Generalisation of structural knowledge in the hippocampal-entorhinal system

James C.R. Whittington*
University of Oxford, UK
james.whittington@magd.ox.ac.uk

Timothy H. Muller*
University of Oxford, UK
timothymuller127@gmail.com

Shirley Mark
University College London, UK
s.mark@ucl.ac.uk

Caswell Barry
University College London, UK
caswell.barry@ucl.ac.uk

Timothy E.J. Behrens
University of Oxford, UK
behrens@fmrib.ox.ac.uk

Abstract

A central problem to understanding intelligence is the concept of generalisation. This allows previously learnt structure to be exploited to solve tasks in novel situations differing in their particularities. We take inspiration from neuroscience, specifically the hippocampal-entorhinal system known to be important for generalisation. We propose that to generalise structural knowledge, the representations of the structure of the world, i.e. how entities in the world relate to each other, need to be separated from representations of the entities themselves. We show, under these principles, artificial neural networks embedded with hierarchy and fast Hebbian memory, can learn the statistics of memories and generalise structural knowledge. Spatial neuronal representations mirroring those found in the brain emerge, suggesting spatial cognition is an instance of more general organising principles. We further unify many entorhinal cell types as basis functions for constructing transition graphs, and show these representations effectively utilise memories. We experimentally support model assumptions, showing a preserved relationship between entorhinal grid and hippocampal place cells across environments.

1 Introduction

Animals have a remarkable ability to flexibly take knowledge from one domain and generalise it to another. This is not yet the case for machines. The advantages of generalising knowledge are clear - it allows one to make quick inferences in new situations, without having to always learn afresh. Generalisation of statistical structure (the relationships between objects in the world) imbues an agent with the ability to fit things/concepts together that share the same statistical structure, but differ in the particularities, e.g. when one hears a new story, they can fit it in with what they already know about stories in general, such as there is a beginning, middle and end - when the funny story appears while listening to the the news, it can be inferred that the programme is about to end.

Generalisation is a topic of much interest. Advances in machine learning and artificial intelligence (AI) have been impressive [23, 28], however there is scepticism over whether 'true' underlying structure is being learned. We propose that in order to learn and generalise structural knowledge, this structure must be represented explicitly, i.e. separated from the representations of sensory objects in

the world. In worlds that share the same structure but differ in sensory objects, explicitly represented structure can be combined with sensory information in a conjunctive code unique to each environment. Thus sensory observations are fit with prior learned structural knowledge, leading to generalisation.

In order to understand how we may construct such a system, we take inspiration from neuroscience. The hippocampus is known to be important for generalisation, memory, problems of causality, inferential reasoning, transitive reasoning, conceptual knowledge representation, one-shot imagination, and navigation [13, 7, 17, 25]. We propose the statistics of memories in hippocampus are extracted by cortex [27], and future hippocampal representations/memories are constrained to be consistent with the learned structural knowledge. We find this an interesting system to model using artificial neural networks (ANNs), as it may offer insights into generalisation for machines, further our understanding of the biological system itself, and continue to link neuroscience and AI research [18, 38].

1.1 Background

In spatial navigation there is a good understanding of neuronal representations in both hippocampus (e.g. place, landmark cells) and medial entorhinal cortex (e.g. grid, border, object vector cells). Thus when modelling this system, we start with problems akin to navigation so we can both leverage and compare our results to these known representations (noting our approach is more general). Place [29] and grid cells [16] have had a radical impact in neuroscience, leading to the 2014 Nobel Prize in Physiology and Medicine. Place and grid cells are similar in that they have a stable firing pattern for specific regions of space. Place cells tend to only fire in a single (or couple) location in a given environment, whereas grids cells fire in a regular lattice pattern tiling the space. These cells cemented the idea of a 'cognitive map', where an animal holds an internal representation of the space it navigates [36]. Traditionally these cells were believed to be spatial only. It has since emerged that place and grid cells code for both spatial and entirely non-spatial dimensions such as sound frequency [2], and furthermore grid-like codes for two dimensional (2D) non-spatial coordinate systems exist [9]. It therefore seems that place and grid codes may provide a general way of representing information. Other entorhinal cell types (border [31], object vector cells [19]) appear to have disparate roles in coding space. Here we unify them, along with grid cells, as basis functions for transition statistics.

Grid cells may offer a generalisable structural code. Indeed grid cell representations are similar in environments that share structure ([14], section 5). Recent results suggest such codes summarise statistics of 2D space, either via a PCA of hippocampal place cells [12] or as eigenvectors of the successor representation [32]. These summary statistics represent rules of 2D-ness (not just 'spatial' space), e.g. if A is close to B, and B is close to C, we can infer A and C are close. Place cells may offer a conjunctive representation. Their activity is modulated by the sensory environment as well as location [39, 22]. Additionally the place cell code is different for two structurally identical environments - this is called remapping [6, 26]. Remapping is traditionally thought to be random. However, we propose place cells form a conjunctive representation between structural (grid cells) and sensory input, and therefore remap to non-random locations consistent with this conjunction.

1.2 Contributions

We implement our proposal in an ANN tasked with predicting sensory observations when walking on 2D graph worlds, where each vertex is associated with a sensory experience. To make accurate predictions, the agent should learn the underlying hidden structure of the graphs. We separate structure from sensory identity, proposing grid cells encode structure, and place cells form a conjunctive representation between sensory identity and structure (Fig 1a). This conjunctive representation forms a Hebbian memory, which bridges structure and identity, allowing the same structural code to be reused across environments that share statistics but differ in sensory experiences. We combine fast Hebbian learning of episodic memories, with gradient descent which slowly learns to extract statistics of these memories. Our network learns representations that mirror those found in the brain, with different entorhinal-like representations forming depending on transition statistics. We further present analyses of a remapping experiment [5], which support our model assumptions, showing place cells remap to locations consistent with a grid code, i.e. not randomly as previously thought.

