UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Contextual Control of Hopfield Networks in a Hippocampal Model

Permalink

https://escholarship.org/uc/item/1bj4d35q

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Chateau-Laurent, Hugo
Alexandre, Frederic

Publication Date

2024

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Contextual Control of Hopfield Networks in a Hippocampal Model

Hugo Chateau-Laurent (hugo.chateaulaurent@gmail.com)
Inria centre at the university of Bordeaux,
Neurodegenerative Diseases Institute

Frédéric Alexandre (frederic.alexandre@inria.fr)
Inria centre at the university of Bordeaux,
Neurodegenerative Diseases Institute

Abstract

Executive functions guide episodic memory to retrieve information essential for adaptive behavior. The prefrontal cortex achieves this by influencing hippocampal processing through anatomical projections targeting the entorhinal cortex and area CA1. However, most computational models of the hippocampus overlook this cognitive control, either neglecting it or implementing implausible direct connections to the hippocampus. This paper explores the contextual control of associative memory implemented by modern Hopfield networks, within a hippocampus-inspired autoencoder. Our experiments underscore the importance of proximity between prefrontal afferences and the locus of memory storage for efficient contextual modulation of episodic memory, challenging the standard model of hippocampal processing. These findings not only advance our understanding of higher-level cognition but also provide design principles for more adaptive machine learning algorithms.

Keywords: Episodic memory; cognitive control; hippocampus; prefrontal cortex; modern Hopfield networks; computational neuroscience

Introduction

Efficient behavioral guidance relies on the selective recall of memories pertinent to the current context. This process is intricately shaped by the dynamic interplay between contextual and sensory information. Contextual information, encompassing internal representations beyond immediate perception, is instrumental in predicting future rewards and states. It includes temporal event sequences, spatial integration, and overarching goals. While substantial research has focused on stimulus-driven memory retrieval, the modulation of episodic memory recall by contextual factors remains less explored. Understanding these mechanisms is crucial not only for advancing episodic memory theory but also for developing machine learning algorithms capable of navigating diverse spatiotemporal contexts flexibly.

The prefrontal cortex is known to guide the recall of relevant memories in the hippocampus (Eichenbaum, 2017). However, in what has been termed the standard model (Nadel & Moscovitch, 1997) or standard framework (Cheng, 2013), episodic memories in the hippocampus are stored in a region called CA3, which lacks a direct connection from the prefrontal cortex (Andrianova et al., 2023). This raises a fundamental question about how the prefrontal cortex influences memory retrieval in CA3. In a modeling study (Pilly, Howard, & Bhattacharyya, 2018), contextual information was fed to the dentate gyrus to modulate recall. This is at odds with prefrontal projections which are thought to terminate in the entorhinal cortex (Eichenbaum, 2017) and potentially reach downstream layer CA1 through the thalamus (Anderson, Bunce, & Barbas, 2016). The aim of this paper is to study how prefrontal signals can modulate memory through the entorhinal cortex or CA1. An autoencoder artificial neural network is used to model the entorhinal–hippocampal circuit. A layer representing the prefrontal cortex sends context to one of the autoencoder layers (target of prefrontal afferences). A modern Hopfield network (Ramsauer et al., 2020) is placed in one of the hippocampal layers and used to recall the closest state of that layer, mimicking the storage of episodic memory, by taking into account sensory and context information. Performance and contextual representations are studied based on the position of the Hopfield memory relative to the target. As the results challenge the current view of episodic storage, the implications for the standard model are discussed in the last section.

