

# Correcting the Hebbian Mistake: Toward a Fully Error-Driven Hippocampus

Yicong Zheng, Xiaonan L. Liu, Satoru Nishiyama, Charan Ranganath, and Randall C. O'Reilly

Departments of Psychology and Computer Science

Center for Neuroscience

University of California, Davis

1544 Newton Ct

Davis, CA 95616

oreilly@ucdavis.edu

October 30, 2021

## Abstract:

The hippocampus plays a critical role in the rapid learning of new episodic memories. Many computational models propose that the hippocampus is an autoassociator that relies on Hebbian learning (i.e., “cells that fire together, wire together”). However, Hebbian learning is computationally suboptimal as it modifies weights unnecessarily beyond what is actually needed to achieve effective retrieval, causing more interference and resulting in a lower learning capacity. Our previous computational models have utilized a powerful, biologically plausible form of error-driven learning in hippocampal CA1 and entorhinal cortex (EC) (functioning as a sparse autoencoder) by contrasting local activity states at different phases in the theta cycle. Based on specific neural data and a recent abstract computational model, we propose a new model called Theremin (Total Hippocampal ERror MINimization) that extends error-driven learning to area CA3 — the mnemonic heart of the hippocampal system. In the model, CA3 responds to the EC monosynaptic input prior to the EC disynaptic input through dentate gyrus (DG), giving rise to a temporal difference between these two activation states, which drives error-driven learning in the EC→CA3 and CA3↔CA3 projections. In effect, DG serves as a teacher to CA3, correcting its patterns into more pattern-separated ones, thereby reducing interference. Results showed that Theremin, compared with our original model, has significantly increased capacity and learning speed. The model makes several novel predictions that can be tested in future studies.

## Introduction

It is well-established that the hippocampus plays a critical role in the rapid learning of new episodic memories (Eichenbaum, Yonelinas, & Ranganath, 2007). Most computational and conceptual models of this hippocampal function are based on principles first articulated by Donald O. Hebb and David Marr (Hebb, 1949; Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995; McNaughton & Nadel, 1990). At the core of this framework is the notion that recurrent connections among CA3 neurons are strengthened when they are co-activated (“cells that fire together, wire together”), essentially creating a cell assembly of interconnected neurons that bind the different elements of an event. As a result of this Hebbian learning, subsequent partial cues can drive pattern completion to recall the entire original memory, by reactivating the entire cell assembly via the strengthened interconnections.

In addition, Marr’s fundamental insight was that sparse levels of neural activity in area CA3 and especially the dentate gyrus (DG) granule cells, will drive the creation of cell assemblies that involve a distinct

---

R. C. O'Reilly is Director of Science at Obelisk Lab in the Astera Institute, and Chief Scientist at eCortex, Inc., which may derive indirect benefit from the work presented here.

Supported by: ONR grants N00014-20-1-2578, N00014-19-1-2684/ N00014-18-1-2116, N00014-18-C-2067 (MDM)

combination of neurons for each event, otherwise known as *pattern separation* (Marr, 1971; O'Reilly & McClelland, 1994; Yassa & Stark, 2011). As a consequence, the DG to CA3 pathway has the capability to minimize interference from learning across even closely overlapping episodes (e.g., where you parked your car today vs. where you parked it yesterday). Note that it is the patterns of activity over area CA3 that constitute the principal hippocampal representation of an episodic memory, and learning in these CA3 synapses is thus essential for cementing the storage of these memories. Overall, the basic tenets established by Hebb and Marr account for a vast amount of behavioral and neural data on hippocampal function, and represents one of the most widely accepted theories in neuroscience (Eichenbaum, 2016; Milner, Squire, & Kandel, 1998; O'Reilly, Bhattacharyya, Howard, & Ketz, 2014; Yonelinas, Ranganath, Ekstrom, & Wiltgen, 2019)

Although almost every biologically-based computational model of hippocampal function incorporates Hebbian plasticity, it is notable that Hebbian learning is computationally suboptimal in various respects, especially in terms of overall learning capacity (Abu-Mostafa & St. Jacques, 1985; Treves & Rolls, 1991). In models that rely solely on Hebbian learning, whenever two neurons are active together, the synaptic weight between them is increased, regardless of how necessary that change might be to achieve better memory recall. As a result, such models do not know when to stop learning, and continue to drive synaptic changes beyond what is actually necessary to achieve effective pattern completion. The consequence of this “learning overkill” is that all those unnecessary synaptic weight changes end up driving more interference with the weights needed to recall other memories, significantly reducing overall memory capacity. Even the high degree of pattern separation in the DG and CA3 pathways might not be sufficient to make up for the interference caused by reliance on Hebbian learning. Although it is difficult to quantitatively assess the capacity of the hippocampus in various species, there is reason to believe that even the high degree of pattern separation in the DG and CA3 pathways might not be sufficient to make up for the interference caused by reliance on Hebbian learning.

One logical alternative to the simple Hebbian approach is to introduce a self-limiting learning mechanism that drives only synaptic changes that are absolutely necessary to support learning. However, determining this minimal amount of learning can be challenging: how can local synaptic changes “know” what is functionally necessary in terms of the overall memory system function? One well-established class of such learning mechanisms are error-driven learning rules: by driving synaptic changes directly in proportion to a functionally-defined error signal, learning automatically stops when that error signal goes to zero. For example, the well-known Rescorla-Wagner learning rule for classical conditioning (Rescorla & Wagner, 1972) is an instance of the delta-rule error-driven learning rule (Widrow & Hoff, 1960):

$$dW = x(r - y), \quad (1)$$

where  $dW$  is the amount of synaptic weight change,  $x$  is the sending neuron activity level (e.g., average firing rate of sensory inputs representing conditioned stimuli),  $r$  is the actual amount of reward received, and  $y$  is the expected amount of reward, computed according to the existing synaptic weights:

$$y = \sum xW. \quad (2)$$

This learning rule drives learning (changes in weights,  $dW$ ) up to the point where the expected prediction of reward ( $y$ ) matches the actual reward received ( $r$ ), at which point learning stops, because the difference term in the parentheses goes to 0. The dependency on  $x$  is critical for *credit assignment*, which ensures that the most active sending neurons change their weights the most, as such weight changes will be the most effective in reducing the error. The widely-used error backpropagation learning algorithm is a mathematical extension of this simpler delta-rule form of learning (Rumelhart, Hinton, & Williams, 1986), and demonstrates the general-purpose power of these error-driven learning principles, underlying the current success in large-scale deep learning models (LeCun, Bengio, & Hinton, 2015).

