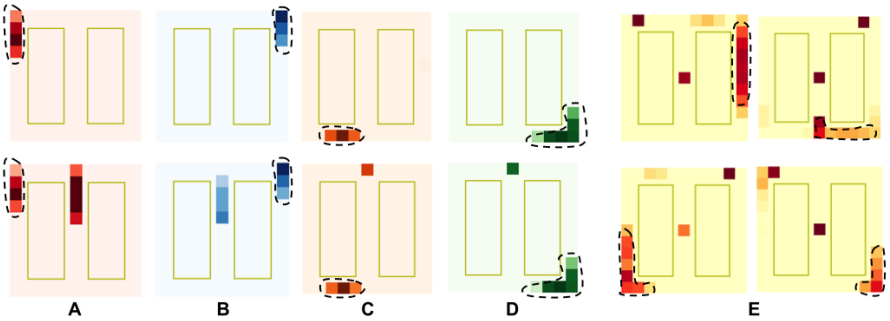

Figure 3: **A**, **B**, **C**, **D**) *Top row*: Activity maps showing well isolated place fields of four LSTM units (acting as place cells) indicated in dotted regions after the reward task. Place fields determined by contiguous locality with average activity exceeding 30% peak unit activity during a single left trajectory followed by a right trajectory. *Bottom row*: LSTM unit activity exceeding 60% of previously averaged peak unit activity for the given neuron when agent run from bottom of maze stem to top of stem and given a low frequency (left) cue tone halfway up the stem, then stationary at choice point with LSTM network repeatedly receiving observations from choice point for timesteps thereafter (shown in addition to previously determined unit place fields in dotted regions). **A** and **B**) Strong extrafield firing contiguously from cue to choice point. **C**, **D**) High extrafield firing at choice point while agent is paused at top of stem. **E**) Place fields (determined from average activity on both trajectories) of four LSTM units outlined in dotted areas after reward based task. High levels of consistent extrafield firing at primary and secondary cue points in 56% of LSTM units.

3.2 Forward moving representation

The internal dynamics of the LSTM network has an inherently forward looking representation of the maze once pre-trained in a predictive manner. As depicted in Fig. 4, whilst the agent is stationary, the dynamics of the LSTM network moves forward through the maze, incorporating the trajectory modulation of the cue played halfway up the maze stem. The forward movement of the representation is also notable for having an inconsistent velocity, where the LSTM inferred agent location jumps (Hasselmo, 2009) from the top of the maze to lower down the arm (timestep 16 in Fig. 4).

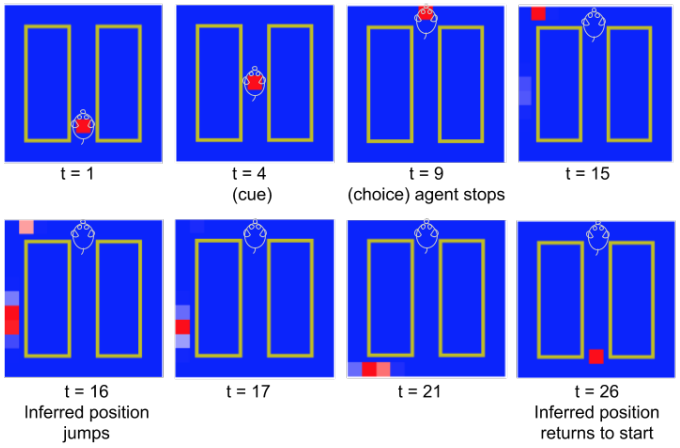

Figure 4: LSTM inferred agent position after pre-training on maze. The agent is run from the start at timestep 1 to timestep 4 where it receives a low frequency cue (indicating a left turn). At timestep 9 the agent is stopped at the top of the maze stem and the LSTM is given the environment observation from this location for the remainder of the shown timesteps. The inferred position then moves left according to the cue with the position seeming to jump abruptly between timesteps 15 and 16. The inferred position then moves back to the starting position at timestep 26. We observe an analogous inferred forward moving representation on the right side of the maze with a high frequency cue.