Methods

Task

Like the model of Pilly et al. (2018), our model performs the tasks described in Peters, David, Marcus, and Smith (2013). In these tasks, object-discrimination problems are presented in two different contexts and used to evaluate context-guided memory recall. In a given task t_i and context c, the network learns to choose a rewarded object $R_{i,c}$ over a non-rewarded object $\bar{R}_{i,c}$. In the alternative context c', it learns to choose between the same objects as in context c, except that the non-rewarded object in context c is rewarded in c', and conversely (i.e. $\bar{R}_i^c = R_i^{c'}$ and $R_i^c = \bar{R}_i^{c'}$). Objects are represented in the network as 6×4 matrices with 6 active entries of value 1 placed at random locations, and other entries with value 0 (Figure 1).

Model

The model is composed of an autoencoder, a modern Hopfield associative memory, and a prefrontal layer (Figure 2).

The autoencoder is a multi-layered fully-connected neural network which follows the anatomy of the medial temporal lobe. The input is first sent to the entorhinal cortex, then flows to CA1 through a trisynaptic pathway ($\mathrm{EC_{II}} \rightarrow \mathrm{DG} \rightarrow \mathrm{CA3} \rightarrow \mathrm{CA1}$) and a monosynaptic pathway ($\mathrm{EC_{III}} \rightarrow \mathrm{CA1}$). Information is sent from CA1 to $\mathrm{EC_V}$, then flows to the output

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

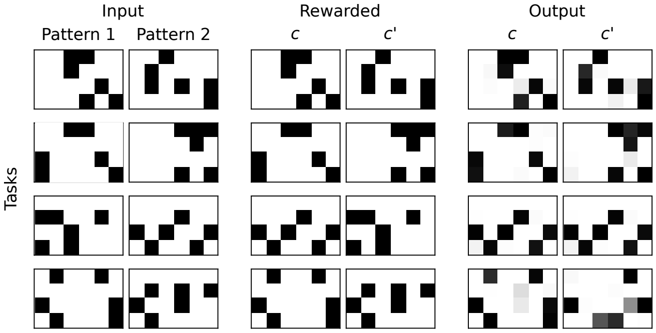

Figure 1: Example tasks in which the model succeeds (first two rows) and fails (last two rows) to recall the correct memory during the test phase. The two choice patterns that are placed in the first slots of the input are shown on the left. The rewarded patterns of contexts c and c' are shown in the middle. On the right, the choice of the network (i.e. third slot of the output) is shown for both contexts. In the first two tasks, the output of the network successfully corresponds to the rewarded pattern in both contexts. In the third task, however, the network chooses the same pattern in both contexts and thus fails to complete the task. In the fourth task, the network outputs the correct pattern in context c but outputs a pattern that is not in the choice set for that task in c'. The trial is incorrect as the output is closer to $\bar{R}_i^{c'}$ than $R_i^{c'}$. For these examples, the network was pretrained on 10000 tasks with a learning rate of $10^{-4.5}$.

layer. ReLU is used as the activation function of hidden layers and sigmoid is used in the output layer (Table 1). The number of neurons in each region is set as the number of neurons in the analog rat region, as retrieved from Cutsuridis, Graham, Cobb, and Vida (2019), scaled down by a factor $\alpha = 0.003$. The network input is the concatenation of the two patterns of the choice objects and the pattern of the rewarded object (hidden during test). The assignment of objects R_i^c and $\bar{R}_i^c$ to the first or second slot is random to prevent the network from learning to always output the same slot. An additional input, corresponding to the prefrontal cortex, is used to modulate recall contextually (Figure 2). It represents the two possible contexts as a two-dimensional one-hot vector. When a layer is targeted by this contextual layer, this simply means that it receives it as an additional input.

One of the hippocampal layers can be paired with a modern Hopfield network. During the acquisition phase described in the next section, the activity vectors of that layer are appended to an autoassociative memory matrix X. During the test phase, the activity of the layer, say a, is replaced by the output of the Hopfield network a^*:

$$a^* = X \mathrm{softmax}(\beta X^\top a), \tag{1}$$

where β is set to 1,000 to strongly separate memories. The memory-augmented activity a^* is then sent to the next layers of the autoencoder.