The key conceptual novelties are as follows. **Neuroscience:** We found an interpretation of grid cells, place cells and remapping that offers a mechanistic understanding for the hippocampal involvement in generalisation of knowledge across domains. We provide a unifying framework for many entorhinal

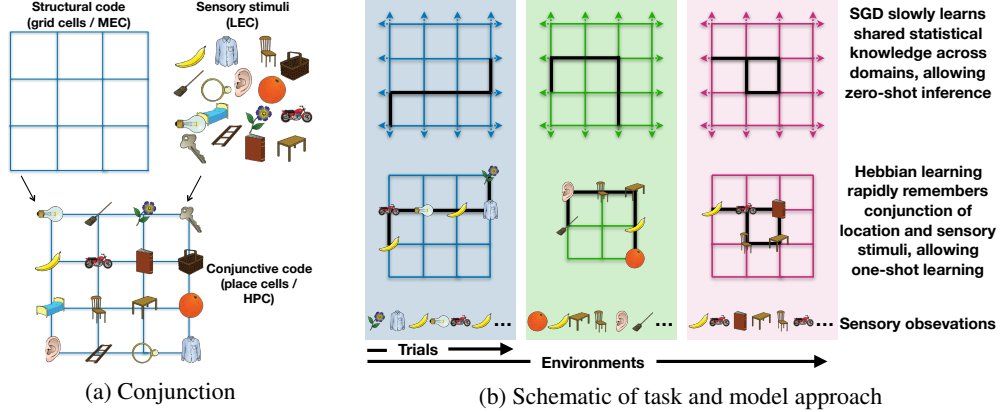| (a) Conjunction | (b) Schematic of task and model approach |

Figure 1: (a) Separated structural and sensory representations combined in a conjunctive code. LEC/MEC: Lateral/Medial entorhinal cortex, HPC: Hippocampus. (b) Problem the model faces - extracting generalisable statistics across domains, while rapidly learning the map within domain.

cell types (grid cells, border cells, object vector cells) as building basis functions for transitions statistics. Our results suggest spatial representations found in the brain may be an instance of a more general coding mechanism organising knowledge across multiple domains. **Machine learning:** We have built a network where fast Hebbian learning interacts with slow statistical learning. This allows us to learn representations whereby memories not only stored in a Hebbian network for one-shot retrieval within domain, but also benefit from statistical knowledge that is shared across domains - allowing zero shot inference.

2   Related work

Concurrently developed papers discovered grid-like and/or place-like representations in ANNs [4, 10]. Neither paper uses memory or explains place cell phenomena. Both, however, use *supervised* learning in order to discover these representations, either supervising on actual x, y coordinates [10] or ground truth place cells [4]. We use *unsupervised* learning, providing the network with only sensory observations and actions. This is information available to a biological agent, unlike ground truth spatial representations. We further propose a role for grid cells in generalisation, not just navigation.

Our modelling approach is simliar to [15 **?**]. However, we choose our memory storage and addressing to be computationally biologically plausible (rather than using other types of differentiable memory more akin to RAM), as well as using hierarchical processing. This enables our model to discover representations that are useful for both navigation and addressing memories. We also are explicit in separating out the abstract structure of the space from any specific content (Fig 1a).

We follow a similar ideology to complementray learning systems [27] where the statistics of memories in hippocampus are extracted by cortex. We additionally propose that this learnt structural knowledge constrains hippocampal representations in new contexts, allowing reuse of learnt knowledge.

3   Model

We consider an agent passively moving on a 2D graph (Fig 1b), observing a non-unique sensory stimulus (e.g. an image) on each vertex. If the agent wishes to 'understand' its environment then it should maximise its model's probability of observing each stimulus. The agent is trained on many environments sharing the same structure, i.e. 2D graphs, however the stimulus distribution is different (each vertex is randomly assigned a stimulus). There are various approaches to this problem, however a generalisable one should exploit the underlying structure of the task - the 2D-ness of the space. One such approach is to have an abstract representation of space encoding relative locations, and then to place a memory of what stimulus was observed at that (relative) location. Since the agent understands where it is in space, this allows for accurate state predictions to previously visited nodes even if the agent has never travelled along that particular edge before (e.g. loop closure in Figs 1b pink and 2c). Although we consider 2D graphs to compare learned representations to those found in the brain, our approach is appropriate for generalising other stuctural/relational/conceptual knowledge [25].

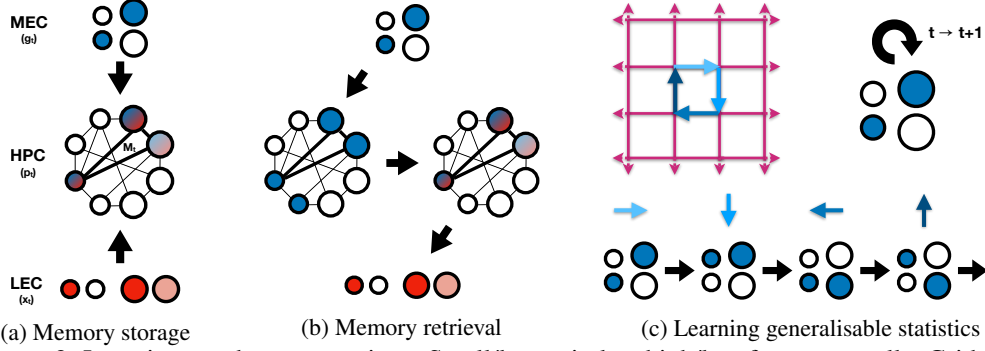| (a) Memory storage | (b) Memory retrieval | (c) Learning generalisable statistics |

Figure 2: Learning good representations. Small/large circles: high/low frequency cells. Grid cells (MEC) need to create (a) conjunctive Hebbian memories (in HPC, weights M_t between place cells) such that the same grid code can reinstate (b) the same memory via attractor dynamics. Grid cells are recurrent (c), and so must learn transition weights such that they have the same code when returning to a state (loop closure). This code must be general enough to work across many environments.