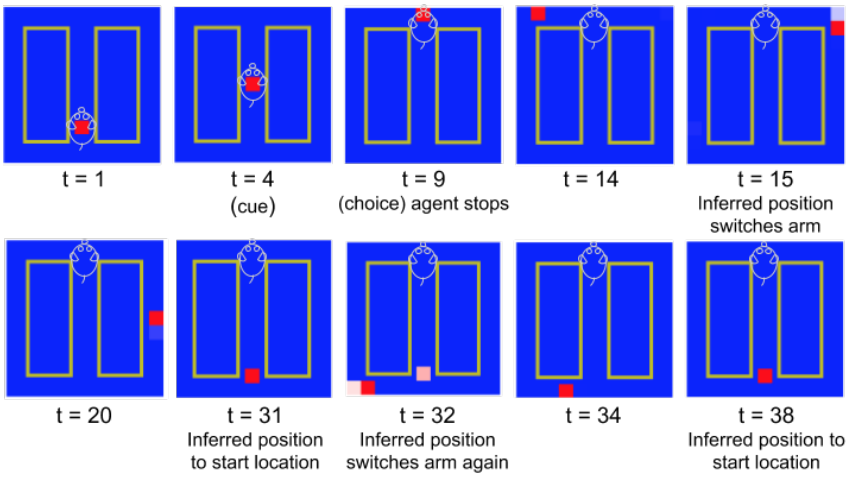

Figure 5: LSTM representation after reward training. As previously, we run the agent from the start position to the top of the stem of the maze at timestep 9 with a low frequency (left) cue tone at timestep 4. Again, the agent is stopped at this position with the LSTM network receiving the environment observations from this position for the remainder of the shown timesteps. As with the network purely trained on the predictive task, the representation moves in the direction corresponding to the frequency of the given cue tone. Then between timesteps 14 and 15, the inferred position jumps from the return arm with active reward sites to the alternate arm, with the inferred position moving from this position to the start location fairly consistently. Then the inferred position jumps again at timestep 32 to the rewarding return arm and moves constantly to the start position.

In stark contrast to the dynamics of the LSTM network following predictive pre-training, following training on the reward task wherein the agent is free to choose the action taken at the three cue points in the maze (the primary and secondary points shown in Figure 2), the forward representation of the LSTM is still clearly looking ahead of the agent but is now displaying sweeping behaviour which is identified experimentally in rats by Johnson & Redish (2007) when performing modulation based tasks. When the agent is stationary at the choice point, we observe the representation moving ahead of the agent - first in the direction corresponding to the cue frequency given at the first cue point and then abruptly down the opposing arm of the maze towards the starting location, thereafter the representation moves down the correct arm (corresponding to the cue) and becomes stationary at the maze start location. The agent appears to be sampling the trajectory concerning the alternate return arm of the maze before ultimately settling on the rewarding return arm.

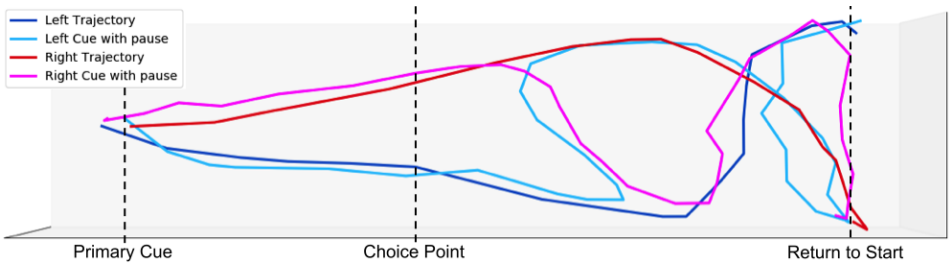

Figure 6: UMAP manifold of LSTM network dynamics of complete left trajectory (dark blue) and complete right trajectory (red) shown along with manifold of dynamics when agent run from start location to choice point with left cue (light blue) and right cue (pink) given at cue point and agent paused in place at the top of the maze stem. A few timesteps after the agent is paused, the dynamics of the left cue paused agent (light blue) switches manifold path abruptly from running alongside the complete left trajectory path (blue) and joins the right trajectory path (red), following this for many timesteps before ultimately resulting at the same manifold end position as the complete left trajectory manifold path (blue). This is analogous for the right cue paths (red and pink).

We further investigate the network representation using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for dimension reduction. Figure 6 shows generally connected manifolds, with closer inspection revealing the dynamics which leads to the sweeping arm behaviour in Figure 5 when the agent is stationary at the primary choice point.

Separately, we find that place fields of particular LSTM units drift forwards from their original firing positions after pre-training, towards the reward locations on the return arms throughout reward training, as shown experimentally in CA1 neurons in Lee et al. (2006). We observe this behaviour in 50 out of 380 network units (13%), with final resting locations of place fields at reward locations (seen in Appendix Figure 9). This is possibly explained by the gradient of Q values (prediction of predicted reward) spreading backwards from reward locations (Hasselmo, 2005) and becoming stronger throughout training.