Table 1: Layers (sizes and activation functions) of the autoencoder. The “Objects” input and output size corresponds to the product of the number, width and height of input patterns.

Layer	Output size	Function
Context Input	2	
Objects Input	$3 \times 6 \times 4$	
$\mathrm{EC_{II}}$	$110{,}000\alpha$	ReLU
$\mathrm{EC_{III}}$	$250{,}000\alpha$	ReLU
DG	$1{,}200{,}000\alpha$	ReLU
CA3	$250{,}000\alpha$	ReLU
CA1	$390{,}000\alpha$	ReLU
$\mathrm{EC_{V}}$	$330{,}000\alpha$	ReLU
Objects Output	$3 \times 6 \times 4$	sigmoid
Context Output	2	sigmoid

Simulation Details

Pretraining Phase All task-context combinations are presented once during each epoch, in random order. Learning is online, as training examples are presented one by one. Similarly to the recently proposed autoencoder model of the hippocampus (Santos-Pata et al., 2021), the network learns to output its input, namely the two input object patterns and the rewarded one, as well as the context vector (unless otherwise mentioned). This is done by updating feedforward autoencoder weights using backpropagation, in a pretraining phase. The loss function to be minimized is the mean squared error between the input and the output.

Acquisition Phase Subsequently, in the acquisition phase, the weights of the autoencoder are fixed and memories are stored in the associative memory. Two training regimes are used to assess the need for the fast associative memory: one-shot and many-shot. In the many-shot regime, pretraining lasts 20 epochs with 50 tasks. The same 50 tasks are then acquired in the memory and used to test performance. Contrastingly, in the one-shot regime akin to episodic memory, pretraining lasts a single epoch with 1000 tasks, such that networks are updated as many times as in the many-shot regime. In the acquisition phase of the one-shot regime, 50 novel tasks are presented to the network for memory acquisition, then used to test performance. Hence, in the one-shot regime, the autoencoder weights are not trained with the acquisition/test tasks.

Test Phase While the rewarded pattern is provided as input to the network during the pretraining and training phases, it is masked during test (i.e. values of the third input slot are set to 0). In the test phases, a trial is considered correct when the pattern stored in the third slot of the output is closer to the rewarded pattern than the non-rewarded one, as assessed by the mean squared error (Figure 1). Thus, the expected performance of a random network, or a network with no contextual target, is 50% and serves as baseline. The learning rate

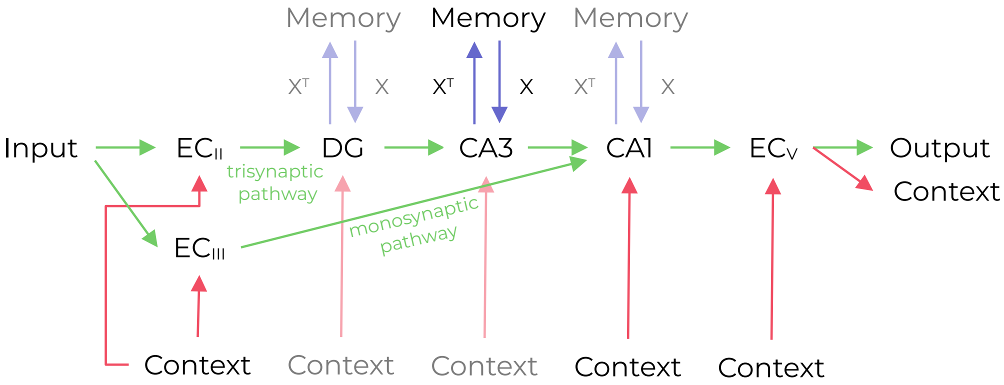

Figure 2: Architecture of the model consisting of an autoencoder (green), an associative memory (blue) and a prefrontal cortex layer targeting the autoencoder with context information (red). The autoencoder is trained to output its input and the context when indicated. The associative memory is modified during the acquisition phase, and used to recall memories during the test phase. In the standard model of the hippocampus, memory is in CA3 (strong blue), but other locations are also tested here (transparent blue). Similarly, context information can target various layers (transparent red), but the current view is that the prefrontal sends context to the entorhinal cortex (Eichenbaum, 2017) or to CA1 (Anderson et al., 2016), as shown in strong red.