We have previously shown that these error-driven learning principles can be applied to the CA1 region of the hippocampus (Ketzel, Morkonda, & O'Reilly, 2013), building on theta-phase dynamics discovered by Hasselmo, Bodelon, and Wyble (2002). The critical *target* value driving this error-driven learning is the full pattern of activity over the entorhinal cortex (EC) representing the current state of the rest of the cortex. Learning stops when the hippocampal encoding of this EC pattern projecting from CA3 through CA1 matches the target version driven by the excitatory projections into the EC. These error-driven learning dynamics have been supported empirically by studies of CA1 learning in various tasks (Schapiro, Turk-Browne, Botvinick, & Norman, 2017; Schapiro, Turk-Browne, Norman, & Botvinick, 2016). However, this prior model retained the standard Hebbian learning for all of the connections within CA3 and DG, because the error signal that drives CA1 learning does not have any way of propagating back to these earlier areas within the overall circuit: the connectivity is only feedforward from CA3 to CA1.

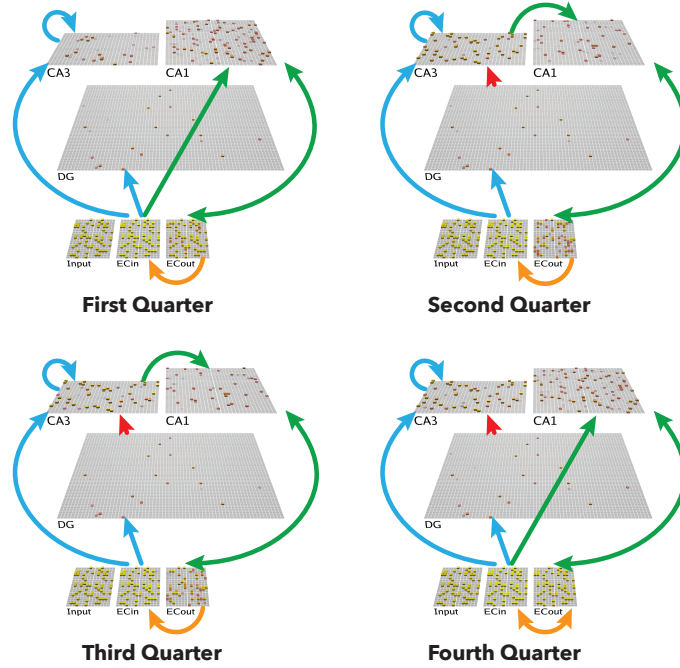
To be able to apply a similar type of self-limiting error-driven learning to the core area CA3 of the hippocampus, we need a suitable target signal available to neurons within the CA3 that determines when the learning has accomplished what it needs to do. Recently, Kowadlo, Ahmed, and Rawlinson (2020) proposed in an abstract, backpropagation-based model that the DG can serve as a kind of teacher to the CA3, driving learning just to the point where CA3 on its own can replicate the same highly pattern-separated representations that the DG imparts on the CA3. We build on this idea here, by showing how error-driven learning based on this DG-driven target signal can emerge naturally within the activation dynamics of the hippocampal circuitry, driving learning in the feedforward and recurrent synapses of area CA3. Thus, we are able to extend the application of error-driven learning to the “heart” of the hippocampus.

We show that this more fully error-driven hippocampal learning system has significantly improved memory capacity and resistance to interference compared to one with Hebbian learning in CA3. Furthermore, we show how these error-driven learning dynamics fit with detailed features of the neuroanatomy and physiology of the hippocampal circuits, and can have broad implications for understanding important learning phenomena such as the *testing effect* (Liu, O'Reilly, & Ranganath, 2021). Overall, this new framework provides a consistent computational and biological account of hippocampal episodic learning, which departs from the tradition of Hebbian learning at a computational level, while retaining the overall conceptual understanding of the essential role of the hippocampus in episodic memory. Thus, we do not throw the baby out with the bathwater here, and our model remains consistent with the vast majority of behavioral and neural data consistent with the classic Hebb-Marr model. Nevertheless, it also does make novel predictions, and at a broad, behavioral level, the improved performance of our model is more compatible with the remarkable capacity of the relatively small hippocampal system for encoding so many distinct memories over the course of our lives.

In the remainder of the paper, we first introduce the computational and biological framework for error-driven learning in the hippocampal circuits, and then present the details of an implemented computational model, followed by results of this model as compared to our previous Hebbian-CA3 version (Ketzel et al., 2013), as well as representational analyses that capture the subregional dynamics in the model. We conclude with a general discussion, including testable predictions from this framework and implications for some salient existing behavioral and neural data on hippocampal learning.

### Sources of Error Driven Learning in the Hippocampal Circuit

We begin by briefly reviewing our earlier work showing how the monosynaptic pathway interconnecting the entorhinal cortex (EC) and area CA1 can support error-driven learning, via systematic changes in pathway strengths across the theta cycle (Hasselmo et al., 2002; Ketzel et al., 2013) (Figure 1). Although nominally a central part of the hippocampus, from a computational perspective it makes more sense to think of this monosynaptic CA1  $\leftrightarrow$  EC pathway (sometimes known as the temporo-ammonic pathway) as an ex-



**Figure 1:** Architecture of the Theremin model. Visual depiction of one full theta-cycle training trial, separated into four different phases within the cycle (i.e., four *quarters*). The CA1 learns to properly decode the CA3 pattern into the corresponding EC representation, while CA3 learns to encode the EC input in a more pattern-separated manner reflecting DG input. Arrows depict pathways of particular relevance for that quarter. First quarter: Blue arrows show initial activation of CA3 and DG via monosynaptic pathways from ECin (superficial layers of EC). Green shows CA1 likewise being monosynaptically driven from ECin, and in turn driving Ecout (deep layers). Second quarter: Red arrow indicates DG driving CA3, providing a target activity state over CA3 relative to the first quarter state. Also, CA3 starts to drive CA1, resulting in full “attempted recall” state over Ecout by the end of the Third quarter. In the Fourth quarter, the ECin drives Ecout (Orange arrow), which in turn drives any resulting changes in CA1. The activation state at the end of this quarter represents the full target state for all error-driven learning throughout the hippocampus.

tension of the neocortex, where the principles of error-driven learning have been well-developed (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020; O’Reilly, 1996; Whittington & Bogacz, 2019). Specifically, this pathway can be thought of as learning to encode the EC information in CA1 in a way that can then support reactivation of the corresponding EC activity patterns when memories are later retrieved via CA3 pattern completion. Computationally, this is known as an *auto-encoder*, and error-driven learning in this case amounts to adjusting the synapses in this monosynaptic pathway to ensure that the EC pattern is accurately reconstructed from the CA1 activity pattern.