We propose grid cells as bases for constructing abstract representations of space, and place cell representations for the formation of fast episodic memories (Fig 2a). To link a stimulus to a given (relative) location, a memory should be a conjunction of abstract (relative) location and sensory stimulus, thus we propose place cells form a conjunctive representation between the sensorium and grid input (Figs 1a and 2a). This is consistent with experimental evidence [39, 22]. We posit that this is done hierarchically across spatial frequencies, such that the higher frequency statistics can be repeatedly used across space. This reduces the number of weights that need to be learnt. This proposition is consistent with the hierarchical scales observed across both grid cells [34] and place cells [21], and with the entorhinal cortex receiving sensory information in hierarchical temporal scales [37]. We consider grid cells to be recurrent through time, allowing predictive state transitions to occur via grid cells (Fig 2c). This is consistent with grid codes being a natural basis for navigation in 2D spaces [33, 8].

We view the hippocampal-entorhinal system as one that performs inference. Grid cells make inferences based on their previous estimate of location in abstract space (and optionally sensory information linked to previous locations via memory). Place cells, a conjunction between the sensory data and location in abstract space, are stored as memories. We consider sensory data, the item/object experience of a state, as coming from the 'what stream' via lateral entorhinal cortex. The grid cells in our model, are the 'where stream' coming from medial entorhinal cortex (Fig 2). Our hippocampal conjunctive memory links 'what' to 'where', such that when we revisit 'where' we remember 'what'.

3.1 Model summary

The model is a neural network and learns structure across tasks. We optimise end-to-end via backpropagation through time. The central (attractor) network employs Hebbian learning to rapidly remember the conjunction of location and sensory stimulus. A generative temporal model learns how to use the Hebbian memory most efficiently given the common statistics of transitions across worlds.

3.2 Generative model

We consider the agent to have a generative model (Fig 3a, schematic in Figs 2b, 2c) which factorises as $p_\theta \left(\mathbf{x}_{\leq T}, \mathbf{p}_{\leq T}, \mathbf{g}_{\leq T} \right) = \prod_{t=1}^{T} p_\theta \left(\mathbf{x}_t \mid \mathbf{p}_t \right) p_\theta \left(\mathbf{p}_t \mid M_{t-1}, \mathbf{g}_t \right) p_\theta \left(\mathbf{g}_t \mid \mathbf{g}_{t-1}, \mathbf{a}_t \right)$ where observed variable $\mathbf{x}_t$ is the instantaneous sensory stimulus and latent variables $\mathbf{g}_t$ and $\mathbf{p}_t$ are grid and place cells. M_t represents the agent's memory composed from past place cell representations. $\mathbf{a}_t$ represents the current action - our version of *head-direction* cells [35]. θ are parameters of the generative model.

We now give concise, but intuitive descriptions of the model components. Expanded details in Supplementary Materials (SM). Sensory data $\mathbf{x}_t$ is a one-hot vector where each of its n_s elements represent a sensory identity. We consider place and grid cells, $\mathbf{p}_t$ and $\mathbf{g}_t$ respectively, to come in different frequencies (hierarchies) indexed by superscript f. Though we already refer to these variables as grid and place cells, it is important to note that grid-ness and place-ness are not hard-coded - all representations are learned. $f(\cdots)$ denotes functions specific to the distribution in question.

4

Grid cells. To predict where we will be, we can transition from our current location based on our heading (i.e. path integration, schematic in Fig 2c). $p_\theta(\mathbf{g}_t \mid \mathbf{g}_{t-1}, \mathbf{a}_t)$ is a Gaussian transition probability density, with transitions taking the form $\mathbf{g}_t = f_{\mu_g}(\mathbf{g}_{t-1} + D_a \mathbf{g}_{t-1}) + \sigma \cdot \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$, $Vec[D_a] = f_D(\mathbf{a}_t)$ and $\sigma = f_{\sigma_g}(\mathbf{g}_{t-1})$. Connections in D_a are from low frequency to the same or higher frequency only (or alternatively only within frequency). We separate into hierarchical scales so that high frequency statistics can be reused across lower frequency statistics, i.e. learning and knowledge is reused across space.

Place cells. $p_\theta(\mathbf{p}_t \mid M_{t-1}, \mathbf{g}_t)$ is a Gaussian probability density for retrieving memories. Stored memories are extracted via an attractor network (Fig 2b) using $f_g(\mathbf{g}_t)$ as input - i.e. grid cells act as an index for memory extraction (Details in Section 3.4).

Data. We classify a stimulus identity using $p_\theta(\mathbf{x}_t \mid \mathbf{p}_t) \sim Cat(f_x(\mathbf{p}_t))$.

(a) Generative model   (b) Inference network

Figure 3: Circled/ boxed variables are stochastic/ deterministic. Red arrows indicate additional inference dependencies. Dashed arrows continue through time. Dotted arrow are optional as explained below.
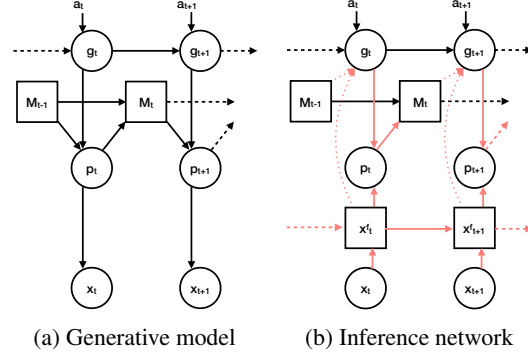
3.3 Inference network