is a sensitive parameter, as its optimal value varies among the conditions tested in this work. Therefore, for each tested condition, the learning rate is set as the average of the five learning rates that best performed during a random search. This purely stochastic optimization process evaluated the performance of the model in each condition, with 300 random learning rate values. After this process, 50 different simulations were performed with the optimal learning rate of each condition.

Results

Behavioral Performance

The need for associative memory is first assessed to ensure that the tasks require episodic processing. With the many-shot regime, a one-tailed *t*-test comparing the performance with Hopfield memory in CA3 (M = 65.06% correct, SD = 7.85) and without modern Hopfield network (M = 70.64%, SD = 4.05) indicates that the models can perform the tasks overall, but that associative memory slightly impairs performance with this training regime ($t(49) = -4.97$, $p < .001$). On the other hand, with the one-shot regime akin to episodic memory, results with CA3 memory (M = 75.06%, SD = 4.64) and without modern Hopfield network (M = 49.54%, SD = 3.19) indicate that memory is necessary when test tasks are not used to pretrain the autoencoder (one-tailed $t(49) = 34.89$, $p < .001$). This demonstrates that the encoding and recall operations of modern Hopfield networks can take context into account. In these conditions, CA3 is directly targeted by context information. To test whether the model can perform the task in accordance with the anatomical review of (Eichenbaum, 2017), context information is then provided to entorhinal layers which receive prefrontal projections in the biological brain. The performance with entorhinal target (M = 51.06%, SD = 3.05) and CA3 target (M = 75.06%, SD = 4.64) indicate that the model is much less performant when context information is given to the entorhinal cortex (one-tailed $t(49) = 28.2$, $p < .001$), and performs almost no better than chance (Figure 3).

When the dentate gyrus is the memory-augmented layer, the performance when the entorhinal cortex is targeted (M = 70.68%, SD = 4.49) and when the dentate gyrus is directly targeted (M = 71.44%, SD = 4.7) indicate that performance is not significantly impaired when targeting entorhinal layers (one-tailed $t(49) = 1.12$, $p = .13$), as shown in Figure 3. Similarly, when the associative memory is in CA1, the model performs well when context is provided to upstream layers (Figure 3). When superficial layers of the entorhinal cortex are targeted and memory is in CA1, the performance of the model with the monosynaptic pathway (M = 70.14%, SD = 4.71) and without it (M = 56.28%, SD = 4.05) indicate that this pathway is necessary to convey context information to the memory (one-tailed $t(49) = 17.55$, $p < .001$), as shown in Figure 3. By combining results of when the dentate gyrus, CA3 or CA1 (no monosynaptic pathway for simplicity) is memory-augmented, it is possible to evaluate performance according to the position of the targeted layer relative to the position of the memory layer. This analysis reveals that targeting regions downstream memory has no effect (Figure 4). Spearman's rank correlation was computed to assess the relationship between the relative position of the target (restricted to relative positions less than or equal to 0) and performance. There was a positive correlation between the two variables, $r(898) = 0.78, p < .001$. Furthermore, a two-way ANOVA was performed to evaluate the interaction effects of relative position (less than or equal to 0) and context reconstruction on performance. Without context reconstruction, only the sensory (objects) input is taken into consideration in the autoencoder loss. The results indicated a significant main effect for relative position ($F(3,899) = 545.59, p < .001$), main effect for context reconstruction ($F(1,899) = 24.29, p < .001$), and interaction between relative position and context reconstruction ($F(3,899) = 9.06, p < .001$). Figure 4 indeed shows that the steepness of performance decline, as target is located more and more upstream memory, is greater when the model only tries to reconstruct sensory input than when it also tries to reconstruct context information in the pretraining phase.

