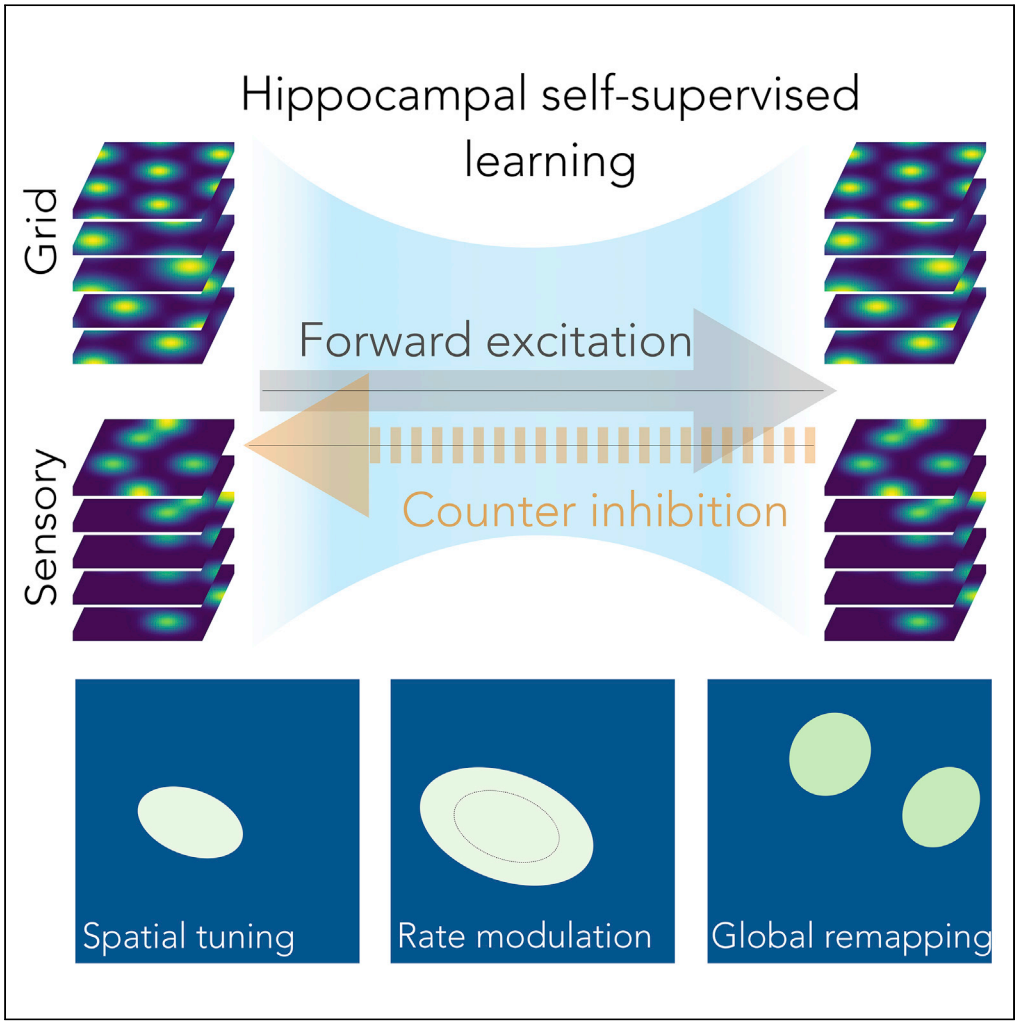
iScience

CellPress
OPEN ACCESS

Article

Entorhinal mismatch: A model of self-supervised learning in the hippocampus



Diogo Santos-Pata, Adrián F. Amil, Ivan Georgiev Raikov, César Rennó-Costa, Anna Mura, Ivan Soltesz, Paul F.M.J. Verschure

pverschure@ibecbarcelona.eu

Highlights

Is backpropagation of the error optimizing biological learning as in neural networks

The hippocampus seems to support gradient descent via countercurrent inhibition

The entorhinal-hippocampal complex error is suggested to drive self-supervised learning

The mismatch learning rule reproduces many of the hippocampal physiological phenomena

Santos-Pata et al., iScience *24*, 102364
April 23, 2021 © 2021
https://doi.org/10.1016/j.isci.2021.102364

iScience

CellPress
OPEN ACCESS

Article

Entorhinal mismatch: A model of self-supervised learning in the hippocampus

Diogo Santos-Pata,[1,6] Adrián F. Amil,[1,2,6] Ivan Georgiev Raikov,[3] César Rennó-Costa,[4] Anna Mura,[1] Ivan Soltesz,[3] and Paul F.M.J. Verschure[1,5,7,*]

SUMMARY

The hippocampal formation displays a wide range of physiological responses to different spatial manipulations of the environment. However, very few attempts have been made to identify core computational principles underlying those hippocampal responses. Here, we capitalize on the observation that the entorhinal-hippocampal complex (EHC) forms a closed loop and projects inhibitory signals "countercurrent" to the trisynaptic pathway to build a self-supervised model that learns to reconstruct its own inputs by error backpropagation. The EHC is then abstracted as an autoencoder, with the hidden layers acting as an information bottleneck. With the inputs mimicking the firing activity of lateral and medial entorhinal cells, our model is shown to generate place cells and to respond to environmental manipulations as observed in rodent experiments. Altogether, we propose that the hippocampus builds conjunctive compressed representations of the environment by learning to reconstruct its own entorhinal inputs via gradient descent.

INTRODUCTION

The hippocampus has been suggested to play a key role in a range of cognitive functions, including spatial navigation (Burgess, Jeffery, & O'Keefe, 1999), memory consolidation (Nadel and Moscovitch, 1997), attentional shift (Devauges and Sara, 1990), working memory maintenance (Axmacher et al., 2010; Olton et al., 1980), and others (Mack et al., 2018). Interestingly, in spite of the variety of cognitive functions at least partially attributed to the hippocampus, a unified theory of hippocampal function anchored in psychology and the cognitive sciences has not been successfully developed. Indeed, descriptions of diverse computational principles inferred mostly from hippocampal connectivity and single-cell firing patterns, such as competitive selection (De Almeida, Idiart and Lisman, 2009), attractor dynamics (Wills, Lever, Cacucci, Burgess, & O'Keefe, 2005), sequences (Pastalkova et al., 2008), short- and long-term plasticity (Alger and Teyler, 1976), and hierarchical processing (Lavenex and Amaral, 2000), have advanced explanations for behavioral and physiological experimental data from multiple paradigms including habituation to novelty (Yamaguchi, Hale, D'Esposito and Knight, 2004) and latent learning (Kimble and BreMiller, 1981). Therefore, a general description of the computational arsenal of the hippocampus should also advance our understanding of hippocampal cognitive processes.

A standout feature of the circuitry of the hippocampus and entorhinal cortex (EC) is its organization in multiple parallel loops (Bartesaghi et al., 2006). The circuitry of the EC allows the convergence of cortical input and hippocampal output pathways within the same structure, setting the conditions for the implementation of a "comparator" between the two (Lörincz and Buzsáki, 2000). Further, such a plastic circuit that minimizes its input-output discrepancy with multiple inner layers may be viewed as implementing a self-supervisory function, i.e., to act as a type of neural network that is capable of learning without a supervising reference (Wang, 2001). A neural network of such characteristics is normally referred to as an autoencoder (Hinton & Salakhutdinov, 2006), which has been previously used as a model of the hippocampus (Gluck et al., 2003; Gluck and Myers, 1993; Japkowicz et al., 1995), showing how sensory cues can be modulated to support learning in other brain regions performing tasks related to classical conditioning. Furthermore, a recent study following a similar principle has also shown that place cells emerge naturally by the information compression carried out in the bottleneck of the autoencoder (Benna and Fusi, 2019). Although it is unclear whether the hippocampal circuitry evaluates the hippocampal-cortical signal discrepancy (i.e., the comparator) and learns accordingly, it can in principle support computations that are classically

[1]Laboratory of Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS), Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain

[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[3]Department of Neurosurgery, Stanford University, Stanford, CA, USA

[4]Digital Metropolis Institute, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil

[5]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

[6]These authors contributed equally

[7]Lead contact

*Correspondence: pverschure@ibecbarcelona.eu

https://doi.org/10.1016/j.isci.2021.102364

associated with the hippocampus: the correction of noisy and incomplete input signals, known as pattern completion (Rolls, 2013), and the generation of sparse divergent representations, known as pattern separation (Yassa and Stark, 2011). Importantly, these two functions are thought to be critical for the role of the hippocampus in forming episodic memories (N. Burgess, Maguire, & O'Keefe, 2002). Therefore, we ask the question whether there is a general computational principle that can explain these phenomena as well as the corresponding physiology and behavioral data in a unified theoretical framework. We argue that the input-output mismatch minimization is the core computation carried out in the entorhinal-hippocampal complex (EHC), which optimizes information compression and transfers in a self-supervised manner.