Using the delta rule equation shown above, this objective can be formulated as:

$$dW = CA1(ECin - Ecout), \quad (3)$$

where *ECin* is the EC input pattern driven by cortical inputs to the EC, and *Ecout* is the output produced by the CA1 encoding of this *ECin* pattern (Figure 1). If these are identical, then the error is 0, and no learning occurs, and any learning that does occur is directly in proportion to the extent of error correction required. The detailed biological story for how the system could implement this error-driven learning mechanism depends on the relative out-of-phase strengths of the two major projections into the CA1 (Hasselmo et al., 2002): in one phase the CA1 is driven more strongly by direct ECin layer 3 inputs, and in another it is driven more by the CA3 projections. Furthermore, the Ecout (deep layers) is initially driven by CA1, but then is driven more by ECin (layer 3).

The net effect is that, by learning based on temporal differences in activation state over time, the delta rule

is realized. Specifically, the difference in E<sub>Cout</sub> activation between their CA1-driven vs. ECin-driven states produces the (ECin – E<sub>Cout</sub>) difference in the above equation. Likewise, CA1 neurons directly experience a “reflection” of this temporal difference, conveyed via the projections from E<sub>Cout</sub> → CA1, and they also can learn using their own local delta-rule-like equation based on this temporal difference. Mathematically, this is a form of error backpropagation (Lillicrap et al., 2020; O’Reilly, 1996; Whittington & Bogacz, 2019). See Ketz et al. (2013) for more details. Consistent with the idea that this monosynaptic pathway is more cortex-like in nature, Schapiro et al. (2017) have shown that this pathway can learn to integrate across multiple learning experiences to encode sequential structure, in a way that depends critically on the error-driven nature of this pathway, and is compatible with multiple sources of data (Schapiro et al., 2016).

To extend this error-driven learning mechanism to area CA3, it is essential to have two different activation states, one that represents a better, target representation (e.g., the actual reward, or the actual ECin input in the examples considered previously), and the other that represents what the current synaptic weights produce on their own. The key idea in our new Theremin model is that the highly pattern-separated activity pattern in the DG drives a target representation as a pattern of activity over CA3, which can be compared to what the CA3 can produce on its own (i.e., prior to the arrival of DG → CA3 inputs). Thus, in effect, the DG, which is the sparsest and most pattern-separated hippocampal layer, is serving as a teacher to the CA3, driving error-driven learning signals there just to the point where CA3 on its own can replicate the DG-driven sparse, pattern-separated representations (Kowadlo et al., 2020).

The delta-rule formulation for this new error-driven learning component is:

$$dW = \text{ECin}(\text{CA3}_{dg} - \text{CA3}_{nodg}), \quad (4)$$

where *ECin* is the sending activity into CA3 (via the perforant pathway (PP) projections),  $\text{CA3}_{dg}$  is the activity of the CA3 neurons when being driven by the strong mossy fiber (MF) inputs from the DG, and  $\text{CA3}_{nodg}$  is the CA3 activity when not being driven by those DG inputs, i.e., from only the PP and CA3 recurrent inputs prior to the arrival of DG → CA3 inputs. Critically, to the extent that CA3 prior to DG input is already matching the DG-driven pattern, no additional learning needs to occur, thus producing the interference minimization benefits of error-driven learning. Note that the same error-driven signal in CA3 trains the lateral recurrent pathway within CA3 in addition to the ECin → CA3 PP projections, so that these recurrent connections also adapt to fit the DG-driven pattern, but no further.

Although this form of error-driven learning might make sense computationally, how could something like this delta error signal emerge naturally from the hippocampal biology? First, as in our prior model of learning in the monosynaptic pathway (Ketz et al., 2013), we adopt the idea that the delta arises as a *temporal difference* between two states of activity over the CA3, which is also consistent with a broader understanding of how error-driven learning works in the neocortex; (O’Reilly, 1996; O’Reilly & Munakata, 2000; O’Reilly, Russin, Zolfaghar, & Rohrlich, 2021). The appropriate temporal difference over CA3 should actually emerge naturally from the additional delay associated with the propagation of the MF signal through the DG to the CA3, compared to the more direct PP signal from ECin to CA3. Thus, as in our prior models, the minus phase term in the delta rule occurs first (i.e.,  $\text{CA3}_{nodg}$ ), followed by the plus phase (this terminology goes back to the Boltzmann machine, which also used a temporal-difference error-driven learning mechanism; Ackley, Hinton, & Sejnowski, 1985).

Neurophysiologically, there are a number of lines of empirical evidence consistent with this temporal difference error-driven learning dynamic in the CA3:

- CA3 pyramidal cells respond to PP stimulation prior to the granule cells (GCs) in the DG, in vivo (Do, Martinez, Martinez, & Derrick, 2002; Yeckel & Berger, 1990), such that the indirect input through the DG will be that much more delayed due to the slower DG response (by roughly 5 msec at least).

- MF inputs from the DG GCs to the CA3 can induce heterosynaptic plasticity at PP and CA3 recurrent connections (Kobayashi & Poo, 2004; McMahon & Barrionuevo, 2002; Rebola, Carta, & Mulle, 2017; Tsukamoto et al., 2003). This is consistent with ability of the later-arriving DG inputs to drive the CA3 synaptic changes toward that imposed by this stronger target-like pattern, compared to the earlier pattern initially evoked by PP and CA3 recurrent inputs.
- Although several studies have found that contextual fear learning is intact without MF input to CA3 (Kitamura et al., 2015; McHugh et al., 2007; Nakashiba et al., 2012), incomplete patterns from DG during encoding impair the function of EC  $\rightarrow$  CA3 pathway in contextual fear conditioning tasks (Bernier et al., 2017), suggesting that DG still plays an important role in heterosynaptic plasticity at CA3.