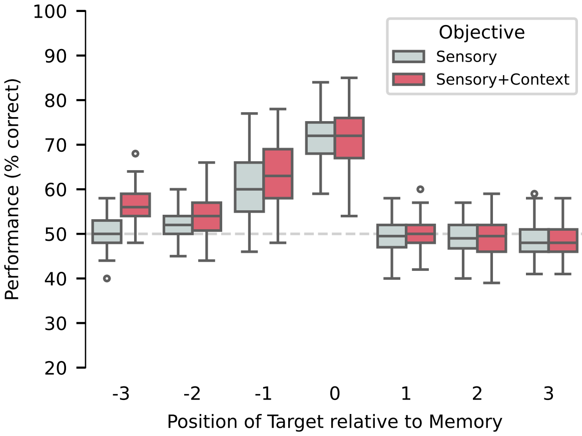

Figure 4: Performance as a function of the location of the target relative to the associative memory, using sensory only and full reconstruction losses. No monosynaptic pathway was used with CA1 memory for clarity. Relative positions P and -P indicate that the target and the memory are P layers apart. The target is located before the memory with -P, and after with P. The gray dashed line indicates the baseline performance of 50%.

Neural Coding of Context

In order to understand how the model either succeeds or fails at performing the task, the representations employed in the different layers can be analyzed. First, a model with CA3 memory and entorhinal target is used. After pretraining and acquisition, principal component projections reveal that contexts are well separated at the targeted EC_{II} population (Figure 5), but this separation diminishes in downstream layer CA3 (Figure 6), to such an extent that the associative memory in CA3 is not able to separate memories contextually (not shown). Downstream layers thus cannot know the context to answer correctly.

One possibility is that some neurons are specifically modulated by context identity and underlie the separation in the principal component space. In order to find such context cells, paired t-tests are used to assess whether the activity of each cell is dependent on the context (either stronger or weaker), treating the 50 test tasks as samples. Neurons with $p < .001$ are considered context cells. This analysis reveals that 68% of EC_{II} neurons in this model are context cells (Figure 5), and that this number decreases in the next populations (e.g. 45% of CA3 cells in Figure 6).

The presence of context-modulated cells is very likely to help the model performing the tasks. In order to ensure that context cells indeed have causal influence on behavioral performance and the contextual separation of activity, inactivation experiments are performed. Since prefrontal information can hardly reach CA3 when the the entorhinal cortex is targeted, the associative memory is placed in DG. After the test phase, 1664 context cells were identified in DG (46%). This might not seem a lot compared to the proportion in CA3, but since DG is bigger than CA3, this represents many more neurons and is enough to separate contexts much better than CA3 in the principal component space (not shown). After identification, the DG context cells were inactivated during an additional test phase by clamping their activity to 0 in all 50 test tasks. With this protocol, the contexts are no longer separated in the principal component space (Figure 7), as compared to when 1664 neurons are randomly selected for inactivation instead (Figure 8). Furthermore, while random inactivation yielded 72% correct answers (similar to results of Figure 3), performance dropped to 52% when context cells were inactivated.

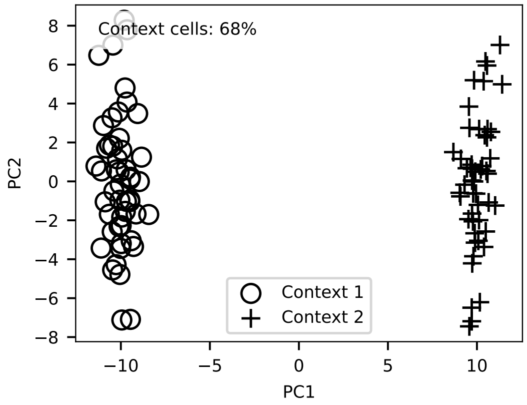

Figure 5: Principal component analysis of EC_{II} activity and proportion of context cells. Here, EC_{II} is the target and CA3 is the memory.