Interestingly, although the concept of the trisynaptic loop emphasizes the role of the sequential feedforward circuitry (Brewer et al., 2013), a growing body of empirical evidence suggests that multiple mechanisms may carry information in the backward or "countercurrent" direction within the hippocampal formation. For example, the phase of theta waves in the subiculum appear to precede the CA1 and CA3 theta phases, with GABAergic mechanisms playing a key role in the temporal coupling (Jackson et al., 2014). Similarly, all major types of CA3 and CA1 GABAergic neurons have been shown to possess significant boundary-crossing axon terminals, carrying CA activity in the form of inhibition back to dentate gyrus neurons (Szabo et al., 2017). In fact, about a fifth of the GABAergic inputs to dentate granule cells have been estimated to originate from the CA3 and CA1 regions, indicating the potential power of this countercurrent GABAergic hippocampal projection system. Feedback inhibition from the CA1 area to CA3 and dentate gyrus is in general agreement with the hypothesis that a countercurrent pathway for synaptic plasticity may exist within the hippocampus (Sik et al., 1994). A further plausible component engaged in promoting synaptic changes in the hippocampal loop as a consequence of the EC comparator is the entorhinal layers IV-VI projections to granule and GABAergic neurons in the dentate gyrus (Deller et al., 1996) and the GABAergic projection from the deeper layers of the EC to the CA1 (Melzer et al., 2012). Together, these experimental observations of multiple pathways providing backward inhibitory signaling with potential roles in modulating synaptic plasticity highlight a potential solution for a biological implementation of the type of error backpropagation needed for self-supervised gradient descent learning within the hippocampus (see (Santos-Pata et al., 2021) for a more thorough review and a related discussion on its biological feasibility).

Here, we evaluate whether the self-supervised model of the hippocampus featuring the EC comparator hypothesis and input-output mismatch minimization can account for diverse physiological findings reported to occur in the hippocampus and adjacent regions, namely, spatial learning and representations (M. B. Moser, Moser et al., 1995), the response to environmental modifications (Colgin et al., 2008), novelty detection (Knight, 1996), and relearning (Clare et al., 2002). To that end, we implemented ENCORE, a self-supervised network with multiple layers associated with each of the main subregions of the EHC: EC, dentate gyrus (DG), CA3, and CA1. The network is fed with realistic spatial signals from medial EC (MEC) and lateral EC (LEC) that propagate sequentially throughout the network. The learning algorithm applied to the synaptic weights, error backpropagation, minimizes the difference between the activity of the first and last layers. We then evaluate whether the emergent spatial representations in the middle layers can capture the reported data considering particular manipulations on the spatial inputs. Therefore, and contrary to previous models (Benna and Fusi, 2019; Gluck and Myers, 1993), we provide a comprehensive set of realistic environmental manipulations allowing us to compare ENCORE responses to numerous physiological benchmarks and in a layer- or subregion-wise manner. Based on our results, we propose that the hippocampus may be able to implement a form of error backpropagation which affords an end-to-end self-supervised optimization of the whole EHC by continuously minimizing its input-output mismatch. In turn, this core principle leads to an information compression throughout the hippocampal bottleneck that is able to reproduce core physiological responses to environmental modifications seen in rodent experiments.

RESULTS

Learning, spatial representations, and responses to environmental manipulations

We have devised a self-supervised autoencoder network to explore the potential role for entorhinal cells in detecting novelty and to modulate synaptic changes throughout the hippocampus through the backpropagation of error (Figures 1A and 1B; see transparent methods). First, we tested the ability of the model to reconstruct the entorhinal inputs (Figures 1C and 1D) at the output layer. As expected from these types of self-supervised models, the model quickly (40 epochs) learned an optimal weight configuration to minimize the error between the activity of the input and output layers (Figure 1E). After learning, the entorhinal

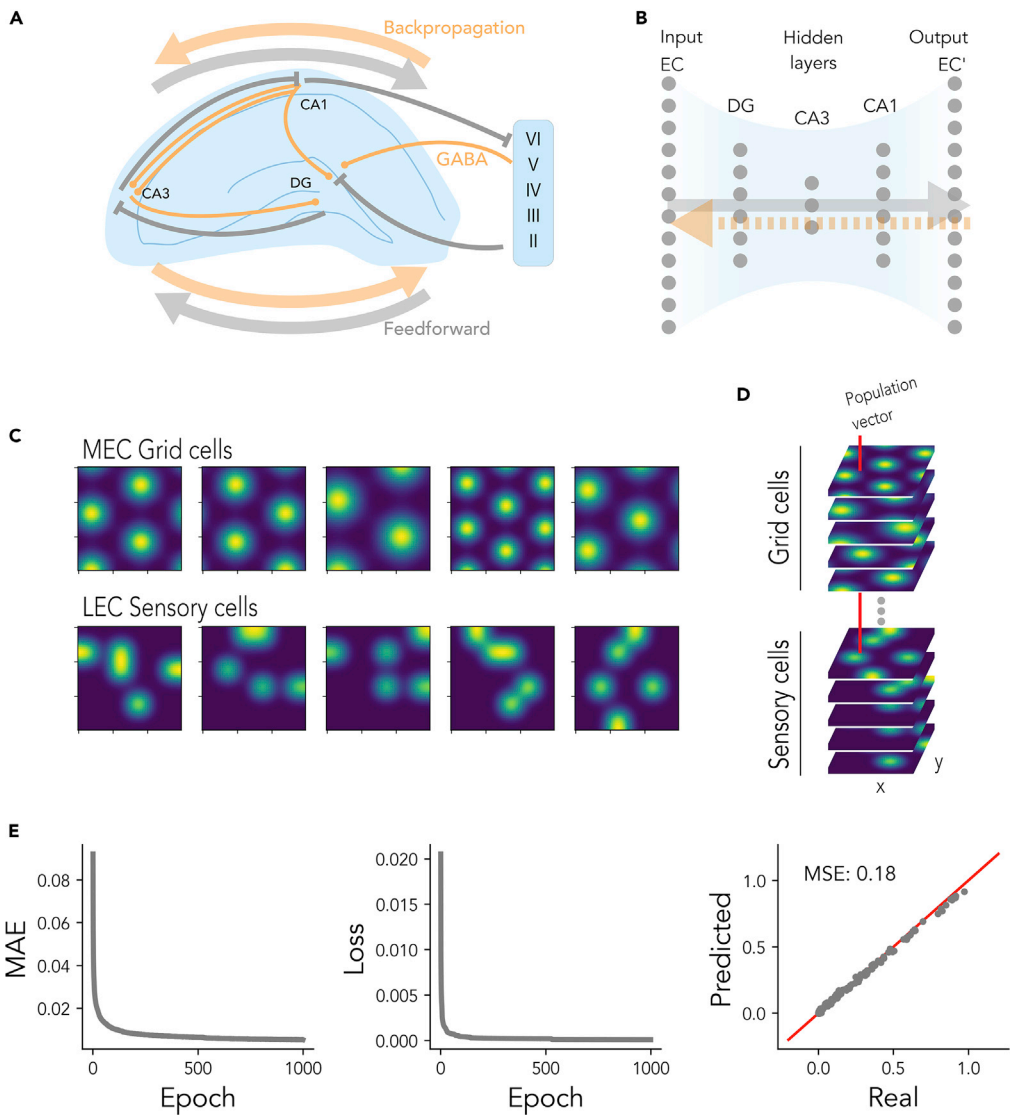


Figure 1. Hippocampal model and training paradigm

(A) Hippocampal trisynaptic circuit with feedback loop through the EC.

(B) Proposed mechanism to learn from the hippocampal loop in the form of a self-supervised model.

(C) Input rate maps mimicking the ones from grid cells (MEC) and sensory cells (LEC). See transparent methods.

(D) Diagram representation of the model input every time (t) as a function of the spatial location.

(E) (Left and center) Loss and mean absolute error (MAE) of the model during learning (epochs). (Right) The correlation between the real input (EC) to the network and the outcome (Pearson-r = 0.998).

reconstruction error at every spatial bin was asymptotically below 0.01 mean absolute error (Figure 1E, left and center), with a strong population vector correlation between input and output at every point in time (Pearson-r, $r = 0.99$, $p < 0.001$, Figure 1E, right). Having established the ability of the model to learn to reconstruct the EC input vector, we analyzed the activity of the hidden layers while exploring the environment. As expected from hippocampal place cells, units in our model tuned their maximal activity to specific locations of the squared arena (Figure 2A). Interestingly, the network's hidden layers (an analogy of the DG, CA3, and CA1 regions) showed place field density distributions broadly reflecting experimental data of the rodent hippocampus (Leutgeb et al., 2007) (Figure 2B).