Due to the inclusion of memories, as well as other non-linearities, the posterior $p(\mathbf{g}_t, \mathbf{p}_t \mid \mathbf{x}_{\leq t}, \mathbf{a}_{\leq t})$ is intractable - we therefore turn to approximate inference. To infer on this generative model, we make critical decisions that respect our proposal of structural information separated from sensory information as well as respecting biological considerations. We use a recognition distribution that factorises as $q_\phi(\mathbf{g}_{\leq T}, \mathbf{p}_{\leq T} \mid \mathbf{x}_{\leq T}) = \prod_{t=1}^T q_\phi(\mathbf{g}_t \mid \mathbf{x}_{\leq t}, M_{t-1}, \mathbf{g}_{t-1}) q_\phi(\mathbf{p}_t \mid \mathbf{x}_{\leq t}, \mathbf{g}_t)$. See Fig 3b for inference network schematic. ϕ denote parameters of the inference network. We learn θ and ϕ, by maximising the ELBO with the variational autoencoder framework [20, 30] (details in SM).

Place cells. We treat these variables as a conjunction between sensorium and structural information from grid cells (Fig 2a). The sensorium is obtained by first compressing the immediate sensory data, $\mathbf{x}_t$, to $f_c(\mathbf{x}_t)$, after which it is filtered via exponential smoothing into different frequency bands, $\mathbf{x}_t^f$. After a normalisation step, each $\mathbf{x}_t^f$ is combined conjunctively with $\mathbf{g}_t^f$ to give the mean of the distribution $q_\phi(\mathbf{p}_t \mid \mathbf{x}_{\leq t}, \mathbf{g}_t)$. The separation into hierarchical scales helps to provide a unique code for each position, even if the same stimulus appears in several locations of one environment, since the surrounding stimuli, and therefore the lower frequency place cells, are likely to be different. Since the place cell representations form memories, one can utilise the hierarchical scales for memory retrieval. We note that although the exponential smoothing appears over-simplified, it approximates the Laplace transform with real coefficients. Cells of this nature have been discovered in LEC [37].

Grid cells. We factorise $q_\phi(\mathbf{g}_t \mid \mathbf{x}_{\leq t}, M_{t-1}, \mathbf{g}_{t-1})$ as $q_\phi(\mathbf{g}_t \mid \mathbf{g}_{t-1}, \mathbf{a}_t) q_\phi(\mathbf{g}_t \mid \mathbf{x}_{\leq t}, M_{t-1})$. To know where we are, we can path integrate ($q_\phi(\mathbf{g}_t \mid \mathbf{g}_{t-1}, \mathbf{a}_t)$ - equivalent to the generative distribution described above) as well as use sensory information that we may have seen previously ($q_\phi(\mathbf{g}_t \mid \mathbf{x}_{\leq t}, M_{t-1})$). The second distribution (optional addition) provides information on location given the sensorium. Since memories link location and sensorium, successfully retrieving a memory given sensory input allows us to refine our location estimate. Experimentally this improves training.

3.4 Hebbian memories

Storage. Memories of place cell representations are stored in Hebbian weights between place cells (M_t in Fig 2a), similar to [3]. We choose Hebbian learning, not only for its biological plausibility, but to also allow rapid learning when entering a new environment . We use the following learning rule to update the memory: $M_t = \lambda M_{t-1} + \eta(\mathbf{p}_t - \hat{\mathbf{p}}_t)(\mathbf{p}_t + \hat{\mathbf{p}}_t)^T$, where $\hat{\mathbf{p}}_t$ represents place cells generated from inferred grid cells. λ and η are the rate of forgetting and remembering respectively. Connections from high to low frequencies are set to zero, so that memories are retrieved hierarchically. We note than many other types of Hebbian rules work. In SM we describe changes to the learning rule if the additional distribution $q_\phi(\mathbf{g}_t \mid \mathbf{x}_{\leq t}, M_{t-1})$ is included for inference of grid cells.

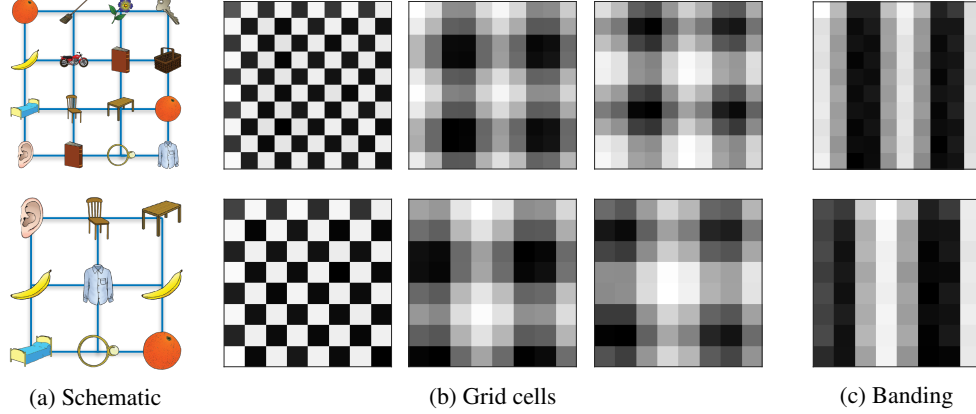| (a) Schematic | (b) Grid cells | (c) Banding |

Figure 4: Top panel all one environment, bottom panels another environment. (a) Schematic of two environments (not actual size). (b) Same three grid cells from each environment. High frequency grid on left, lower frequency on the right. Same grid code used in environments of different sizes implies a general way of representing space, i.e. not just a template of each environment. These are square grids as we have chosen a four way embedding of actions and four connected space. (c) Banded cell.