In addition to this DG-driven error learning in CA3, we explored a few other important principles that also help improve overall learning performance. First, reducing the strength of the MF inputs to the CA3 during memory recall helped shift the dynamics toward pattern completion instead of pattern separation, as was hypothesized in O'Reilly and McClelland (1994). This is consistent with evidence and models showing that MF projections are not necessary in naturally recalling a memory (Bernier et al., 2017; Nakashiba et al., 2012; Rolls, 2013). However, other data suggests that it still plays an important role in increasing recall precision (Bernier et al., 2017; Nakashiba et al., 2012; Pignatelli et al., 2019; Ruediger et al., 2011). Thus, consistent with these data, we found that reducing, but not entirely eliminating MF input to the CA3 during recall was beneficial, most likely because it enabled the other pathways to exert a somewhat stronger influence in favor of pattern completion, while still preserving the informative inputs from the DG.

Second, we experimented with the parameters on the one remaining Hebbian form of learning in the network, which is in the ECin  $\rightarrow$  DG pathway (i.e., the perforant path). This pathway does not have an obvious source of error-driven contrast, given that there is only one set of projections into the DG granule cells. Thus, we sought to determine if there were particular parameterizations of Hebbian learning that would optimize learning in this pathway, and found that shifting the balance of weight decreases over weight increases helped learning overall, working to increase pattern separation in this pathway still further.

Finally, we tested a range of different learning rates for all of the pathways in the model, along with relative strengths of the projections, across a wide range of network sizes and numbers of training items, to determine the overall best parameterization under these new mechanisms.

Next, we describe our computational implementation within the existing Ketz et al. (2013) framework, and then present the results of a systematic large-scale parameter search of all relevant parameters in the model, to determine the overall best-performing configuration of the new model.

## Methods

### *Hippocampal Architecture*

The current model, which we refer to as the Theremin (i.e., Total Hippocampal ERror MINimization) (Figure 1), is based on our previous theta-phase hippocampus model (Ketz et al., 2013), which was developed within the earlier Complementary Learning System (CLS) model of the hippocampus (Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). The broader implementation framework is the Leabra model (Local, Error-driven, and Associative, Biologically Realistic Algorithm), which provides standard point-neuron rate-coded neurons, inhibitory interneuron-mediated competition and sparse, distributed representations, full bidirectional connectivity, and temporal-difference based error-driven learning dynamics (O'Reilly & Munakata, 2000; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012). See

<https://github.com/emer/leabra> for fully-documented equations, code, and several example simulations, including the exact model presented here.

Figure 1 shows the standard hippocampal architecture captured in our models. The EC superficial layer (ECin) is the source of input to the hippocampus, integrated from all over the cortex. Based on anatomical and physiological data, we organize the EC into different pools (also called slots) that reflect the inputs from different cortical areas, and thus have different types of representations reflecting the specializations of these different areas (Witter, Doan, Jacobsen, Nilssen, & Ohara, 2017). In the present model, we assume some pools reflect item-specific information, while others reflect the various aspects of information that together constitute context, which is important for distinguishing different memory lists in our tests.

The ECin projects to the DG and CA3 via broad, diffuse PP projections, which have a uniform 25% random chance of connection. This connectivity is essential for driving conjunctive encoding in the DG and CA3, such that each receiving neuron receives a random sample of information across the full spectrum present in the ECin. Further, the DG and CA3 have high levels of inhibition, driving extreme competition, such that only those neurons that have a particularly favorable conjunction of input features are able to get active in the face of the strong inhibition. This is the core principle behind Marr’s pattern separation mechanism, captured by his simple R-theta codon model (Marr, 1971). Using the standard FFFB (feedforward & feedback) inhibition mechanism in Leabra, DG has a inhibitory conductance multiplier of 3.8 and CA3 has 2.8, compared to the standard cortical value of 1.8 which produces activity levels of around 15%. These resulted in DG activity around 1% and CA3 around 2%. The number of units in DG is roughly 5 times of that in CA3, consistent with the theta-phase hippocampus model.

The CA3 receives strong MF projections from the DG, which have a strength multiplier of 4 (during encoding), giving the DG a much stronger influence on CA3 activity compared to the direct PP inputs from ECin. CA3 also receives recurrent collateral projections which have a strength multiplier of 2, which are the critical Hebbian cell-assembly autoassociation projections in the standard Hebb-Marr model, as captured in Ketz et al. (2013) using a Hebbian learning mechanism. That model also uses Hebbian learning in the PP pathways from ECin to DG and CA3, which also facilitate pattern completion during recall as analyzed in O’Reilly and McClelland (1994).

In the monosynaptic pathway, ECin (superficial) layers project to CA1, which then projects back into the deep layers of EC, called Ecout in the model, such that CA1 encodes the information in ECin and can drive Ecout during recall to drive the hippocampal memory back out into the cortex. This is the auto-encoder function of CA1, which is essential for translating the highly pattern-separated representations in CA3 back into the “language” of the cortex. Thus, a critical locus of memory encoding is in the CA3 → CA1 connections that associate the CA3 conjunctive memory with the CA1 decoding thereof — without this, the randomized CA3 patterns would be effectively unintelligible to the cortex.

Unlike the broad and diffuse PP projections, the EC ↔ CA1 connections obey the pool-wise organization of EC, consistent with the focal, point-to-point nature of the EC ↔ CA1 projections (Witter et al., 2017). Thus, each pool is separately auto-encoding the specific, specialized information associated with a given cortical area, which enables these connections to slowly learn the systematic “language” of that area. The entire episodic memory is thus the distributed pattern across all of these pools, but the monosynaptic pathway only sees a systematic subset, which it can efficiently and systematically auto-encode.