Next, to quantifying the model's response to environmental modifications, we froze learning and tested the model on a second stack of LEC rate maps, representing a novel environment (Figure 2C). Thus, we were

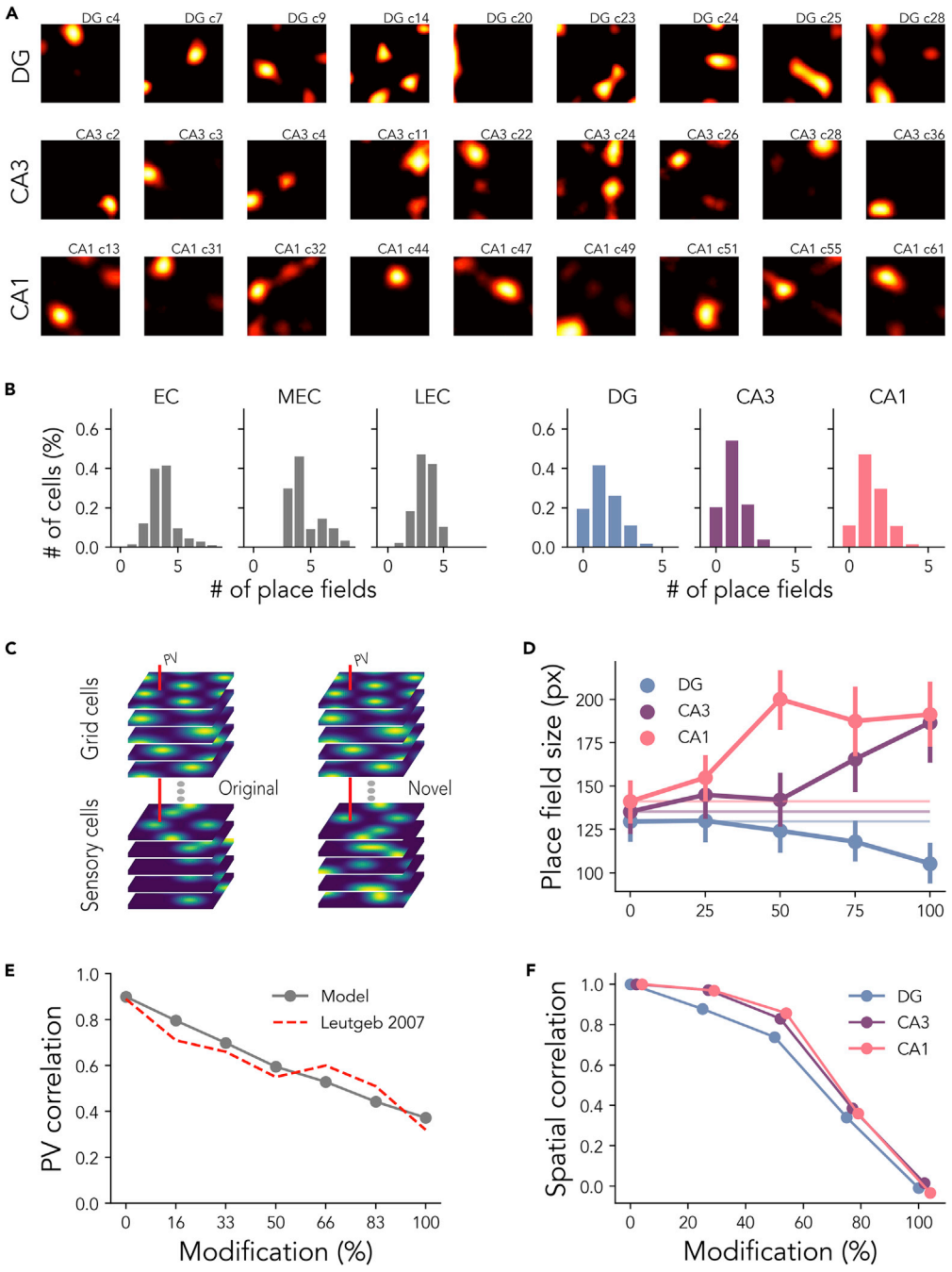


Figure 2. Place cell modulation after environmental morphing

(A) Example of hidden layers (DG, CA3, CA1) activity resembling the ones of hippocampal place cells.

(B) Distribution of number of place fields per cell in the EC input layer (LEC and MEC) and hidden layers.

(C) Procedure to study the model response to environmental modifications. (Left) The EC input used to train the model. (Right) Replacing a percentage of the sensory cells rate maps mimics environmental modifications. Environmental modifications of different degrees (levels) could go from slightly distinct (10%) to a completely novel environment (100%). The population vector (PV) at each spatial bin is fed to the model.

(D) Modifying the environment leads to an increase of place fields size (mean ± standard deviation) in CA3 and CA1 layers but not in the DG (Barry et al., 2012).

(E) The DG layer responds with rate modulation (Leutgeb et al., 2007).

(F) Spatial correlation decreases with increased environmental modifications, less drastically for CA layers, as expected from later hippocampal subregions.

able to test the model's accuracy and the effects of environmental modification at different magnitudes, from familiar (0% modification) to completely novel (100% modification). This approach mimics the modulating the incoming sensory signals due to alteration of objects displaced within the environment. The activity of hippocampal neural ensembles has been observed to be modulated as a function of the degree of environmental modifications, suggesting a rate remapping code mediating spatial maps learned in the DG (Leutgeb et al., 2007). At the computational level, rate remapping has been hypothesized to be mediated by the interaction between spatial and sensory inputs arriving at the hippocampus by entorhinal projections (Rennó-Costa et al., 2010). An interesting question, therefore, is whether input-output mismatch error learning is sufficient to reproduce rate remapping. We quantified the response of units in the model DG in relation to the rate of environmental modifications (Figure 2E). As in rodent physiology, we observed a decrease in the population activity correlation as the environment was progressively modified (Pearson test, $r = 0.99$, $p < 0.001$; one-way analysis of variance [ANOVA] $F(1,278) = 75{,}635.479$, $p = 0.0$, $\eta 2 = 0.996$), closely matching the decay function observed in hippocampal rate remapping experiments. Furthermore, we quantified the changes in spatial correlation in cells at the distinct stages of our network with respect to environmental modification. The simulations indicated a steeper decrease for early (DG) compared to later (CA) stages (Figure 2F), a result that directly speaks to the observations that CA3 populations sustain spatial representations over increased levels of sensory modifications (Leutgeb et al., 2007) due to their attractor dynamics (Reno Costa CA3 paper). Moreover, the spatial correlation dropped significantly for the 3 stages after 50% modification, a signature consistent with global remapping (Sanders et al., 2020). Thus, these results suggest that the internal network configuration (i.e., intra-hippocampal connectivity) learned from the input-output mismatch error derived from the EC comparator is sufficient to explain key features of rate and global remapping.

The scale of spatially tuned place fields of hippocampal CA1 place cells has been observed to temporally expand after environmental modifications (Barry, Ginzberg, O'Keefe and Burgess, 2012). To test whether our model responded similarly, we quantified the size place fields along the hippocampal hidden layers during a progressive morphing of the environment (Figure 2D). We observed that our model's CA1 place fields increased their spatial scale when the LEC sensory stream is altered (one-way ANOVA $F(1,1430) = 24.34$, $p = 0.0$, $\eta 2 = 0.017$). Interestingly, this modulation was specific to CA1 cells, with CA3 cells resisting up to 50% of environmental changes (one-way ANOVA $F(1,988) = 21.309$, $p = 0.0$, $\eta 2 = 0.021$), and with DG cells decreasing their scale at greater modifications (one-way ANOVA $F(1,767) = 10.465$, $p = 0.00127$, $\eta 2 = 0.013$). Moreover, a similar observation concerning the changes in spatial scale has also been observed in MEC grid cells during rodent exploration upon environmental modifications (Barry et al., 2012). Because our model is based on self-supervision, we measured the differences in spatial scaling in the predicted MEC grid cells' fields (the output layer of the model). Surprisingly, and as observed in physiological data, we also identified an increase in grid cells firing fields after environmental sensory modifications (Student's t-test, statistic = -8.881363, $p < 0.001$, Figure 3). Moreover, we believe that these observations based on spatial alterations are in agreement with the grid-place cell interdependence hypothesis which proposes that grid cells rely on excitation activity from hippocampal neurons (Bonnevie et al., 2013).

The aforementioned modulation of place fields upon environmental modification has been previously demonstrated during environmental morphing (O'Keefe and Burgess, 1996). Specifically, it has been observed that environmental boundary elongation or stretching after navigational exploration is sufficient to control the position and geometry of hippocampal place fields, suggesting that boundary signals, likely originating in MEC layer III (Solstad et al., 2008), are involved in modulating the hippocampal spatial firing activity. Moreover, those observations are also in line with the hypothesis that the hippocampus is involved in both allocentric and egocentric spatial representations (Feigenbaum and Rolls, 1991). Because our model simply received local inputs, i.e., no distal cue information, we next aimed to quantify whether the internally generated network dynamics were sufficient to infer short-distance egocentric representations while still modulating the position and geometry of place fields as observed in the rodent hippocampus. To do so, we generated a horizontally elongated arena mimicking the input signals processed during environmental stretching, where the spatial scale of grid cells remained intact, but sensory signals were extended proportional to environmental morphing (Figure 4A). We exposed the model to the modified arena but without allowing further synaptic updates, thus avoiding learning of the new configuration. Visual inspection of the model's reconstructed place fields revealed that the firing activity of its place cells accompanied the direction and extent of the environmental elongation (Figure 4B). The size (area) of place fields increased after environmental stretching, affecting all three hippocampal populations (Figure 4D, DG:

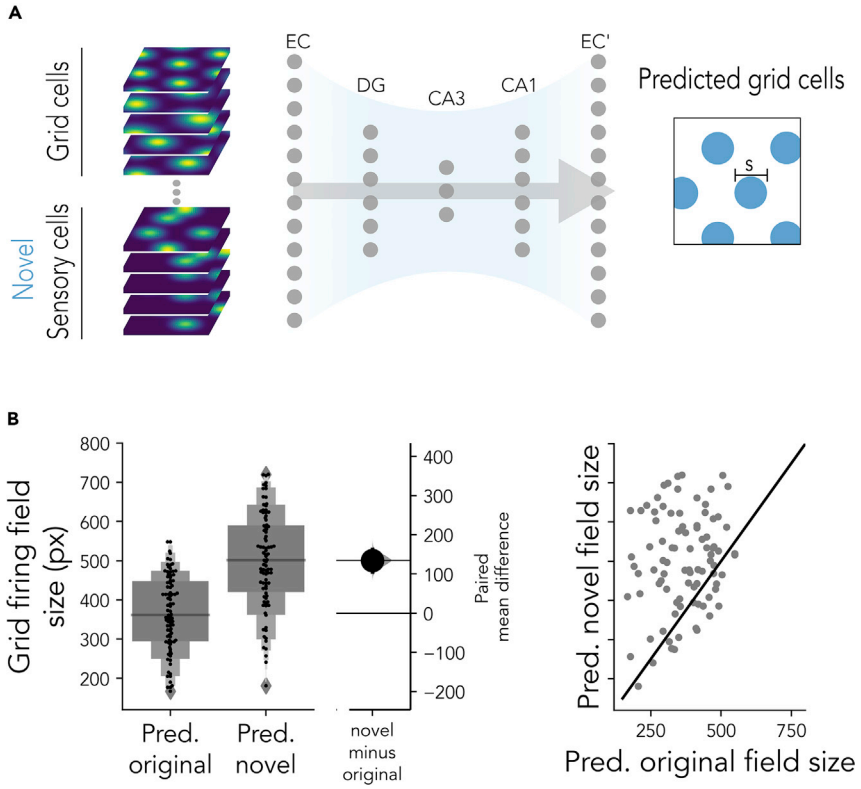


Figure 3. Grid cells firing field expansion after environmental modification

(A) Hypothesized modulation of grid cell (MEC) reconstruction activity in EC′ after altering 10% LEC input (sensory cells) (see transparent methods).

(B) Grid cells' firing field size in response to SPECIFIY MANIPULATION. Px = pixels.

original 127 x 84 pixels, stretched 190 x 107 pixels, t test, statistic = −6.9193, $p < 0.001$; CA3: original 139 x 79 pixels, stretched 200 x 138 pixels, t test, statistic = −5.4390, $p < 0.001$; CA1: original 137 x 85 pixels, stretched 214 x 160 pixels, t test, statistic = −6.8463, $p < 0.001$). Moreover, the number of place cells within each stage also increased after environmental stretching, an effect strongly affecting CA1 cells (Figure 4C, t test, DG: t- statistic = − 3.2163, $p < 0.001$; CA3: t-statistic = −4.4941, $p < 0.001$; CA1: t-statistic = −3.7907, $p < 0.001$).

Novelty detection, generalization, and relearning

Once the hippocampal loop has been optimized by minimizing its input-output mismatch in a certain environment, changes in the environmental configuration should lead to changes in activity, serving novelty detection. Therefore, we next studied how activity changed across the stages of the model dependent on the amount of environmental modification (see transparent methods section for a detailed description of how the environment was modified). We used the mean squared error (MSE) of the activity vectors per layer between the original and the novel environment as an approximation of the local error signal since it captures the difference in population activity between conditions.

By continuously varying the environment, we observed a strong relationship between the extent of the environmental modification and the MSE in the CA1 output stage (Figures 5A 5C). This indicates that the error signal produced in the EC by the hippocampal output may not only be detecting changes in the environment but also their magnitude. Furthermore, although this linear dependence between environmental change and error is also maintained across layers, we observed that the magnitude of the error was differentially distributed, with the DG having the highest slope (Figure 5D). This is in agreement with previous physiological reports demonstrating a larger change in neural activity in the DG as compared to CA3 and CA1 due to novel environments (Hunsaker et al., 2008; Lee et al., 2005). Hence, our results emphasize the role of the DG as the layer within the hippocampal loop in both signaling novelty and undergoing

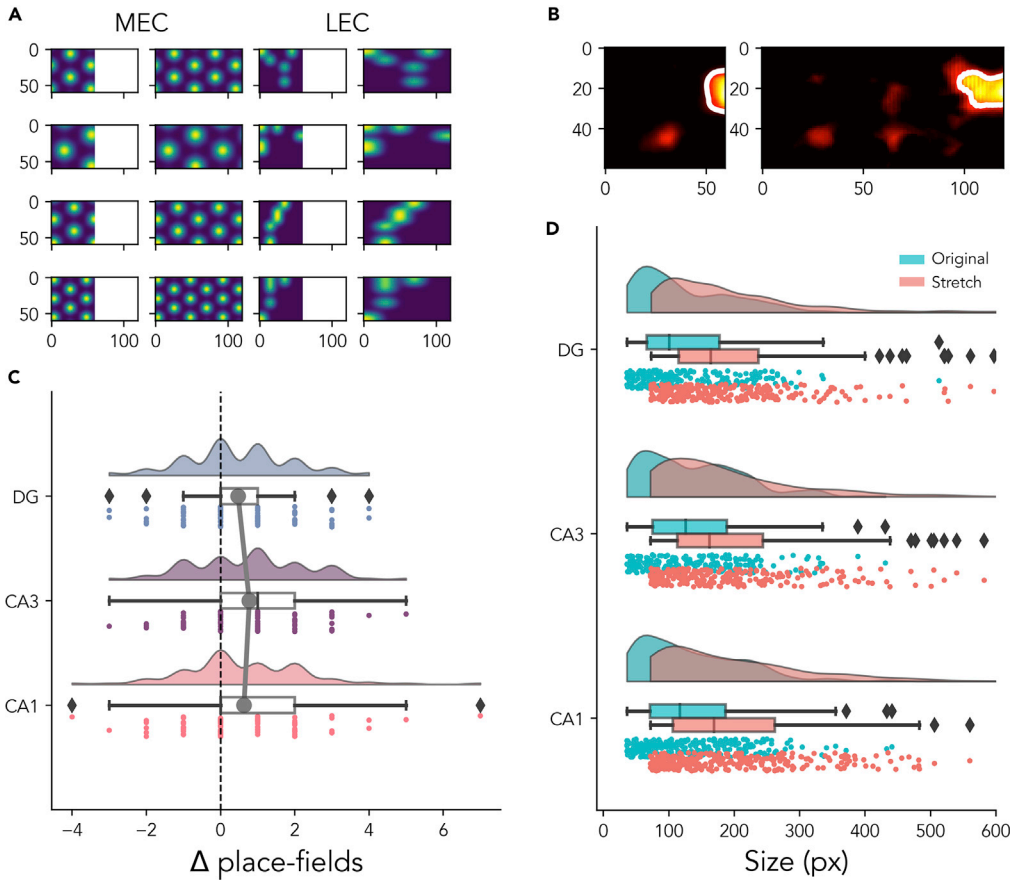


Figure 4. Place field elongation after environmental stretching

(A) Procedure to test the model during environmental stretching as in (O' Keefe and Burgess, 1996).

(B) Example cell with a stretched place field after environmental manipulation.

(C) Difference in the number of place fields between original and stretched environment, suggesting that more place fields emerge.

(D) As in (C), place fields tend to increase their size after environmental stretch. Px = pixels.

error driven plastic changes (Davis et al., 2004; E. Moser, Moser and Andersen, 1993). Moreover, by changing the environment in a space-specific manner and analyzing the error signal during spatial navigation, we also found that increases in the error signal of the network coincided well with the position of the modified place fields (Figure 5E). The latter result shows that the error signal generated at the EC that drives learning, can encode novelty during navigation and, therefore, be used to trigger learning at specific locations in the environment. Thus, we also studied the performance of the hippocampal model during the relearning of a novel, modified environment (Figures 5F and 5G). We observed that, even when the environment is completely novel, the network maintains a relatively small initial error, demonstrating the generalization capabilities of its previously learned place fields and their distribution (Figure 5G). Moreover, it can be seen that error decreases very rapidly after the initial novelty detection, showing fast relearning under novelty. However, the number of trials (i.e., epochs) to reach the error asymptote of the naive condition (i.e., first learning) varies with the extent of environmental modification. Interestingly, relearning is faster for both small and large changes as compared to intermediate changes (Figure 5F). That is, highly novel environments drive the place fields to readapt faster than similar environments with medium levels of modification. The latter suggests a nonlinear relationship between the extent of environmental modification and relearning speed that could be tested experimentally.