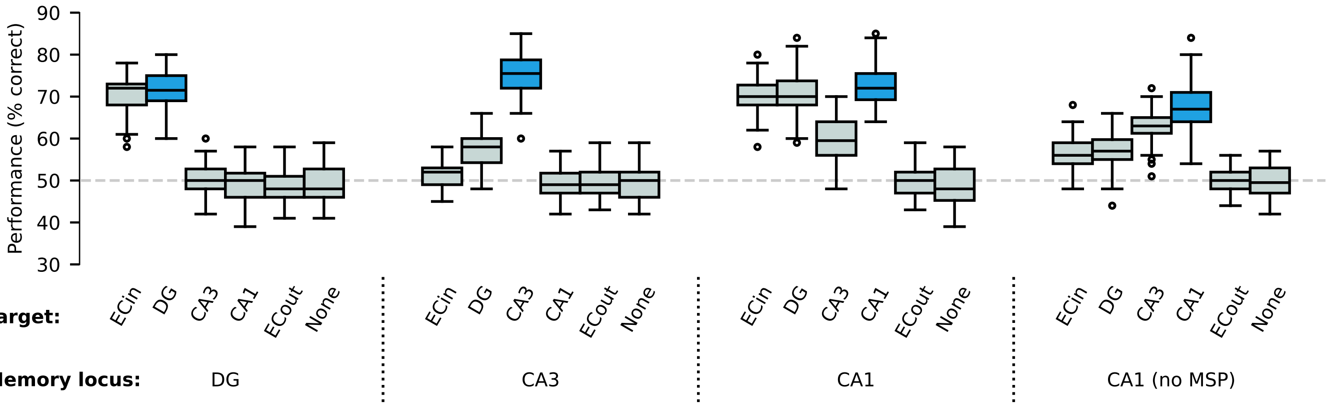

Figure 3: Performance of the model with different memory loci (blue) and targets. The performance with CA1 memory and no monosynaptic pathway is also shown on the right. The gray dashed line indicates the baseline performance of 50%.

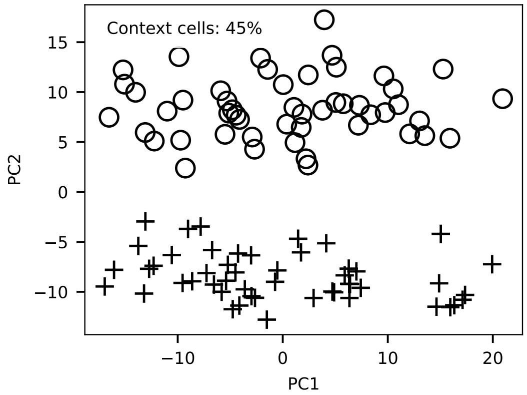

Figure 6: Principal component analysis of CA3 activity and proportion of context cells. Here, EC_{II} is the target and CA3 is the memory.

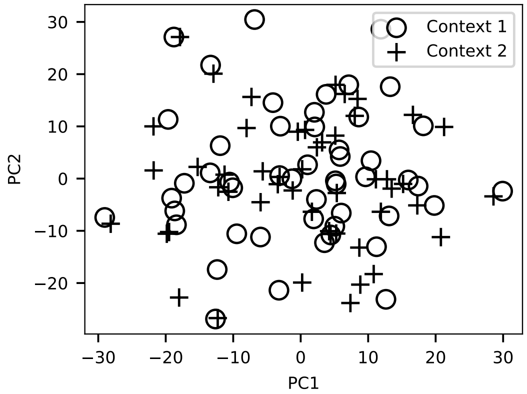

Figure 7: Principal component analysis of DG with EC_{II} as a target, DG as a memory, and 1664 DG context cells inactivated.