The purpose of the theta-phase error-driven learning in the Ketz et al. (2013) model is to shape these synaptic weights to support this systematic auto-encoding of information within the monosynaptic pathway. Specifically, CA1 patterns at peaks and troughs of theta cycles come from CA3-retrieved memory and ECin inputs, respectively. As shown in equation 3 above, the target plus-phase activation comes from the Ecout being strongly driven by the ECin superficial patterns, in contrast to the minus phase where Ecout is being driven directly by CA1. Thus, over iterations of learning, this error-driven mechanism shapes synapses so that the CA1 projection to Ecout will replicate the corresponding pattern on ECin.

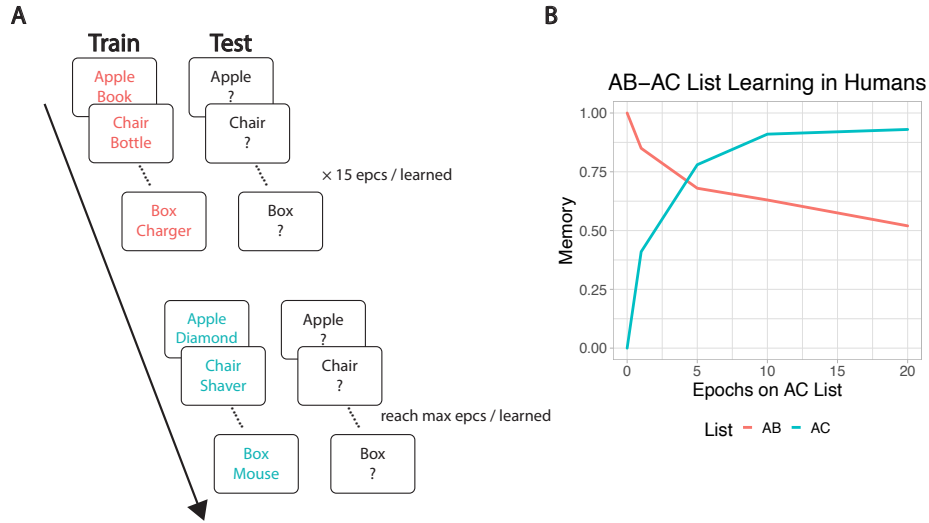


Figure 2: AB-AC list learning paradigm diagram and human data reproduced from an empirical experiment (Barnes & Underwood, 1959). A) The first AB list is trained until memory accuracy reaches 100% or 15 epochs, whichever is less; the second AC list is then trained to same criterion, while continuing to test AB and AC items. Detailed procedure is described in Methods. B) Human participants show moderate interference of the AB list after learning the AC list.

Relative to this theta-phase model, the new Theremin model introduces error-driven learning in the CA3, using equation 4 as shown above, which was achieved by delaying the output of  $DG \rightarrow CA3$  until the 2nd theta phase cycle (Figure 1). Thus, the plus phase represents the CA3 activity in the presence of these strong DG inputs, while the minus phase is the activity prior to the activation of these inputs. In addition, as noted earlier, we tested the effects of reducing the strength of MF inputs to CA3 during recall testing, along with testing all other relevant parameters in a massive grid search.

### Model Testing

The task used in the current study is a standard AB-AC paired-associates list-learning paradigm, widely used to stress interference effects (Barnes & Underwood, 1959; McCloskey & Cohen, 1989) (Figure 2). In these paradigms, typically, a participant learns a list of word pairs, with each pair referred to as *A-B*. Once the pairs are learned to a criterion or a fixed number of repetitions, participants learn a new list of *A-C* word pairs, in which the first word in each *A-B* pair is now associated with a new word. Learning of *A-C* pairs is typically slowed due to competition with previously learned *A-B* pairs (*proactive interference*), and once the *A-C* pairs are learned, retention of *A-B* pairs is reduced (*retroactive interference*).

To simulate the AB-AC paradigm, each pair of A and B items (unique random bit patterns in the model) was trained, and then tested by probing with the A item and testing for recall of the associated B item. A list context representation was also present during training and testing, to distinguish the AB vs. AC list. Once recall accuracy for all AB pairs reached 100%, or 15 epochs of the whole AB list have been trained, the model switched to learn the AC list, where previously learned A items were paired with novel C items and AC list context. Similarly, if memory for all AC pairs reached 100%, or 30 epochs have been trained in total, that run was considered complete. We ran 30 different simulated subjects (i.e., runs) on each configuration and set of parameters, with each such subject having a different set of random initial synaptic weights.

There are several central questions that we address in turn. First, we compared the earlier theta-phase hippocampus model with the new Theremin model to determine the overall improvement resulting from the new error-driven learning mechanism and other optimized parameters. This provides an overall sense



of the importance of these mechanisms for episodic memory performance, and an indication of what kinds of problems can now be solved using these models, at a practical level. In short, the Theremin model can be expected to perform quite well learning challenging, overlapping patterns, opening up significant new practical applications of the model.

Next, we tested different parameterizations of the Theremin model, to determine the specific contributions of: 1) error-driven learning specifically in the CA3, compared to Hebbian learning in this pathway, with everything else the same; 2) reduced MF strength during testing (cued recall); 3) the balance of weight decreases vs. increases in the ECin  $\rightarrow$  DG projections; 4) the effect of pretraining on the monosynaptic pathway between EC and CA1, which simulates the accumulated learning in CA1 about the semantics of EC representations, reflecting in turn the slower learning of cortical representations. In other words, human participants have extensive real life experience of knowing the A/B/C list items, enabling the CA1 to already be able to invertably reconstruct the corresponding EC patterns for them, and pretraining captures this prior learning. Pretraining has relatively moderate benefits for the Theremin model, and was used by default outside of this specific test. The pretraining process involved turning DG and CA3 off, while training the model with items and context separately only in the monosynaptic EC  $\leftrightarrow$  CA1 pathway for 5 epochs.

The learning capacity of a model is proportional to its size, so we tested a set of three network sizes (small, medium, large, see Appendix for detailed parameters) to determine the relationship between size and capacity in each case. The list sizes ranged from 20 to 100 pairs of AB–AC associations (for comparison, Barnes and Underwood (1959) used 8 pairs of nonsense syllables). For the basic performance tests, the two dependent variables were NEpochs and ABmem. NEpochs measures the total number of epochs used to finish one full run through AB and AC lists, which measures the overall speed of learning (capped at 30 if the network failed to learn). ABmem is the memory for AB pairs after learning the AC list, thus representing the models’ ability to resist interference.