DISCUSSION

Over the last four decades, a variety of cell types have been found to encode egocentric and allocentric spatial features and these have highlighted the role of the hippocampus in mapping, localization and

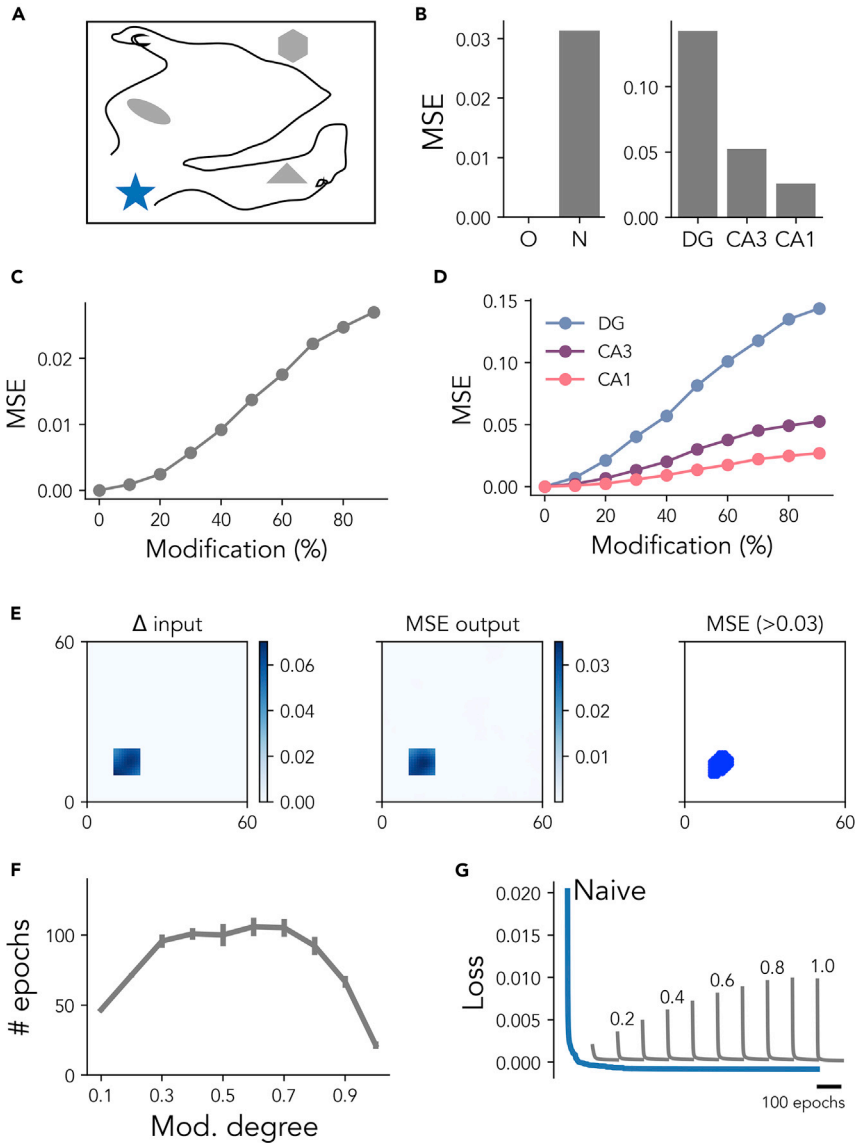


Figure 5. Novelty detection and relearning

(A) Novelty can be simulated by manipulating the rate maps of the sensory cells (LEC) at specific locations.

(B) MSE for the model output (left) and for individual stages (right) showing that error magnitude decreases along the DG to CA1 layers.

(C and D) The error increases as the environmental modification increases and is stage specific.

(E) Example of novelty detection during navigation showing that the model can detect environmental modifications by monitoring its reconstruction error. By altering the activity of the sensory cells within a portion of the environment (hotspot in the left plot), the model increases its error at that same location (center plot). Thresholding the model's output error allows us to detect modified locations, threshold set to > 0.03.

(F) Number of epochs require to XYZ versus the degree of environmental morphing. The number of epochs needed for learning (stabilization) increases with the environmental modification level. Notably, the model converges quicker for largely different, novel environments.

(G) The number of epochs needed for learning (stabilization) increases with the environmental modification level. Notably, the model converges quicker for largely different, novel environments.

planning (Howard and Eichenbaum, 2015). Indeed, the hippocampal trisynaptic circuit and its physiological responses have been extensively studied with special emphasis on spatial navigation (Ekstrom et al., 2003). In this regard, the associative binding of both sensory- and self-motion- related signals (coming from LEC and MEC, respectively) within the hippocampal network serves the animal's ability to learn the statistical

regularities of the environment and to recall previously experienced episodes. Thus, we decided to explore whether the wide variety of reported physiological responses to environmental modifications during spatial navigation in rodents could be accounted for by a small set of core computational principles operating in the hippocampus: self-supervision via EC input reconstruction (comparator hypothesis), information compression along the trisynaptic pathway, and gradient descent learning via countercurrent inhibition. For that purpose, we used an neural network optimized by error backpropagation exposed to realistic stimuli coming from both LEC (sensory) and MEC (space). We observed that our model was capable of generating place cell-like receptive fields, modulating individual neuron's and population activity to environmental modifications , performing novelty detection, and generalizing to novel environments reflecting physiological data.

The implementation of hippocampus-like computations in the form of an autoencoder dates back almost three decades (Gluck and Myers, 1993), for instance, suggested that the hippocampus develops new stimulus representations that enhance the discriminability of predictive cues. Importantly, and unlike these previous studies, we do not make an ontological commitment to the autoencoder as a comprehensive and faithful model of the hippocampus. Instead, we argue that a optimized by error backpropagation captures a unifying computational principle operating in the hippocampus: the gradual compression of external information through gradient descent over the self-generated reconstruction error. We call this architecture ENCORE (Entorhinal Compression Reconstruction). The rationale behind considering ENCORE model as an informative model of the EHC is threefold. Firstly, we select the network topology by grounding our understanding of hippocampal information processing on the aforementioned comparator hypothesis (Lörincz and Buzsáki, 2000), whereby the EC generates an error signal from comparing the neocortical inputs with their respective hippocampal reconstructions. Secondly, we propose that the mismatch error signal generated in the EC comparator is fed back to and optimizes the hippocampal trisynaptic pathway by gradient descent and, more specifically, by error backpropagation. We selected the learning rule of error backpropagation after reviewing recent anatomical and physiological literature pointing to a broad inhibitory (GABAergic) network that runs countercurrent to the mainly excitatory trisynaptic pathway (Santos-Pata et al., 2021). This countercurrent inhibitory network consists of boundary-crossing interneurons that could plausibly modulate synaptic plasticity in pyramidal cells in a way that is consistent with recent proposals about how error backpropagation could be carried out in the brain, i.e., by means of somato-dendritic interactions like plateau potentials and backpropagating action potentials (Lillicrap et al., 2020). Thirdly, ENCORE is tested against an up-to-date range of behavioral and physiological benchmarks. ENCORE also provides an updated account of the comparator and error reconstruction hypothesis with respect to our contemporary knowledge of hippocampal physiology and its related behavior in rodents. In this respect, our model exhibits surprising features mimicking its biological counterpart that are neither trivial nor enforced by training: e.g., the expansion of the receptive fields of the reconstructed grid cells right after environmental morphing (Barry et al., 2012) and the quantitative fit to the rate remapping phenomena (Leutgeb et al., 2007).

One important component of our model, which at the same time operates as one of the main principles of our proposal, is the use of gradient descent learning in the form of error backpropagation. Despite the skepticism in using machine learning methods and computational abstractions to approximate functions of brain regions, recent studies have listed biological mechanisms potentially involved in the backpropagation of error (Guerguiev et al., 2017; Lillicrap et al., 2020). Nonetheless, the pervasive use of neural networks to emulate brain function has also raised concerns about their capability to reliably capture the computational principles and mechanisms by which biological neural systems operate (Massaro, 1988). Indeed, the concern that algorithms like error backpropagation can be super powerful seems to be justified given how they excel at solving complicated tasks achieving super-human performance (e.g. (Silver et al., 2016)). However, recent reviews highlight the conditions where neural networks can indeed be highly informative about brain mechanisms, especially when they provide specific falsifiable predictions and hypothesis (Saxe et al., 2021). In this regard, our model not only exhibits a variety of responses comparable to its biological counterpart without explicitly being trained to do so (i.e., it only learns to reconstruct its own inputs) but also is able to make falsifiable predictions that go beyond fitting specific data sets during training. For instance, our model predicts that place cell stabilization and behavioral performance should be achieved faster under completely novel environments than in moderately modified environments of a previously familiar one (Figures 5F and 5G). Indeed, this could explain the differences in pace between the

phenomena of partial and global remapping (Sanders et al., 2020). We thus suggest that, under a completely novel environment, the circuit generates error signals much more frequently with respect to its predictions (i.e., reconstructions) and that these error signals in turn boost learning in the hippocampus (e.g., by means of enhanced plasticity mediated by acetylcholine release, which is indeed driven by novelty (Jeewajee, Lever, Burton, O'Keefe and Burgess, 2008)).

By considering the role of the entorhinal comparator in generating error signals mediating learning, we emphasize the role of the EC in performing novelty detection in the hippocampus. Unlike previous models relying on oscillatory interference networks under resonance amplification (Borisyuk et al., 2001), sequential network models relying on self-supervised learning have already shown to be capable of performing match-mismatch judgments and thus novelty detection by virtue of input reconstruction and input-output comparison (Japkowicz et al., 1995). Notably, recent evidence shows that the comparison operation underlying novelty detection is likely to occur at least partly via interneuron populations within the EC (Miao et al., 2017). However, unlike previous models, we also emphasize the putative role of the DG in encoding novelty (Figure 5D) as being the region exhibiting the most prominent changes in spatial tuning after environmental modifications (Figure 2F). These phenomena indeed point to the DG as a very plastic subregion that would be more prone to rate remapping (Leutgeb et al., 2007), with its neuronal activity signaling novelty (Maass et al., 2014).