3.3 Selectivity of neuronal units

In addition to a forward sweeping representation, this trained network also exhibits neural activity that acutely parallels hippocampal circuits, as demonstrated by Griffin et al. (2007), where the authors show that these circuits tend to be much more strongly selective for the encoding or choice phase of a task rather than the direction of the organism's trajectory. We garner the preference of selectivity of each neuronal unit in our network using a discrimination index used by Griffin et al. (2007) for the turn direction selectivity (DI_{turn}) and the phase selectivity (DI_{phase}):

$$DI_{\text{turn}} = \frac{FR_{\text{right}} - FR_{\text{left}}}{FR_{\text{right}} + FR_{\text{left}}} \tag{4}$$

$$DI_{\text{phase}} = \frac{FR_{\text{cue}} - FR_{\text{choice}}}{FR_{\text{cue}} + FR_{\text{choice}}} \tag{5}$$

where FR_{right} for a particular LSTM unit is the mean firing rate from the cue point on the central stem to the choice point at the top of the stem on trajectories where the agent turns right at the choice point. Similarly FR_{left} is the mean stem firing rate when the agent turns left. FR_{cue} is the firing rate at the cue (encoding) point averaged over both left and right trajectories and similarly FR_{choice} is the averaged firing rate at the choice (sampling) point.

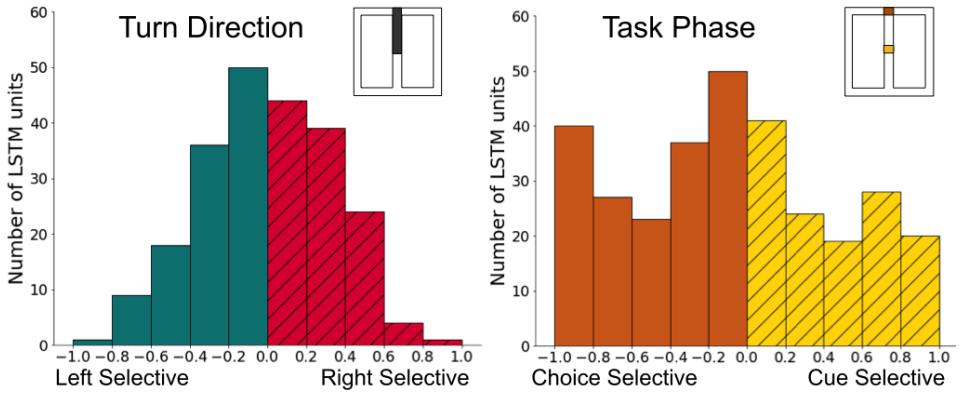

Figure 7: Histograms showing LSTM unit discrimination index for turn direction selectivity (DI_{turn}) vs task phase selectivity (DI_{phase}). A highly negative selectivity index for turn direction indicates a neuronal unit which exhibits high levels of selectivity (uniquely high network activity) for a leftward trajectory and a highly positive selectivity index indicates selectivity for a rightward trajectory. A negative selectivity for task phase indicates a neuron which is highly selective for the choice (retrieval) phase of the goal based task whereas a positive index indicates a neuron which is highly selective for the cue (encoding) phase of the task.

The firing areas used for selectivity measurement are insets in Figure 7. This disparity in areas chosen for selectivity scrutiny is due to the entirety of the stem above the cue point being used for dynamics related to turn direction selectivity, whereas we aim to isolate the encoding and sampling

phases of the task by examining the cue and choice points individually for DI_{phase}. Figure 7 shows a higher proportion of LSTM units are strongly task selective than strongly turn selective, with significantly more LSTM units having extreme DI_{phase} indices than extreme DI_{turn} indices.

In addition, the reward trained network is found to have a disproportionately high number of units (163 of 380 LSTM units) with place fields at the start location of the maze. Moreover, we find evidence of conditional destination encoding in these units which were heavily differentiated in their firing with respect to particular rewarding locations, as shown experimentally in CA1 hippocampal place cells (Ainge et al., 2007; Wood et al., 2000; Ferbinteanu & Shapiro, 2003). 59.5% of units with a place field at the maze start location fired uniquely at this point for rewarding locations on a particular return arm.