In addition to these performance tests, we ran representational analyses on different network layers (i.e., hippocampal subregions) over the course of learning. This enabled us to directly measure the temporal difference error signals that drove learning in Theremin, and how representations evolved through learning. Furthermore, by comparing across differences in learning algorithm and other parameters, we can directly understand the overall performance differences. The main analytic tool here is to compute cycle-by-cycle correlations between the activity patterns present at that cycle and the patterns present at the end of a trial of processing (100 cycles), which provides a simple 1-dimensional summary of the high-dimensional layer activation patterns as they evolve over time.

## Results

### *Overall memory performance*

First, we examined the broadest measure of overall learning performance improvements in Theremin compared to the earlier theta-phase model from Ketz et al. (2013). Figure 3 shows the results across all three network sizes and numbers of list items. For all three network sizes, the ABmem results show that Theremin was better at counteracting interference and retained more memory for AB pairs than the theta-phase hippocampus model across all list sizes and network sizes ( $ps < .01$  except SmallHip List100 ( $p = 0.736$ )). Moreover, the full Theremin model completed learning significantly faster (i.e., the NEpochs measure) than the theta-phase hippocampus model across all list sizes and network sizes ( $ps < .01$  except SmallHip List100 (all NEpochs = 30)).

To more specifically test the effects of the new error-driven CA3 mechanism in the Theremin model, we directly compared the Theremin model with another Theremin model without error-driven CA3 component (labeled as NoEDL), but with everything else the same. For this and subsequent comparisons, we focus

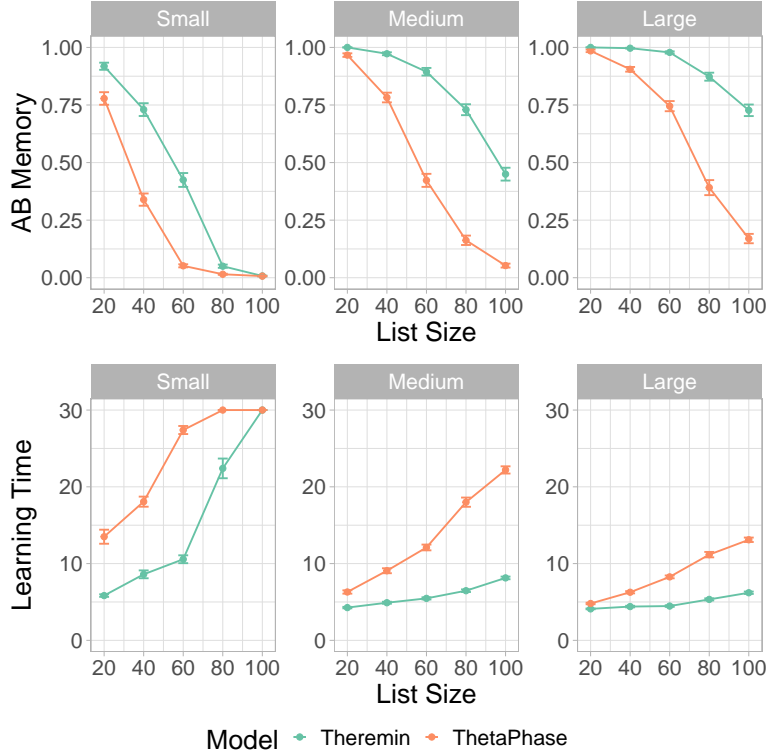


Figure 3: Theremin vs. ThetaPhase on AB memory and learning time for all three network sizes. The Theremin model was better at counteracting interference across all list sizes and network sizes, and had significantly faster training time across all list sizes and network sizes.

on the medium and large network sizes, as the small case often failed to learn at all for larger list sizes. Figure 4 shows that, except for the smallest list size (20 items), Theremin retained significantly more AB memory ( $ps < .01$ , except large network with list size of 20 ( $p = 0.321$ )) and learned faster ( $ps < .01$ , except large network with list size of 20 ( $p = 0.343$ )) than NoEDL. Thus, it is clear that this error-driven learning mechanism is responsible for a significant amount of the improved performance of the Theremin model relative to the earlier theta-phase model.

To determine the contributions of the other new mechanisms included in the Theremin model, we compared the full Theremin to versions without each of these mechanisms (Figure 4). The NoDynMF version eliminated the mechanism of dynamically decreasing the strength of MF inputs from DG to the CA3 during recall, and the results show a significant effect on performance for all but the smallest list size (20 items) (NEpoch  $ps < .01$ , except large network with list size of 20 ( $p = .013$ ), ABMem  $ps < .01$ , except large network with list size of 20 and 80 ( $p = 0.127$ )).

To determine the importance of learning in ECin  $\rightarrow$  DG pathway overall, we tested a NoDGLearn variant with no learning at all in this pathway. In principle, the DG could support its pattern separation function without any learning at all, relying only on the high levels of pattern separation and random PP connectivity. However, we found that learning in this pathway is indeed important, with an overall decrease in performance for larger list sizes (above 40 items) (NEpoch  $ps < .01$ , BigHip List40  $p = .011$ ; ABMem  $ps < .01$ , BigHip List40  $p = .025$ ). Interestingly, as the list size scaled up, the NoDGLearn model learned increasingly more slowly, such that it was even slower than the theta-phase model at a list size of 100. This effect is attributable to the strong effect of DG on training the CA3, and when the DG’s ability to drive strong pattern separation is compromised, it significantly affects CA3 and thus the overall memory performance.