Despite the set of benchmarks chosen here to test the ENCORE's ability to capture physiological properties characteristic of the rodent hippocampus, it is far from a complete approximation of its biological counterpart. Among many untested hippocampal features, the ordering of neuronal activity in the form of theta sweeps (Gupta et al., 2012), the role of sharp wave ripples in memory formation (Buzsáki, 2015), as well as the generation of cross- frequency coding schemes (Lisman and Jensen, 2013), or the structured relation between individual neuron activity and the overall population activity (e.g., phase precession (Skaggs et al., 1996)) could not, by design, be emulated with the presented model. We argue however that some of these prominent physiological features of the hippocampus might be cast as being of an implementational nature, rather than of a computational one (i.e., functional principles). For instance, theta oscillations could provide the necessary time multiplexing for forward predictions and error backpropagation to take place alternatively in different theta phases within each cycle (O'Reilly, 1996). Similarly, phase coding enabled by cross-frequency coupling could order hippocampal ensembles in time by the magnitude of synaptic activity (Mehta et al., 2002) to facilitate credit assignment during error backpropagation. In addition, we also want to emphasize another relevant feature corresponding to the sequence generation capacity of the hippocampus, which has been argued to be its main function (Buzsáki and Tingley, 2018), acting as a glue to form episodes from temporally contiguous experiences. Further work would include an extension of the ENCORE model to capture this fundamental property by adding recurrent connections to the middle layer (CA3), which indeed is known to be highly recursive (Lebovitz et al., 1971). Also the distinct anatomical and physiological constraints of the hippocampus must be included including its pattern separation and completion capabilities. Also, generative variants such as variational autoencoders (Kingma and Welling, 2014) could be explored in relation to the putative hippocampal role in sequence generation or sampling from internal world models. These features (generativity and recurrency) would potentially shed light onto what properties of the hippocampus are explained by information compression alone or by (sequence-based) prediction.

Limitations of the study

The present study does not address all available benchmarks within the hippocampal literature. Among many untested experiments, we emphasize here the need for a more comprehensive account of the diverse remapping phenomena (Sanders et al., 2020). We believe that a full understanding of remapping and how it relates to generative processes is likely to provide further critical insights into hippocampal function. Also, we did not explicitly test for the presence of the typical range of hippocampal cells usually reported in literature (e.g., border cells, object-vector cells, head direction cells, etc.). In fact, some of them, like head direction cells, are just not possible to account for given the testing paradigm used here, where there is a lack of active navigation within the environment. Hence, actively sampling the respective EC population vectors (Figure 1D) based on velocity or direction vectors and thus mimicking what would be an active exploration of the arena would probably overcome this limitation and possibly lead to the emergence of such variety of hippocampal cells within the hidden layers of the network as we have shown in previous models (Maffei et al., 2015, DACX).

Features that might be relevant to include in further generations of the model would increase its explanatory value include: oscillations and theta-based sequences (Gupta et al., 2012), leading to phase precession (Skaggs et al., 1996) and cross-frequency coding schemes (Lisman and Jensen, 2013), and sharp wave ripples for replay and consolidation (Buzsáki, 2015). As discussed above, the generative properties of the hippocampus, together with the predictive nature of its sequence generation, are probably the most important computational features that should be added to the present model to attempt a more comprehensive account of the related literature.

Overall, the hippocampus is known to play a role in learning features and their associations over multiple dimensions, ranging from spatial location encoding (E. I. Moser, Kropff and Moser, 2008) to semantic relationships (Solomon et al., 2019) to abstract concepts (Quiroga, 2012). Even though our study largely focused on the spatial domain, similar computational mechanisms mediating position and environmental encoding should generalize across domains. The simple principle of EC mismatch error minimization, when applied to biologically realistic EC inputs, is sufficient to explain key physiological phenomena seen in the rodent hippocampus during spatial navigation. Furthermore, it demonstrates how machine learning systems could realize novelty detection, generalization, and rapid re-adaptation to environmental contingencies in an autonomous manner during navigation, thus leading to the holy grail of artificial intelligence: epistemic autonomy (Santos-Pata et al., 2021).

Resource availability

Lead contact

Further information and requests should be directed to and will be fulfilled by Paul FMJ Verschure (pverschure@ibecbarcelona.eu).

Materials availability

This study did not generate new unique reagent.

Data and code availability

The code used to simulate the model and generate the results included in the figures is available at: https://gitlab.com/diogo.santos.pata/encore.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102364.

ACKNOWLEDGMENTS

This research was supported by the EC grants Virtual Brain Cloud (number 826421) and iNavigate (number 873178), and by NPAD/UFRN. The contributions by I.S. and I.G.R. were supported by an NIH BRAIN Initiative grant (U19 NS104590).

AUTHOR CONTRIBUTIONS

D.S.P. conceived the study; D.S.P. and A.F.A. wrote the code and performed the simulations and analyses; all authors contributed to writing the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 21, 2020
Revised: January 30, 2021
Accepted: March 24, 2021
Published: April 23, 2021

REFERENCES

Alger, B.E., and Teyler, T.J. (1976). Long-term and short-term plasticity in the CA1, CA3, and dentate regions of the rat hippocampal slice. Brain Res. *110*, 463–480, https://doi.org/10.1016/0006-8993(76)90858-1.

Axmacher, N., Henseler, M.M., Jensen, O., Weinreich, I., Elger, C.E., and Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. Proc. Natl. Acad. Sci. U S A *107*, 3228–3233, https://doi.org/10.1073/pnas.0911531107.

Barry, C., Ginzberg, L.L., O'Keefe, J., and Burgess, N. (2012). Grid cell firing patterns signal environmental novelty by expansion. Proc. Natl. Acad. Sci. U S A *109*, 17687–17692, https://doi.org/10.1073/pnas.1209918109.

Bartesaghi, R., Migliore, M., and Gessi, T. (2006). Input-output relations in the entorhinal cortex-dentate-hippocampal system: evidence for a non-linear transfer of signals. Neuroscience *142*, 247–265, https://doi.org/10.1016/j.neuroscience.2006.06.001.

Benna, M.K., and Fusi, S. (2019). April 30). Are Place Cells Just Memory Cells? Memory Compression Leads to Spatial Tuning and History Dependence (BioRxiv), p. 624239, https://doi.org/10.1101/624239.

Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J.L., et al. (2013). Grid cells require excitatory drive from the hippocampus. Nat. Neurosci. *16*, 309–317, https://doi.org/10.1038/nn.3311.

Borisyuk, R., denham, M., Hoppensteadt, F., Kazanovich, Y., and Vinogradova, O. (2001). Oscillatory model of novelty detection. Network *12*, 1–20, https://doi.org/10.1080/net.12.1.1.20.

Brewer, G.J., Boehler, M.D., Leondopulos, S., Pan, L., Alagapan, S., DeMarse, T.B., and Wheeler, B.C. (2013). Toward a self-wired active reconstruction of the hippocampal trisynaptic loop: DG-CA3. Front. Neural Circuits *7*, 165, https://doi.org/10.3389/fncir.2013.00165.

Burgess, N.E., Jeffery, K.J., and O'Keefe, J. (1999). The Hippocampal and Parietal Foundations of Spatial Cognition. https://academic.oup.com/brain/article/124/1/238/286412.

Burgess, N., Maguire, E.A., and O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. Neuron *35*, 625–641, https://doi.org/10.1016/S0896-6273(02)00830-9.

Buzsáki, G. (2015). Hippocampal sharp wave-ripple: a cognitive biomarker for episodic memory and planning. Hippocampus *25*, 1073–1188, https://doi.org/10.1002/hipo.22488.

Buzsáki, G., and Tingley, D. (2018). space and time: the Hippocampus as a sequence generator. Trends Cogn. Sci. *22*, 853–869, https://doi.org/10.1016/j.tics.2018.07.006.

Clare, L., Wilson, B.A., Carter, G., Roth, I., and Hodges, J.R. (2002). Relearning face-name associations in early Alzheimer's disease. Neuropsychology *16*, 538–547, https://doi.org/10.1037/0894-4105.16.4.538.

Colgin, L.L., Moser, E.I., and Moser, M.B. (2008). Understanding memory through hippocampal remapping. Trends Neurosci. *31*, 469–477, https://doi.org/10.1016/j.tins.2008.06.008.

Davis, C.D., Jones, F.L., and Derrick, B.E. (2004). Novel environments enhance the induction and maintenance of long-term potentiation in the dentate gyrus. J. Neurosci. *24*, 6497–6506, https://doi.org/10.1523/JNEUROSCI.4970-03.2004.

De Almeida, L., Idiart, M., and Lisman, J.E. (2009). A second function of gamma frequency oscillations: an E%-max winner-take-all mechanism selects which cells fire. J. Neurosci. *29*, 7497–7503, https://doi.org/10.1523/JNEUROSCI.6044-08.2009.