4 Discussion

In this work we show that networks trained with a combined predictive and goal-based objective exhibit functional dynamics and selectivity behaviour coinciding with that of hippocampal neurons. We demonstrate that extrafield firing activity of network units emerge when a simulated agent, which is trained on a goal based reward task in a T-shaped maze, pauses at decision points - suggesting intrinsic dynamics are encoding the future planned trajectory of the agent. This mirrors experimental results in hippocampal place cells in rats (Johnson & Redish, 2007; Frank et al., 2000).

Previous work shows that metric neural representations of environments form when an RNN is optimised to predict agent position from agent velocity (Cueva & Wei, 2018; Banino et al., 2018) and non-metric representations form when an RNN is trained on a predictive task with implicit agent velocity given (Recanatesi et al., 2019). When training our model we do not give the LSTM network any explicit notion of current movement direction or location, it only receives sensory information. This is similar to the purely contextual input received by the model pre-trained by Xu & Barak (2020) where no velocity input is given, however, the network here is still trained on position and landmark prediction in a supervised way.

Instead, our training paradigm coerces the LSTM to maintain an implicit notion of current velocity within its internal state in relation to environmental observations. This, in conjunction with the consideration that model-free RL methods such as Q-learning perform poorly on tasks in dynamic environments such as ours (Dolan & Dayan, 2013), and the long term dependency on the delayed cue in perspective of the choice location, makes the task outlined in Figure 2 particularly challenging.

Training on a sensory predictive task causes the formation of a non-metric place cell like representation in the activations of network units, similarly to Recanatesi et al. (2019). These units demonstrate nonlocal extrafield firing, as shown in Appendix Figure 8 and after reward training in Figure 3. Johnson & Redish (2007) find that this extrafield firing is particularly striking at consequential decision points where rats usually pause in order to sample previously seen trajectories. We observe that cue or choice point extrafield activity is evident in most LSTM units after training on the reward task. This is likely due to the increased precedence these points have in the agent reaching reward locations. Together the trained LSTM network units form a representation which sweeps along the paths available to the agent, first down the reward path and then the other, as shown in Figure 5 and demonstrated in rats in Johnson & Redish (2007).

Although place cells such as CA3 cells have been extensively shown to be critical to spatial memory, especially in terms of consolidation (Nakazawa et al., 2002; Florian & Roullet, 2004; Sandi et al., 2003; Redish & Touretzky, 1998; Miller et al., 2020), it is currently unclear by what mechanism an ensemble of place cells contributes to a representation of a rat's goal location (Morris, 1990). Our model and training paradigm is in keeping with the hypothesis that the hippocampus is involved in maintaining a conjunctive representation of cognitive maps and sensory information (Whittington et al., 2019). We show that this paradigm can be extended with predictive learning of Q-values of anticipated future reward, and show that the resulting representation is well suited for learning actions leading from a cue to a reward. Importantly, this representation emerges solely from sampling sensory inputs and predicted rewards, while reinforcement learning itself remains model-free and is initially random. The surprising similarity of the task-dependent activity in our simulations and experimentally recorded neural activity in similar tasks suggests that the model may replicate central aspects of learning and planning in the hippocampus.

REFERENCES

James A. Ainge, Minija Tamosiunaite, Florentin Woergoetter, and Paul A. Dudchenko. Hippocampal CA1 place cells encode intended destination on a maze with multiple choice points. *Journal of Neuroscience*, 27(36), 2007. ISSN 02706474. doi: 10.1523/JNEUROSCI.2011-07.2007.

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*, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0102-6.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014. ISBN 9781937284961. doi: 10.3115/v1/d14-1179.

Christopher J. Cueva and Xue Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2018.

Christopher J. Cueva, Peter Y. Wang, Matthew Chin, and Xue-Xin Wei. Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HklSeREtPB`.

Ray J. Dolan and Peter Dayan. Goals and habits in the brain, 2013. ISSN 08966273.

Janina Ferbinteanu and Matthew L. Shapiro. Prospective and retrospective memory coding in the hippocampus. *Neuron*, 40(6), 2003. ISSN 08966273. doi: 10.1016/S0896-6273(03)00752-9.

Cédrick Florian and Pascal Roullet. Hippocampal CA3-region is crucial for acquisition and memory consolidation in Morris water maze task in mice. *Behavioural Brain Research*, 154(2), 2004. ISSN 01664328. doi: 10.1016/j.bbr.2004.03.003.

Loren M. Frank, Emery N. Brown, and Matthew Wilson. Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, 27(1), 2000. ISSN 08966273. doi: 10.1016/S0896-6273(00)00018-0.