Retrieval. To retrieve memories, similarly to [3], we use an attractor network of the form $\mathbf{h}_\tau = f_p(\alpha \mathbf{h}_{\tau-1} + \mathrm{M}_t \mathbf{h}_{\tau-1})$, where τ is the iteration of the attractor network and α is a decay term. The input to the attractor, $\mathbf{h}_0$, is from the grid cells or sensorium (with their dimensions scaled appropriately) depending on whether memories are being retrieved for generative or inference purposes respectively. The output of the attractor is the retrieved memory (place cell code).

3.5 Model implications

We offer a solution to the problem of how structural codes are shared, via grid cells, to remapped place cells. Even with identical structure, since sensory stimuli across environments are different, the conjunctive code is different. Thus we believe that place cell remapping is not random, instead place cells are chosen that are consistent with both grid and sensory codes. This is a different notion to other remapping models [1], where random grid modular realignment produces a new set of place cells, and learning, that anchors these new representations, starts afresh in each environment. Our method allows for dramatically faster learning, as learnt structure can be re-used in new environments. In section 5, we present experimental evidence in concordance with our model.

In addition to offering a novel theory for place cell remapping, our model also provides an explanation for what determines place field sizes. Specifically, a given place cell will be active in the regions of space that are consistent with *both* with the grid representation (structure) received by that place cell and the sensory experience coded for by that place cell. It further offers explanation for why a given place cell may have multiple place fields within one environment, as there may be multiple locations where this consistency holds. Therefore our model offers a novel framework for designing experiments to manipulate place field sizes and locations, for example, based on simultaneously recorded grid cells and environmental cues.

We believe that using more biologically realistic computational mechanisms (e.g. Hebbian Memory, no LSTM) will facilitate further incorporation of neuroscience-inspired phenomena, such as successor representations or replay, which may be useful for building AI systems.

4 Model experiments

We show that by predicting sensory observations in environments that share structure, the model learns to generalise structural knowledge. This knowledge is represented similarly to spatial cells observed in the brain, suggesting these cells play a key role in generalisation. We further show our model exhibits fast (one-shot) learning and performs inference of unseen relationships. Although we presented a probabilistic formulation, best results were obtained when only considering the means of each distribution. Further implementation details in SM. We have taken a didactic approach to our model, thus we do not expect stellar model performance, nevertheless the model performs well.

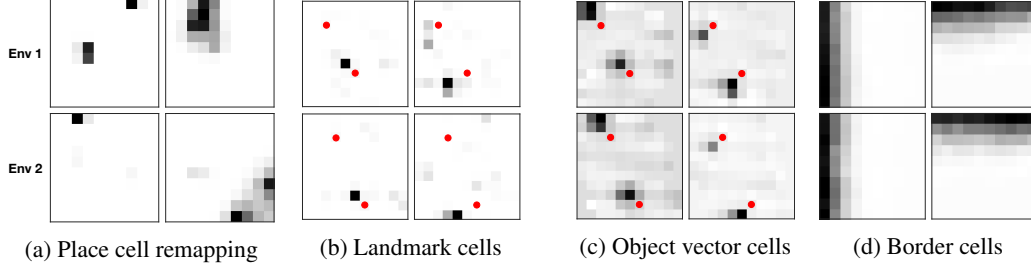(a) Place cell remapping   (b) Landmark cells   (c) Object vector cells   (d) Border cells

Figure 5: Hippocampal cells (a, b) depend on sensory experience, whereas entorhinal cells (c, d) generalise over sensory experience. Example cells from two different environments (top/bottom). a) Place cells demonstrating remapping, as also observed in the brain. These are typical in the model. Left/right: High/low frequency cell. b) Hippocampal landmark cells fire at a specific distance and direction from objects, not generalising over object identities. c) Object vector cells, however, generalise both within and across environments. d) Border cells.

Learned spatial representations. We show the representations learned by our network in Fig 4 and 5 by plotting spatial activity maps of particular neurons. In Fig 4b we present grid cells. The top panel shows cells from one environment, and the bottom panels from a different and slightly smaller environment. We see that our network chooses to represent information in a grid-like pattern (square-grids as the statistics of our space is square). We can also observe spatial firing fields at different frequencies. Representations are consistent across environments, regardless of their size - thus we have a generalisable representation of 2D space, not just a template of a particular sized environment. Fig 4c shows banded cells from our model which are also observed in the brain alongside grid cells [24]. Further learned representations are shown in SM.

We observe the appearance of phases in the grid cells (middle and right panels of Fig 4b), i.e. we find grid representations that are shifted versions of each other, as in the brain [16]. The separation into different phases means that two conjunctive place cells that respond to the same stimulus, will not necessarily be active simultaneously - each cell will only be active when their corresponding grid phase is active. Thus one can uniquely code for the same stimulus in many different locations. Across two environments, a given stimulus may occur at the same grid phase but at a different location. Thus, due to their conjunctive nature, place cells may remap across environments, as in the brain. We show this in Fig 5a. Further learned place representations are shown in SM.

Transition statistics determine basis functions. By changing transition statistics, other entorhinal cell types are observed in our model. Encouraging the agent to spend more time near boundaries leads to the emergence of border cells [31] (Fig 5d). Biasing towards particular sensory experiences leads to the discovery of object vector cells [19] (Fig 5c). Similarly to experimental evidence, although these object vector cells in our 'grid' cell layer generalise over objects, the equivalent landmark cells [11] in our 'place' cell layer do not - they are object specific (Fig 5b). Our results suggest that the 'zoo' of different cell types found in entorhinal cortex may be viewed under a unified framework - summarising the common statistics of tasks into basis functions that can be flexibly combined depending on the particular structural constraints of the environment the animal/agent faces. After an initial guess of task structure, appropriate weighting of the bases can be inferred on-line (e.g. by sensory cues / performance) to parsimoniously describe the current task structure.