Deller, T., Martinez, A., Nitsch, R., and Frotscher, M. (1996). A novel entorhinal projection to the rat dentate gyrus: direct innervation of proximal dendrites and cell bodies of granule cells and GABAergic neurons. J. Neurosci. *16*, 3322–3333, https://doi.org/10.1523/jneurosci.16-10-03322.1996.

Devauges, V., and Sara, S.J. (1990). Activation of the noradrenergic system facilitates an attentional shift in the rat. Behav. Brain Res. *39*, 19–28, https://doi.org/10.1016/0166-4328(90)90118-X.

Ekstrom, A.D., Kahana, M.J., Caplan, J.B., Fields, T.A., Isham, E.A., Newman, E.L., and Fried, I. (2003). Cellular networks underlying human spatial navigation. Nature *425*, 184–187, https://doi.org/10.1038/nature01964.

Feigenbaum, J.D., and Rolls, E.T. (1991). Allocentric and egocentric spatial information processing in the hippocampal formation of the behaving primate. Psychobiology *19*, 21–40, https://doi.org/10.1007/BF03337953.

Gluck, M.A., Meeter, M., and Myers, C.E. (2003). June 1). Computational models of the hippocampal region: linking incremental learning and episodic memory. Trends Cogn. Sci. *7*, 269–276, https://doi.org/10.1016/S1364-6613(03)00105-0.

Gluck, M.A., and Myers, C.E. (1993). Hippocampal mediation of stimulus representation: a computational theory. Hippocampus *3*, 491–516, https://doi.org/10.1002/hipo.450030410.

Guerguiev, J., Lillicrap, T.P., and Richards, B.A. (2017). Towards deep learning with segregated dendrites. ELife *6*, e22901, https://doi.org/10.7554/eLife.22901.

Gupta, A.S., Van Der Meer, M.A.A., Touretzky, D.S., and Redish, A.D. (2012). Segmentation of spatial experience by hippocampal theta sequences. Nat. Neurosci. *15*, 1032–1039, https://doi.org/10.1038/nn.3138.

Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507, https://doi.org/10.1126/science.1127647.

Howard, M.W., and Eichenbaum, H. (2015). Time and space in the hippocampus. Brain Res. *1621*, 345–354, https://doi.org/10.1016/j.brainres.2014.10.069.

Hunsaker, M.R., Rosenberg, J.S., and Kesner, R.P. (2008). The role of the dentate gyrus, CA3a,b, and CA3c for detecting spatial and environmental novelty. Hippocampus *18*, 1064–1073, https://doi.org/10.1002/hipo.20464.

Jackson, J., Amilhon, B., Goutagny, R., Bott, J.B., Manseau, F., Kortleven, C., et al. (2014). Reversal of theta rhythm flow through intact hippocampal circuits. Nat. Neurosci. *17*, 1362–1370, https://doi.org/10.1038/nn.3803.

Japkowicz, N., Japkowicz, N., Myers, C., & Gluck, M. (1995). A Novelty Detection Approach to Classification. In Proceedings of the Fourteenth Joint Conference on Artificial Intelligence, 518–523. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.3663

Jeewajee, A., Lever, C., Burton, S., O'Keefe, J., and Burgess, N. (2008). Environmental novelty is signaled by reduction of the hippocampal theta frequency. Hippocampus *18*, 340–348, https://doi.org/10.1002/hipo.20394.

Kimble, D.P., and BreMiller, R. (1981). Latent learning in hippocampal-lesioned rats. Physiol. Behav. *26*, 1055–1059, https://doi.org/10.1016/0031-9384(81)90209-2.

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. Retrieved from https://arxiv.org/abs/1312.6114v10

Knight, R.T. (1996). Contribution of human hippocampal region to novelty detection. Nature *383*, 256–259, https://doi.org/10.1038/383256a0.

Lavenex, P., and Amaral, D.G. (2000). Hippocampal-neocortical interaction: a hierarchy of associativity. Hippocampus *10*, 420–430, https://doi.org/10.1002/1098-1063.

Lebovitz, R.M., Dichter, M., and Spencer, W.A. (1971). Recurrent excitation in the ca3 region of cat hippocampus. Int. J. Neurosci. *2*, 99–107, https://doi.org/10.3109/00207457109146996.

Lee, I., Hunsaker, M.R., and Kesner, R.P. (2005). The role of hippocampal subregions in detecting spatial novelty. Behav. Neurosci. *119*, 145–153, https://doi.org/10.1037/0735-7044.119.1.145.

Leutgeb, J.K., Leutgeb, S., Moser, M.B., and Moser, E.I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. Science *315*, 961–966, https://doi.org/10.1126/science.1135801.

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

Lisman, J.E., and Jensen, O. (2013). the theta-gamma neural code. Neuron *77*, 1002–1016, https://doi.org/10.1016/j.neuron.2013.03.007.

Lörincz, A., and Buzsáki, G. (2000). Two-phase computational model training long-term memories in the entorhinal-hippocampal region. Ann. N. Y. Acad. Sci. *911*, 83–111. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.9669.

Maass, A., Schütze, H., Speck, O., Yonelinas, A., Tempelmann, C., Heinze, H.J., et al. (2014). Laminar activity in the hippocampus and entorhinal cortex related to novelty and episodic encoding. Nat. Commun. *5*, 1–12, https://doi.org/10.1038/ncomms6547.

Mack, M.L., Love, B.C., and Preston, A.R. (2018). Building concepts one episode at a time: the hippocampus and concept formation. Neurosci. Lett. *680*, 31–38, https://doi.org/10.1016/j.neulet.2017.07.061.

Massaro, D.W. (1988). Some criticisms of connectionist models of human performance. J. Mem. Lang. *27*, 213–234, https://doi.org/10.1016/0749-596X(88)90074-5.

Mehta, M.R., Lee, A.K., and Wilson, M.A. (2002). Role of experience and oscillations in transforming a rate code into a temporal code. Nature *417*, 741–746, https://doi.org/10.1038/nature00807.

Melzer, S., Michael, M., Caputi, A., Eliava, M., Fuchs, E.C., Whittington, M.A., and Monyer, H. (2012). Long-range-projecting gabaergic neurons modulate inhibition in hippocampus and entorhinal cortex. Science *335*, 1506–1510, https://doi.org/10.1126/science.1217139.

Miao, C., Cao, Q., Moser, M.B., and Moser, E.I. (2017). Parvalbumin and somatostatin interneurons control different space-coding networks in the medial entorhinal cortex. Cell *171*, 507–521.e17, https://doi.org/10.1016/j.cell.2017.08.050.

Moser, E.I., Kropff, E., and Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. Annu. Rev. Neurosci. *31*, 69–89, https://doi.org/10.1146/annurev.neuro.31.061307.090723.

Moser, E., Moser, M.B., and Andersen, P. (1993). Synaptic potentiation in the rat dentate gyrus during exploratory learning. Neuroreport *5*, 317–320, https://doi.org/10.1097/00001756-199312000-00035.

Moser, M.B., Moser, E.I., Forrest, E., Andersen, P., and Morris, R.G.M. (1995). Spatial learning with a minislab in the dorsal hippocampus. Proc. Natl. Acad. Sci. U S A *92*, 9697–9701, https://doi.org/10.1073/pnas.92.21.9697.

Nadel, L., and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. Curr. Opin. Neurobiol. *7*, 217–227, https://doi.org/10.1016/S0959-4388(97)80010-4.

O' Keefe, J., and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. Nature *381*, 425–428, https://doi.org/10.1038/381425a0.

O'Reilly, R.C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. Neural Comput. *8*, 895–938, https://doi.org/10.1162/neco.1996.8.5.895.

Olton, D.S., Becker, J.T., and Handelmann, G.E. (1980). Hippocampal function: working memory or cognitive mapping? Physiol. Psychol. *8*, 239–246, https://doi.org/10.3758/BF03332855.

Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. Science *321*, 1322–1327, https://doi.org/10.1126/science.1159775.

Quiroga, R.Q. (2012). Concept cells: the building blocks of declarative memory functions. Nat. Rev. Neurosci. *13*, 587–597, https://doi.org/10.1038/nrn3251.

Rennó-Costa, C., Lisman, J.E., and Verschure, P.F.M.J. (2010). The mechanism of rate remapping in the dentate gyrus. Neuron *68*, 1051–1058, https://doi.org/10.1016/j.neuron.2010.11.024.

Rolls, E.T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. Front. Syst. Neurosci. *7*, 74, https://doi.org/10.3389/fnsys.2013.00074.

Sanders, H., Wilson, M.A., and Gershman, S.J. (2020). Hippocampal remapping as hidden state inference. Elife *9*, 1–31, https://doi.org/10.7554/eLife.51140.

Santos-Pata, D., Amil, A.F., Raikov, I.G., Rennó-Costa, C., Mura, A., Soltesz, I., and Verschure, P.F.M.J. (2021). Epistemic autonomy: self-supervised learning in the mammalian hippocampus. Trends Cogn. Sci., In press. https://doi.org/10.1016/j.tics.2021.03.016.

Saxe, A., Nelli, S., and Summerfield, C. (2021). January 1). If deep learning is the answer, what is the question? Nat. Rev. Neurosci. *22*, 55–67, https://doi.org/10.1038/s41583-020-00395-8.