Marianne Fyhn, Sturla Molden, Menno P. Witter, Edvard I. Moser, and May Britt Moser. Spatial representation in the entorhinal cortex. *Science*, 2004. ISSN 00368075. doi: 10.1126/science.1099901.

Amy L. Griffin, Howard Eichenbaum, and Michael E. Hasselmo. Spatial representations of hippocampal CA1 neurons are modulated by behavioral context in a hippocampus-dependent memory task. *Journal of Neuroscience*, 27(9), 2007. ISSN 02706474. doi: 10.1523/JNEUROSCI.4083-06.2007.

Michael E. Hasselmo. A model of prefrontal cortical mechanisms for goal-directed behavior. *Journal of Cognitive Neuroscience*, 17(7), 2005. ISSN 0898929X. doi: 10.1162/0898929054475190.

Michael E. Hasselmo. A model of episodic memory: Mental time travel along encoded trajectories using grid cells. *Neurobiology of Learning and Memory*, 92(4), 2009. ISSN 10747427. doi: 10.1016/j.nlm.2009.07.005.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.

Dan Hu and Abram Amsel. A simple test of the vicarious trial-and-error hypothesis of hippocampal function. *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), 1995. ISSN 00278424. doi: 10.1073/pnas.92.12.5506.

Adam Johnson and A. David Redish. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45), 2007. ISSN 02706474. doi: 10.1523/JNEUROSCI.3761-07.2007.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Inah Lee, Amy L Griffin, Eric A Zilli, Howard Eichenbaum, and Michael E Hasselmo. Gradual Translocation of Spatial Correlates of Neuronal Firing in the Hippocampus toward Prospective Reward Locations. *Neuron*, 51(5):639–650, 2006. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2006.06.033. URL `http://www.sciencedirect.com/science/article/pii/S0896627306005836`.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), 2018. ISSN 2475-9066. doi: 10.21105/joss.00861.

Thomas D. Miller, Trevor T.J. Chong, Anne M.Aimola Davies, Michael R. Johnson, Sarosh R. Irani, Masud Husain, Tammy W.C. Ng, Saiju Jacob, Paul Maddison, Christopher Kennard, Penny A. Gowland, and Clive R. Rosenthal. Human hippocampal CA3 damage disrupts both recent and remote episodic memories. *eLife*, 9, 2020. ISSN 2050084X. doi: 10.7554/eLife.41836.

RGM Morris. Does the hippocampus play a disproportionate role in spatial memory. *Discussions in Neuroscience*, 6:39–45, 1990.

Karl F. Muenzinger. Vicarious Trial and Error at a Point of Choice: I. A General Survey of its Relation to Learning Efficiency. *Pedagogical Seminary and Journal of Genetic Psychology*, 53 (1), 1938. ISSN 08856559. doi: 10.1080/08856559.1938.10533799.

Kazu Nakazawa, Michael C. Quirk, Raymond A. Chitwood, Masahiko Watanabe, Mark F. Yeckel, Linus D. Sun, Akira Kato, Candice A. Carr, Daniel Johnston, Matthew A. Wilson, and Susumu Tonegawa. Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science*, 297(5579), 2002. ISSN 00368075. doi: 10.1126/science.1071795.

J. O'Keefe and L. Nadel. *The hippocampus as a cognitive map*. Clarendon Press, Oxford, United Kingdom, 1978.

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Signatures of low-dimensional neural predictive manifolds. *Cosyne Abstracts 2019, Lisbon, PT.*, 2019.

A. David Redish and David S. Touretzky. The Role of the Hippocampus in Solving the Morris Water Maze. *Neural Computation*, 10(1), 1998. ISSN 08997667. doi: 10.1162/089976698300017908.

Carmen Sandi, Heather A. Davies, M. Isabel Cordero, Jose J. Rodriguez, Victor I. Popov, and Michael G. Stewart. Rapid reversal of stress induced loss of synapses in CA3 of rat hippocampus following water maze training. *European Journal of Neuroscience*, 17(11), 2003. ISSN 0953816X. doi: 10.1046/j.1460-9568.2003.02675.x.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-Learning. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016. ISBN 9781577357605.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4), 1992. ISSN 0885-6125. doi: 10.1007/bf00992698.

James CR Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalisation in the hippocampal formation. *bioRxiv*, pp. 770495, September 2019. doi: 10.1101/770495.