One-shot learning. We test whether the network can remember what it has just seen. We consider occasions when the agent stays still at a node for the first time, as a function of the number of times that node has previously been visited (Fig 6a). We see that the agent is able to predict at a high accuracy even if it has only just visited the node for the first time. This indicates we are able to do one-shot-learning with Hebbian memory, demonstrating our model can learn episodic memories.

Zero-shot inference. Having learned the structure of our space, we should be able to correctly predict previously visited nodes even if we approach from a non-traversed edge - i.e. infer a link in the graph on loop closure. We present such data in Fig 6b. We plot the prediction accuracy of such link inferences as a function of the fraction of the total nodes visited in the graph. We achieve considerably better than chance ($1/n_s = 0.02$) prediction, which remains stable throughout graph traversal. This shows that structural information is used for inferring unseen relationships.

Long term memories. Despite using BPTT truncated at 25 steps, we retain memories for much longer (Fig 6c), indicating our grid code allows efficient storage and retrieval of episodic memories.
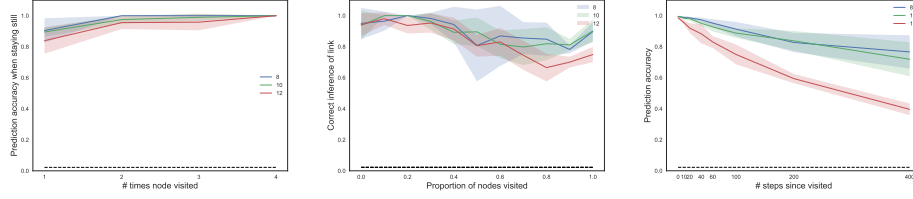
(a) One-shot learning   (b) Zero-shot link inference   (c) Long term memories

Figure 6: Prediction accuracy for different box widths. (a), (b) are previously unseen links only. Black dashed line is chance. (a) Attractor network is immediately able to retrieve Hebbian memories. (b) Unobserved graph links are inferred, implying the network has successfully learned and generalised structural knowledge. (c) Memories are successfully retrieved a long time after initial storage.

To reiterate, no representations are hard-coded. Place-like representations are learned in the attractor. Grid-like (and other entorhinal) representations are learned in the generative temporal model. These emerge from end-to-end training. These grid-like representations allow zero shot inference in new worlds demonstrating structural generalisation.

5   Analysis of data from a remapping experiment

Our framework predicts place cells and grid cells retain their relationship across environments, allowing generalisation of structure encoded by grid cells. We empirically test this prediction in data from a remapping experiment [5] where both place and grid cells were recorded from rats in two different environments. The environments were of the same dimensions (1m by 1m) but differed in their sensory (texture/visual/olfactory) cues so the animals could distinguish between them. Each of seven rats has recordings from both environments. Recordings on each day consist of five twenty-minute trials in the environments: the first and last trials in one environment and the intervening three trials in a second environment.

5.1   Methods

We test the prediction that a given place cell retains its relationship with a given grid cell across environments using two measures. First, whether grid cell activity at the position of peak place cell activity is correlated across environments (gridAtPlace), and second, whether the minimum distance between the peak place cell activity and a peak of grid cell activity is correlated across environments (minDist; normalised to corresponding grid scale). To account for potential confounds or biases (e.g. border effects, inaccurate peaks), we fit the recorded grid cell rate maps to an idealised grid cell equation [33], and use this ideal grid rate map to give grid cell firing rates and locations of grid peaks. Only grid cells with high grid scores (> 0.8) were used to ensure good ideal grid fits to the data, and we excluded grid cells with large scales (> 50cm), both computed as in [5]. Locations of place cell peaks were simply defined as the location of maximum activity in a given cell's rate map. To account for border effects, we removed place cells that had peaks close to borders (< 10cm from a border).

Our framework predicts a preserved relationship between place and grid cells of the same spatial scale (module). However, since we do not know the modules of the recorded cells, we can only expect a non-random relationship across the entire population. For each measure, we compute its value for every place cell-grid cell pair (from two trials). A correlation across trials is then performed on these values. To test the significance of this correlation and ensure it is not driven by bias in the data, we generate a null distribution by randomly shifting the place cell rate maps and recomputing the measures and their correlation across trials. We then examine where the correlation of the non-shuffled data lies relative to the null.

5.2   Results

We present analyses for both the gridAtPlace measure (Fig 7a) and the minDist measure (Fig 7b). The scatter plots show the correlation of a given measure across trials, where each point is a place cell-grid cell pair. The histogram plots show where this correlation (green line) lies relative to the null distribution of correlation coefficients. The p value is the proportion of the null distribution that is greater than the unshuffled correlation.

8

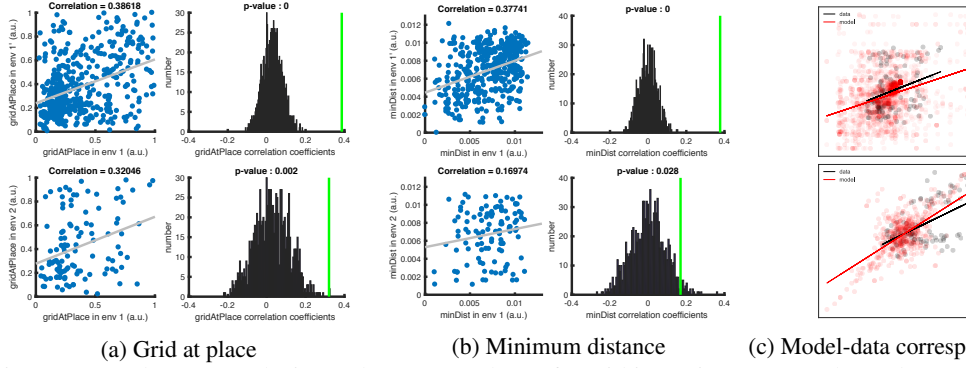(a) Grid at place   (b) Minimum distance   (c) Model-data correspondence

Figure 7: (a), (b) Data analysis results: top panels are for within environment analyses, bottom panels across environment analyses. (c) Black/ Red: Data/ Model. Top: gridAtPlace across environments. Bottom: Scatter of elements of correlation matrices across environments.