Sik, A., Ylinen, A., Penttonen, M., and Buzsáki, G. (1994). Inhibitory CA1-CA3-hilar region feedback in the hippocampus. Science *265*, 1722–1724, https://doi.org/10.1126/science.8085161.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489, https://doi.org/10.1038/nature16961.

Skaggs, W.E., McNaughton, B.L., Wilson, M.A., and Barnes, C.A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus *6*, 149–172, https://doi.org/10.1002/(SICI)1098-1063.

Solomon, E.A., Lega, B.C., Sperling, M.R., and Kahana, M.J. (2019). Hippocampal theta codes for distances in semantic and temporal spaces. Proc. Natl. Acad. Sci. U S A *116*, 24343–24352, https://doi.org/10.1073/pnas.1906729116.

Solstad, T., Boccara, C.N., Kropff, E., Moser, M.B., and Moser, E.I. (2008). Representation of geometric borders in the entorhinal cortex. Science *322*, 1865–1868, https://doi.org/10.1126/science.1166466.

Szabo, G.G., Du, X., Oijala, M., Varga, C., Parent, J.M., and Soltesz, I. (2017). Extended interneuronal network of the dentate gyrus. Cell Rep. *20*, 1262–1268, https://doi.org/10.1016/j.celrep.2017.07.042.

Wang, D. (2001). Unsupervised learning: foundations of neural computation. AI Mag. *22*, 101, https://doi.org/10.1609/AIMAG.V22I2.1565.

Wills, T.J., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. Science *308*, 873–876, https://doi.org/10.1126/science.1108905.

Yamaguchi, S., Hale, L.A., D'Esposito, M., and Knight, R.T. (2004). Rapid prefrontal-hippocampal habituation to novel events. J. Neurosci. *24*, 5356–5363, https://doi.org/10.1523/JNEUROSCI.4587-03.2004.

Yassa, M.A., and Stark, C.E.L. (2011). pattern separation in the hippocampus. Trends Neurosci. *34*, 515–525, https://doi.org/10.1016/j.tins.2011.06.006.

iScience, Volume *24*

Supplemental information

Entorhinal mismatch: A model of self-supervised learning in the hippocampus

Diogo Santos-Pata, Adrián F. Amil, Ivan Georgiev Raikov, César Rennó-Costa, Anna Mura, Ivan Soltesz, and Paul F.M.J. Verschure

Transparent Methods

To capture the putative hippocampal principles of input-output mismatch minimization, information compression, and gradient descent learning, we devised a self-supervised model in the form of an autoencoder neural network subject to error backpropagation. The input was constrained by the physiological activity arriving at DG from EC, namely grid (MEC) and sensory (LEC) cells, in the form of a conjunctive population vector.

Entorhinal cells

The activity of individual grid and sensory cells was represented by a 1x1 meter 2-dimensional array of 2 squared centimeters resolution bins. Grid cells were built based on the analytic expression implemented by (Blair, Welday, & Zhang, 2007) so that the firing rate activity of simulated cells at each location $r = (x, y)$ was given by:

$$G(r, \lambda, \theta, c) = g\left(\sum_{k=1}^{3}\left(\frac{4\pi}{\sqrt{3\lambda}}\, u(\theta_k + \theta)\cdot(r - c)\right)\right)$$

where the summation of three patterns of angles $\theta_{1:3} \in \{-30, +30, +90\}$ forms the characteristic hexagonal pattern found in grid cells with scale size defined by λ. The unitary vector pointing to the direction $\theta_k$is given by $u(\theta_k) = \big(\cos(\theta_k), \sin(\theta_k)\big)$. Samples of MEC rate maps are shown in Figure 1C. Importantly, MEC maps, unlike LEC, were set invariant to the morphing of the environment.

The LEC rate maps were generated similarly as in (Rennó-Costa, Lisman, & Verschure, 2010). First, for each rate map, the arena was divided into a 6x6 grid. Then, one third of the bins were randomly selected as active, respecting the expected spatial specificity of these cells (Hargreaves, Rao, Lee, & Knierim, 2005). The value of all active bins was set to 1. The 6x6 grid was then projected into a 60x60 new grid, conserving the relative positions of the active bins. Finally, the rate map was generated by convolving the grid with a gaussian kernel with a standard deviation of 4 bins. Samples of LEC rate maps are shown in Figure 1C. The number of grid cells (MEC) and sensory cells (LEC) was 90 and 210 correspondingly, following the LEC/MEC ratio reported in (Rennó-Costa et al., 2010).

Hippocampal model

The hippocampal model was designed as a 5-layer autoencoder. The input and output layers correspond to the population vectors of the EC (Figure 1D), which instantiates the closed-loop of the entorhinal-hippocampal system (Figure 1A). The three middle layers correspond to the DG, CA3, and CA1 and their relative sizes follow the classical shape of an autoencoder (Figure 1B), with 100, 80, and 100 units, correspondingly. All the units had a ReLU activation function. The loss function implementing the comparator hypothesis was the mean squared error (MSE) between the input and the respective output after feedforward propagation. Concretely, the loss was minimized by error backpropagation (RMSprop). The initial learning rate was set at 0.001 and the batch size at 32. One epoch corresponded to an entire pass through the arena (i.e., all the

population vectors). For each experiment, the model was trained for 1000 epochs, although convergence was normally assured within much fewer epochs (Figure 1E). After learning, the rate maps for the units in the middle layers were extracted by recording their activity values at each bin of the arena (i.e., for each population vector as input). Examples of extracted rate maps can be seen in Figure 2A.

Place field counting and size

Given a particular rate map, place fields were detected and counted by a simple clustering procedure. First, the rate map was smoothened by passing it through a gaussian kernel with a standard deviation of 3 bins. Then, the values were normalized between [0,1] by subtracting the minimum value and dividing by the maximum. The bins with values below a threshold of 0.3 were set to 0, whilst the ones above were set to 1 (i.e., active bins). The clustering was done just by assigning a cluster identity number to the different groups of active bins that were direct neighbors. The size of a cluster was determined with the total number of bins assigned to it. Finally, the number of place fields was calculated as the number of clusters that had a size within the range of [0.01, 0.2] of the total size of the rate map.

Reshaping analysis

To test how place fields changed as a function of environmental reshaping after learning, we stretched the environments as described in (O' Keefe & Burgess, 1996). MEC rate maps were horizontally expanded by maintaining the same spatial scale, thus naturally extending their corresponding hexagonal patterns (see Figure 4A, left). LEC rate maps were expanded by nearest-neighbor interpolation, effectively expanding their receptive fields (see Figure 4A, right). Then, the number of place fields and their sizes were computed for each layer (Figure 4C-D).

Rate remapping in the Dentate Gyrus

The rate remapping was computed as the correlation between the population vectors (i.e., PV correlation) of the DG units in the original and modified environments (as in (Leutgeb, Leutgeb, Moser, & Moser, 2007); see Figure 2E). Novel environments modified by a certain degree were created by generating new rate maps for the corresponding proportion of LEC cells. To control for the actual degree of modification between the novel and original environments, the remapping metric was corrected by multiplying it by the PV correlation between the LEC rate maps from both environments. Also, a general offset of -0.1 was applied to account for the imperfect correlation under identical environments reported in animal studies.

Novelty detection and relearning

A systematic analysis of the error distribution across the network when embedded in modified environments was performed. The error for the model (Figure 5B, left; Figure 5C) was computed as the MSE between the new population vector inputs and the outputs in the novel environment (i.e., reconstruction error). Moreover, error metrics across layers (Figure 5B, right; Figure 5D) were computed as the MSE of their layer-specific activity vectors between the original and modified environments. In this way, we could quantify how much error was associated with each

layer, in terms of expected place fields across layers compared to the actual ones elicited by the new inputs. Finally, for the relearning analysis (Figure 5F-G), we computed the number of epochs that the model needed to re-stabilize the learning curve (Figure 5F) to the 5% of the baseline curve (Figure 5G, "Naïve"), after different degrees of environmental modification. Data points are averages with their respective SEM of 50 independent experiments.

References

Blair, H. T., Welday, A. C., & Zhang, K. (2007). Scale-invariant memory representations emerge from moiré interference between grid fields that produce theta oscillations: A computational model. *Journal of Neuroscience*, *27*(12), 3211–3229. https://doi.org/10.1523/JNEUROSCI.4724-06.2007

Hargreaves, E. L., Rao, G., Lee, I., & Knierim, J. J. (2005). Neuroscience: Major dissociation between medial and lateral entorhinal input to dorsal hippocampus. *Science*, *308*(5729), 1792–1794. https://doi.org/10.1126/science.1110449

Leutgeb, J. K., Leutgeb, S., Moser, M. B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science*, *315*(5814), 961–966. https://doi.org/10.1126/science.1135801

O' Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*(6581), 425–428. https://doi.org/10.1038/381425a0

Rennó-Costa, C., Lisman, J. E., & Verschure, P. F. M. J. (2010). The Mechanism of Rate Remapping in the Dentate Gyrus. *Neuron*, *68*(6), 1051–1058. https://doi.org/10.1016/j.neuron.2010.11.024