Emma R. Wood, Paul A. Dudchenko, R. Jonathan Robitsek, and Howard Eichenbaum. Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron*, 27(3), 2000. ISSN 08966273. doi: 10.1016/S0896-6273(00)00071-4.

Tie Xu and Omri Barak. Implementing inductive bias for different navigation tasks through diverse rnn attrractors. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Byx4NkrtDS`.

A Appendix

A.1 Extrafield place cell firing after sensory prediction task

After pre-training on the sensory prediction task outlined in Figure 1, we observe that when the agent is paused at the top of the stem of the maze the network representation moves far ahead of the agent, caused by extrafield activity of many neurons in the network.

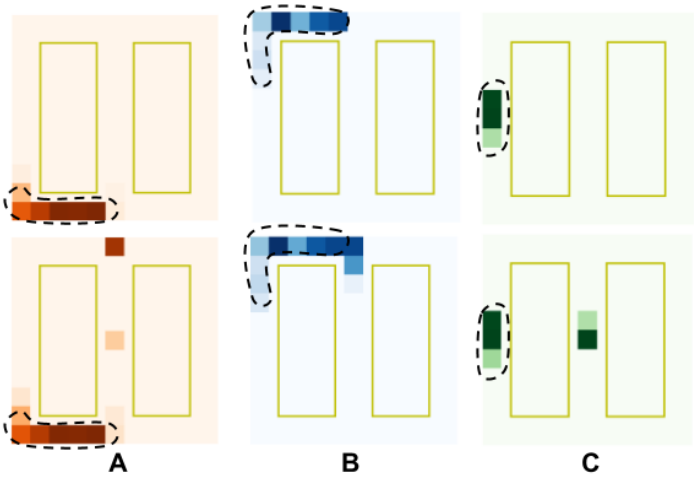

Figure 8: **A**, **B**, **C**) Top row: well isolated place fields of three LSTM units indicated in dotted regions after pre-training. Place fields determined by contiguous locality with average activity exceeding 30% peak field activity during a single left trajectory followed by a right trajectory. Bottom row: agent run from bottom of maze stem to top of stem (and given a low frequency cue tone halfway up the stem) and paused at choice point with LSTM network repeatedly receiving observations from choice point for timesteps thereafter. **A**) Strong extrafield firing at choice point with some activity at the cue point. **B**) Extrafield activity at choice point and at position below. **C**) High extrafield firing at cue point before agent pauses at top of stem.

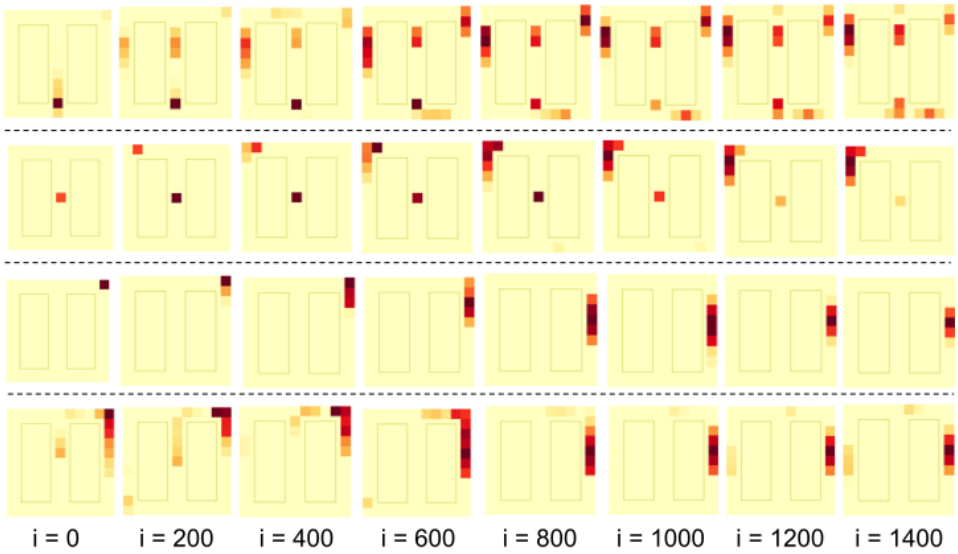

Figure 9: Place fields of four LSTM units, starting from $i = 0$ where the network has been pre-trained on the sensory prediction task, drifting forwards towards reward locations throughout reward training (where i is the number of training iterations). The place fields ultimately rest at maze reward locations at the end of reward training ($i = 1400$).