As a sanity check, we first confirm these measures are significantly correlated *within* environments (i.e. across two visits to the same environment - trials 1 and 5), when the cell populations should not have remapped (see Fig 7a, top and 7b, top). We then test *across* environments (i.e. two different environments - trials 1 and 4), to asses whether our predicted non-random remapping relationship between grid and place cells exists (Fig 7a, bottom and 7b, bottom). Here we also find significant correlations for both measures for the 115 place cell-grid cell pairs. We note the gridAtPlace result holds across environments ($p < 0.005$) when not fitting an ideal grid and using a wide range of grid score cut-offs (minDist not calculated without the ideal grid due to inaccurate grid peaks). Finally performing the across environment gridAtPlace analysis with our *model* rate maps (Fig 7c top), we observe correlations of 0.3-0.35, which are consistent with that of the data.

To share structure, the relationship between grid cells should be preserved across environments. The grid cell correlation matrix is preserved (i.e. itself correlated) across environments ($p < 0.001$ from null), both in the data [5] as well as in our model (Fig 7c bottom). These results are consistent with the notion that grid cells encode generalisable structural knowledge.

These are the first analyses demonstrating non-random place cell remapping based on neural activity, and provide evidence for a key prediction of our model: that place cells, despite remapping across environments, retain their relationship with grid cells.

6 Conclusions

We proposed a mechanism for generalisation of structure inspired by the hippocampal-entorhinal system. We proposed that one can generalise state-space statistics via explicit separation of structure and stimuli, while using a conjunctive representation with fast memory to link the two. We proposed that spatial hierarchies are utilised to allow for an efficient combinatorial code. We have shown that hierarchical grid-like and place-like representations emerge naturally from our model in a purely unsupervised learning setting. We have shown that these representations are effective at both generalising the state-space (zero-shot link inference), but also for hierarchical memory addressing. We have proposed that entorhinal cortex provides a basis set for describing the current transition structure, unifying many entorhinal cell types. We have suggested that spatial coding is just one instance of a broader framework organising knowledge. Our framework incorporates numerous phenomena or functions ascribed to the hippocampal formation (spatial cognition and representations, conceptual knowledge representation, hierarchical representations, episodic memory, inference, and generalisation). We have also presented experimental evidence that demonstrates grid and place cells retain their relationships across environment, which supports our model assumptions. We hope that this work can provide new insights that will allow for advances in AI, as well as providing new predictions, constraints and understanding in neuroscience.

7 Author Contributions

JCRW developed the model, performed simulations and drafted paper. CB collected data. JCRW, THM analysed data. JCRW, THM, SM, TEJB conceived project and edited paper. TEJB supervised.

8   Acknowledgements

References

[1] Larry F. Abbott, J. D. Monaco, and Larry F. Abbott. Modular Realignment of Entorhinal Grid Cell Activity as a Basis for Hippocampal Remapping. *Journal of Neuroscience*, 31(25):9414–9425, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1433-11.2011. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1433-11.2011.

[2] Dmitriy Aronov, Rhino Nevers, and David W. Tank. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722, 2017. ISSN 0028-0836. doi: 10.1038/nature21692. URL http://www.nature.com/doifinder/10.1038/nature21692.

[3] Jimmy Lei Ba, Geoffrey Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using Fast Weights to Attend to the Recent Past. *Advances in Neural Information Processing Systems*, pages 1–10, 2016. URL http://arxiv.org/abs/1610.06258.

[4] 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, may 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0102-6. URL http://dx.doi.org/10.1038/s41586-018-0102-6http://www.nature.com/articles/s41586-018-0102-6.

[5] C. Barry, L. L. Ginzberg, J. O'Keefe, and N. Burgess. Grid cell firing patterns signal environmental novelty by expansion. *Proceedings of the National Academy of Sciences*, 109(43):17687–17692, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1209918109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1209918109.

[6] E Bostock, R U Muller, and J L Kubie. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, 1(2):193–205, 1991. ISSN 1050-9631. doi: 10.1002/hipo.450010207.

[7] Cindy A. Buckmaster, Howard Eichenbaum, David G Amaral, Wendy A Suzuki, and Peter R Rapp. Entorhinal Cortex Lesions Disrupt the Relational Organization of Memory in Monkeys. *Journal of Neuroscience*, 24(44):9811–9825, 2004. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1532-04.2004. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1532-04.2004.

[8] Daniel Bush, Caswell Barry, Daniel Manson, and Neil Burgess. Using Grid Cells for Navigation. *Neuron*, 87(3):507–520, 2015. ISSN 10974199. doi: 10.1016/j.neuron.2015.07.006. URL http://dx.doi.org/10.1016/j.neuron.2015.07.006.

[9] Alexandra O. Constantinescu, Jill X. O'Reilly, and Timothy E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016. ISSN 10959203. doi: 10.1126/science.aaf0941.

[10] Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. pages 1–19, 2018. URL http://arxiv.org/abs/1803.07770.

[11] Sachin S Deshmukh and James J Knierim. Influence of local objects on hippocampal representations: Landmark vectors and memory. *Hippocampus*, 23(4):253–67, apr 2013. ISSN 1098-1063. doi: 10.1002/hipo.22101. URL http://www.ncbi.nlm.nih.gov/pubmed/23447419http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3869706.

[12] Yedidyah Dordek, Daniel Soudry, Ron Meir, and Dori Derdikman. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife*, 5(MARCH2016):1–36, 2016. ISSN 2050084X. doi: 10.7554/eLife.10094.