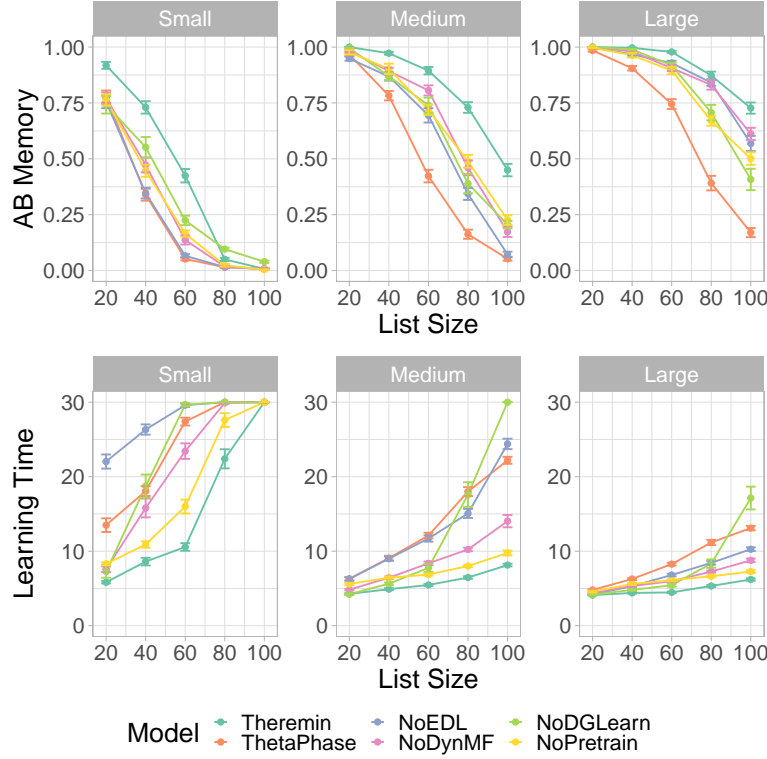


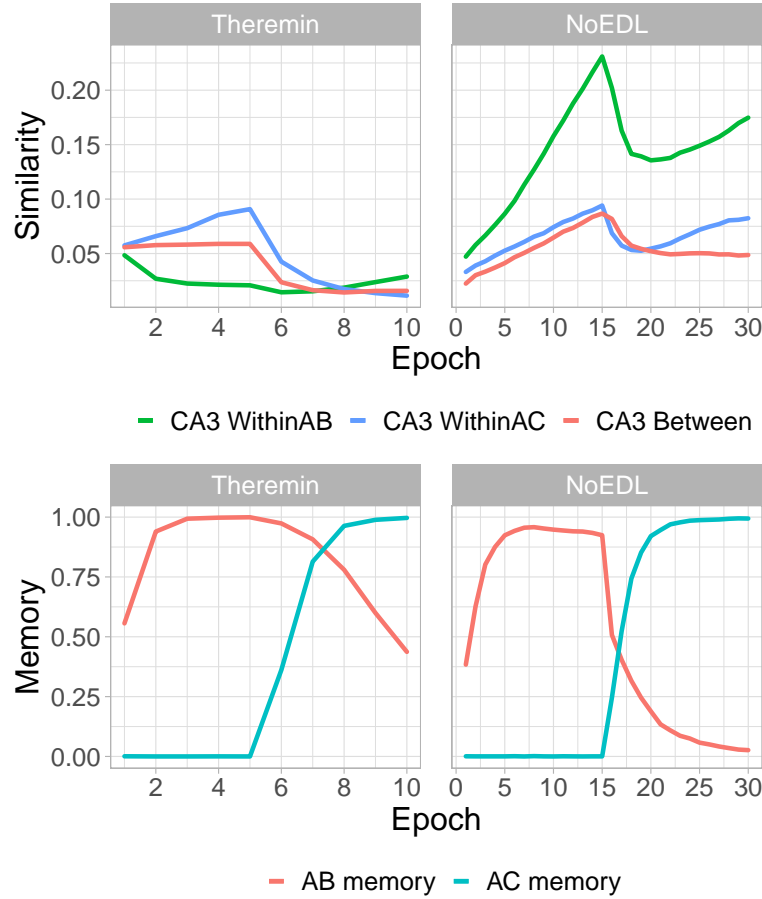
Figure 4: Theremin vs. other Models on AB memory and learning time. NoEDL is the Theremin without the new error-driven learning mechanism. NoDynMF is the Theremin with same mossy fiber strength during training and testing. NoDGLearn is the Theremin with ECin  $\rightarrow$  DG learning off. NoPretrain is the Theremin without pretraining CA1. Each of these factors makes a significant contribution as seen in decrements relative to the full Theremin in interference resistance (AB memory) and learning time.

The higher rate of weight decrease (LTD = long-term depression in biological terms) relative to weight increases in the ECin  $\rightarrow$  DG pathway were also important: eliminating this asymmetry significantly decreased performance for larger list sizes (above 60 items) (NEpoch  $ps < .01$ , SmallHip List60  $p = .051$ ; ABMem  $ps < .05$ , BigHip List80  $p = .129$ ). We also found that a lower learning rate in the ECin  $\rightarrow$  DG pathway improved the ABMem score (reducing interference), but resulted in slower learning, and vice-versa for higher learning rates, consistent with the fundamental tradeoff between learning rate and interference that underlies the complementary learning systems framework (McClelland et al., 1995). Likewise, due to optimized parameters in Theremin, comparing it to a lower or higher learning rate model would result in significant improvement in ABMem or NEpoch, respectively, but not both. Thus, we compared two Theremin variants that had dramatic differences in both ABMem and NEpoch. Higher learning rate resulted in faster learning ( $ps < .01$ ) but less ABMem ( $ps < .01$ ) compared to lower learning rate for list sizes over 40, vice versa.

The final mechanism we tested was the pretraining of the EC  $\leftrightarrow$  CA1 encoder pathway, to reflect long-term semantic learning in this pathway. The NoPretrain variant showed significantly worse performance at all but the smallest list sizes (NEpoch  $ps < .01$ ; ABMem  $ps < .05$  except BigHip List20 ( $p = .155$ )).