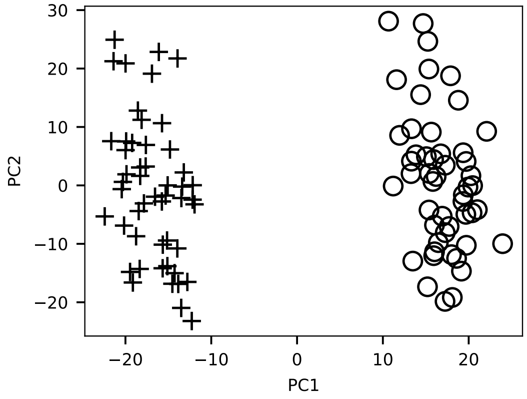

Figure 8: Principal component analysis of DG with EC_{II} as a target, DG as a memory, and 1664 random DG cells inactivated.

Discussion

A task requiring one-shot episodic memory was introduced by Peters et al. (2013) and used in the present work to study context-guided retrieval. Although rewards were encoded artificially using a third stimulus slot to match the encoding scheme of Pilly et al. (2018), this task enabled us to evaluate the performance of diverse anatomical configurations in a simple hippocampus-inspired network. The behavioral results presented here reveal that modern Hopfield networks can encode and retrieve contextual memories. The fast associative memory is especially necessary to perform the task when stimuli are only presented once. In the many-shot regime, however, pairing a memory with CA3 impairs performance. Perhaps when training examples are shown multiple times like in the many-shot regime, the network learns to generalize and ignore contextual information. An alternative explanation is that this training regime develops networks that associate object representations more downstream than in the one-shot regime, such that the CA3 memory is less able to recall the rewarded pattern when masked during test.

Furthermore, it is found that the position of the target of contextual information is a key factor in correctly performing the task. The locus of the associative memory is also important. In fact, performance seems to be governed by the distance between these two elements, such that the target must be close to, but not downstream of, the associative memory layer. Thus, the model that follows the standard model of hippocampal processing and the anatomical review of prefrontal-hippocampal interactions of Eichenbaum (2017) does not perform well. This is due to the associative memory in CA3 being too far from prefrontal projections to superficial layers of the entorhinal cortex. While there is a direct connection from the entorhinal cortex to CA3 in the anatomy, which could potentially improve performance, it is not active during encoding in the standard model. These results, along with the evidence put forward by Cheng (2013), challenge the standard claim that memories are stored in the recurrent collaterals of CA3.

Some alternative models are much better at performing the task. This is for example the case of models whose associative memory is located in a layer directly targeted by the prefrontal context. Alternatively, models with associative memory in DG and CA1 work well with the entorhinal cortex as the target. This is due to DG and CA1 being connected to the entorhinal cortex through the trisynaptic and monosynaptic pathways respectively. It is unlikely that CA1 is the only locus of memory because information sent by $\mathrm{EC}_{\mathrm{II}}$ lacks the temporal nature necessary for sequence learning (Hafting, Fyhn, Bonnevie, Moser, & Moser, 2008), and the trisynaptic pathway is known to participate in memory. The alternative hypothesis is that the role of DG in memory storage is underestimated, and that a new model could employ it to store and retrieve contextual memories (Chateau-Laurent, 2024).

Another factor influencing performance is whether the networks learn to reconstruct context during the pretraining phase. Models that take context into account in the reconstruction error perform better. This is not surprising, as backpropagation then favors the transmission of context information across more layers to reconstruct it in the output layer. Since the hippocampus has been hypothesized to learn to autoencode its entorhinal input (Santos-Pata et al., 2021; Ketz, Morkonda, & O'Reilly, 2013), and since the entorhinal cortex both receives input from and projects back to the prefrontal cortex, it is reasonable to assume that context reconstruction is a training objective. Machine learning models aiming to leverage associative memory in a context-dependent manner should harness these results. More precisely, the architectural distance between contextual cues and memory module should be small and objective functions should include a cue reconstruction term whenever possible.