[13] J. A. Dusek and H. Eichenbaum. The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences*, 94(13):7109–7114, 1997. ISSN 0027-8424. doi: 10.1073/pnas.94.13.7109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.94.13.7109.

[14] Marianne Fyhn, Torkel Hafting, Alessandro Treves, May Britt Moser, and Edvard I. Moser. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, 446(7132):190–194, 2007. ISSN 14764687. doi: 10.1038/nature05601.

[15] Mevlana Gemici, Chia-Chun Hung, Adam Santoro, Greg Wayne, Shakir Mohamed, Danilo J. Rezende, David Amos, and Timothy Lillicrap. Generative Temporal Models with Memory. pages 1–25, 2017. ISSN 1702.04649. URL http://arxiv.org/abs/1702.04649.

[16] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-britt Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005. ISSN 00280836. doi: 10.1038/nature03721.

[17] Demis Hassabis, Dharshan Kumaran, S. D. Vann, and E. A. Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0610561104. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0610561104.

[18] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258, 2017. ISSN 10974199. doi: 10.1016/j.neuron.2017.06.011. URL http://dx.doi.org/10.1016/j.neuron.2017.06.011.

[19] Øyvind Arne Høydal, Emilie Ranheim Skytøen, M B Moser, and Edvard I Moser. Object-vector cells in the medial entorhinal cortex. *bioRxiv*, 2018. doi: 10.1101/286286.

[20] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. (Ml):1–14, 2013. ISSN 1312.6114v10. doi: 10.1051/0004-6361/201527329. URL http://arxiv.org/abs/1312.6114.

[21] Kirsten Brun Kjelstrup, Trygve Solstad, Vegard Heimly Brun, Torkel Hafting, Stefan Leutgeb, Menno P. Witter, Edvard I. Moser, and May-Britt Moser. Finite Scale of Spatial Represenation in the Hippocampus. *Science*, 321(July):140 – 143, 2008. ISSN 0036-8075.

[22] Robert W. Komorowski, Joseph R. Manns, and Howard Eichenbaum. Robust Conjunctive Item-Place Coding by Hippocampal Neurons Parallels Learning What Happens Where. *Journal of Neuroscience*, 29(31):9918–9929, 2009. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1378-09.2009. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1378-09.2009.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012. ISSN 10495258. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[24] Julia Krupic, Neil Burgess, and John O'Keefe. Neural Representations of Location Composed of Spatially Periodic Bands. 337(August):853–857, 2012. URL http://www.sciencemag.org/content/337/6096/853.full.pdf.

[25] Dharshan Kumaran, Jennifer J. Summerfield, Demis Hassabis, and Eleanor A. Maguire. Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, 63(6):889–901, 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.07.030. URL http://dx.doi.org/10.1016/j.neuron.2009.07.030.

[26] Stefan Leutgeb, Jill K. Leutgeb, Carol A. Barnes, Edvard I. Moser, Bruce L. McNaughton, and May-Britt Moser. Independent Codes for Spatial and Episodic Memory in Hippocampal Neuronal Ensembles. (July):619–624, jul 2005.

[27] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. ISSN 0033295X. doi: 10.1037/0033-295X.102.3.419.

[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Bellemare Marc G, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. ISSN 10450823. doi: 10.1038/nature14236. URL http://dx.doi.org/10.1038/nature14236.

[29] John O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, nov 1971. ISSN 00068993. doi: 10.1016/0006-8993(71)90358-1. URL http://linkinghub.elsevier.com/retrieve/pii/0006899371903581.

[30] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. 2014. ISSN 10495258. doi: 10.1051/0004-6361/201527329. URL http://arxiv.org/abs/1401.4082.

[31] Trygve Solstad, Charlotte N Boccara, Emilio Kropff, May-Britt Moser, and Edvard I Moser. Representation of geometric borders in the entorhinal cortex. *Science (New York, N.Y.)*, 322(5909):1865–8, dec 2008. ISSN 1095-9203. doi: 10.1126/science.1166466. URL papers3://publication/doi/10.1126/science.1166466http://www.ncbi.nlm.nih.gov/pubmed/19095945.

[32] Kimberley L. Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017. ISSN 15461726. doi: 10.1038/nn.4650.

[33] Martin Stemmler, Alexander Mathis, and Andreas Herz. Connecting Multiple Spatial Scales to Decode the Population Activity of Grid Cells. *Science Advances*, in press(December):1–12, 2015. ISSN 2375-2548. doi: 10.1126/science.1500816.

[34] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian FrØland, May-Britt Britt Moser, and Edvard I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, 2012. ISSN 00280836. doi: 10.1038/nature11649. URL http://www.nature.com/doifinder/10.1038/nature11649http://dx.doi.org/10.1038/nature11649.

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

[36] Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. ISSN 1939-1471. doi: 10.1037/h0061626. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0061626.

[37] Albert Tsao, Jørgen Sugar, Li Lu, Cheng Wang, James J. Knierim, May-Britt Moser, and Edvard I. Moser. Integrating time from experience in the lateral entorhinal cortex. *Nature*, 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0459-6. URL http://www.nature.com/articles/s41586-018-0459-6.

[38] James C. R. Whittington and Rafal Bogacz. An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Computation*, 29(5):1229–1262, may 2017. ISSN 0899-7667. doi: 10.1162/NECO_a_00949. URL http://www.mitpressjournals.org/doi/10.1162/NECO{_}a{_}00949https://www.biorxiv.org/content/early/2016/12/23/035451https://www.ncbi.nlm.nih.gov/pubmed/28333583{%}0Ahttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467749/pdf/emss-73.

[39] Emma R. Wood, Paul A. Dudchenko, and Howard Eichenbaum. The global record of memory in hippocampal neuronal activity. *Nature*, 397(6720):613–616, 1999. ISSN 00280836. doi: 10.1038/17605.