Beyond behavioral performance, intermediate representations have been analyzed. The model has been found to encode context explicitly, as suggested by the separation of contexts in the principal component space and the discovery of context-modulated cells. Most importantly, these context cells have been found to be necessary to separate contexts and perform the task correctly. Context information must indeed be encoded in the input of the modern Hopfield network for it to encode and recall memories contextually. Context neurons are reminiscent of splitter cells discovered in the hippocampus (Wood, Dudchenko, Robitsek, & Eichenbaum, 2000; Frank, Brown, & Wilson, 2000; Duvelle, Grieves, & Van der Meer, 2023), which split place representation according to the past or future trajectory. The context cells found in this work can be considered splitter cells without explicit temporal and spatial components. Future work could explore controlled episodic memory in more details with more elaborate contextual modulation and navigation tasks, in order to provide a more extensive account of splitter cells. The potential role of DG in memory storage should also be explored in more depth.

For simplicity and epistemological parsimony, the networks employed here were unconstrained feedforward networks and modern Hopfield networks. However, biological plausibility elements such as sparsity constraints and more detailed associative memory mechanisms could be incorporated to assess whether the same results hold in networks closer to the biological hippocampus.

Acknowledgements

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see `https://www.plafrim.fr`).

References

Anderson, M. C., Bunce, J. G., & Barbas, H. (2016). Prefrontal–hippocampal pathways underlying inhibitory control over memory. *Neurobiology of learning and memory*, *134*, 145–161.

Andrianova, L., Yanakieva, S., Margetts-Smith, G., Kohli, S., Brady, E. S., Aggleton, J. P., & Craig, M. T. (2023). No evidence from complementary data sources of a direct glutamatergic projection from the mouse anterior cingulate area to the hippocampal formation. *Elife*, *12*, e77364.

Chateau-Laurent, H. (2024). *Computational modeling of the interactions between episodic memory and cognitive control*. Unpublished doctoral dissertation, University of Bordeaux.

Cheng, S. (2013). The crisp theory of hippocampal function in episodic memory. *Frontiers in neural circuits*, *7*, 88.

Cutsuridis, V., Graham, B. P., Cobb, S., & Vida, I. (2019). *Hippocampal microcircuits: a computational modeler's resource book*. Springer.

Duvelle, É., Grieves, R. M., & Van der Meer, M. A. (2023). Temporal context and latent state inference in the hippocampal splitter signal. *ELife*, *12*, e82357.

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558.

Frank, L. M., Brown, E. N., & Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, *27*(1), 169–178.

Hafting, T., Fyhn, M., Bonnevie, T., Moser, M.-B., & Moser, E. I. (2008). Hippocampus-independent phase precession in entorhinal grid cells. *Nature*, *453*(7199), 1248–1252.

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS computational biology*, *9*(6), e1003067.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, *7*(2), 217–227.

Peters, G. J., David, C. N., Marcus, M. D., & Smith, D. M. (2013). The medial prefrontal cortex is critical for memory retrieval and resolving interference. *Learning & memory*, *20*(4), 201–209.

Pilly, P., Howard, M., & Bhattacharyya, R. (2018). Modeling contextual modulation of memory associations in the hippocampus. *Frontiers in Human Neuroscience*, *12*, 442.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., … others (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Santos-Pata, D., Amil, A. F., Raikov, I. G., Rennó-Costa, C., Mura, A., Soltesz, I., & Verschure, P. F. (2021). Entorhinal mismatch: A model of self-supervised learning in the hippocampus. *Iscience*, *24*(4).

Wood, E. R., Dudchenko, P. A., Robitsek, R. J., & Eichenbaum, H. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron*, *27*(3), 623–633.