### Representational dynamics

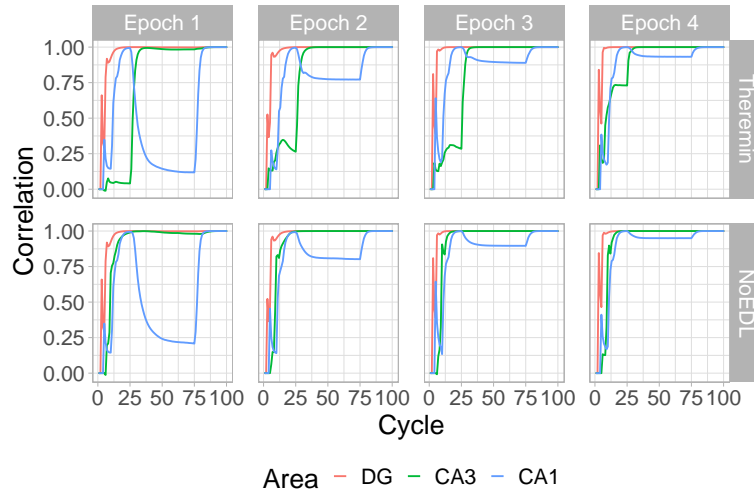
Having established the basic memory performance effects of the error-driven CA3 and other mechanisms in the Theremin model, we now turn to some analyses of the network representations and dynamics to attempt to understand in greater detail how the error-driven learning shapes representations during learning,



**Figure 5:** Statistics for area CA3 over the course of testing (List 100, Medium sized network). Representational similarity analyses (RSA) for area CA3 for Theremin vs. NoEDL show how the error-driven learning in Theremin reduces the representational overlap (top left) whereas the Hebbian learning in NoEDL increases the representational overlap (top right). This explains the differential interference as shown in the AB Memory plot for each case (bottom row). The number of epochs used in Theremin training was set to a fixed number (i.e., 10) that enabled complete learning of AB and AC lists, while in NoEDL was set to the maximum amount used in the current paper (i.e., 30).

how the activation dynamics unfold over the course of the theta cycle within a single trial of learning, and how these dynamics change over multiple iterations of learning. For these analyses, we focus on the 100-item list size, and the medium sized network, comparing the full Theremin model vs. the NoEDL model, to focus specifically on the effects of error-driven learning in the CA3 pathways.

Figure 5 shows a representational similarity analysis (RSA) of the different hippocampal layers over the course of learning, comparing the average correlation of representations in CA3 within each list (all AB items and all AC items, e.g., A1B1 vs. A2B2) and between lists (AB vs. AC, e.g., A1B1 vs. A1C1). These plots also show the proportion of items correctly recalled from each list, with the switch over from the AB to AC list happening half-way through the run (we fixed this crossover point to enable consistent averaging across 30 simulated subjects, using a number of epochs that allowed successful learning for each condition). These results show that the error-driven learning in the full Theremin model immediately learns to decrease the similarity of representations within the list (e.g., WithinAB when learning AB) and between lists over training, while the Hebbian learning in the NoEDL model fails to separate these representations and results in increases in similarity over time. This explains the reduced interference and improved learning times for the error-driven learning mechanism, and is consistent with the idea that the continuous weight changes



**Figure 6:** Changes in hippocampal subregions’ pattern similarities over the course of the first 4 epochs of learning, within a full trial for an example AB pair (model timing equivalent to 100 ms), for the Theremin (top row) and NoEDL models (bottom row). Each line reflects the correlation of the current-time activity pattern relative to the activity pattern at the end of the trial. Two major effects are evident. First, the CA1 pattern learns over epochs to quickly converge on the final plus-phase activation state, based on learning in the CA3  $\rightarrow$  CA1 pathway. Second, the Theremin model shows how the CA3 pattern learns over epochs to converge on the DG-driven activation state that arises after cycle 25, reflecting CA3 error-driven learning. Additionally, big-loop signals from ECont back to ECin could be observed from cycle 25 to 75 in the first epoch for both models, shifting CA3 patterns slightly off its final patterns. Drops seen within the first few cycles were due to the settling of temporally different patterns and were not of interest to the current paper.

associated with Hebbian learning are deleterious.

Figure 6 shows an example AB pair plot, with each layer’s correlation with the final activation state at the end of the trial across 4 training epochs. As illustrated in the plot, the learning dynamics in DG, CA3 and CA1 layers follow different learning rules across 4 quarters in one trial. In the CA3, error-driven learning in the Theremin model causes its activation to converge over the course of learning based on the target DG input that arrives starting after cycle 25. This learning progression is not evident in the NoEDL model, where Hebbian learning in the CA3 establishes a relatively stable representation early on. The CA1 shows increasing convergence to the final plus-phase pattern starting in the second and third quarter (cycle 26-75), when CA3 starts to drive CA1. Interestingly, there is evidence for a “big-loop” error signal (Kumaran & McClelland, 2012) reflecting activation circulating through the trisynaptic pathway and out through the EC, and back into the CA3, deviating its pattern from the stabilized one, as depicted by the slightly curved green line in the first epoch.

To elaborate on the error-driven learning dynamics in the Theremin model, as learning progresses, the CA3 pattern in the first quarter becomes increasingly similar to its final pattern (Figure 6). In effect, this similarity signal reflects how close the CA3 pattern is to its final DG-dominated pattern, before DG starts to have an effect on CA3. In the first epoch, CA3 is driven only by ECin  $\rightarrow$  CA3 and CA3  $\rightarrow$  CA3 inputs, resulting in a large temporal difference (error signal), which in turn modifies these connections (i.e., heterosynaptic plasticity). This error becomes smaller fast, and learning will stop when there is no more error. On the other hand, the NoEDL model continually increases the synaptic weights between CA3 and other regions whenever two neurons are active together, according to the Hebbian learning principle.

## Discussion

By incorporating biologically plausible error-driven learning mechanisms at the core CA3 synapses in our computational model of the hippocampus, along with a few other important optimizations, we have been able to significantly improve learning speed and memory capacity (resistance to interference) compared to our previous model that used Hebbian learning in these synapses. These results demonstrate the critical ability of error-driven learning to automatically limit further learning once it has achieved sufficient changes to support effective memory recall, which then significantly reduces the amount of interference that would otherwise occur from continued synaptic changes. Furthermore, representational similarity analysis (RSA) was used to illustrate temporal dynamics within the hippocampal formation, which explains the effects of the error-driven learning mechanism, making it possible for the model to make specific subregional predictions that could be tested in experiments.

Another contribution of the current model is to test several computationally-motivated mechanisms that improve overall performance, and could plausibly be implemented in the hippocampal biology. First, decreasing the DG  $\rightarrow$  CA3 strength during recall improves performance, because the DG otherwise biases more strongly toward pattern separation rather than the pattern completion needed for recall. Second, as opposed to the intuitive idea that the DG naturally forms highly separated patterns, learning in ECin  $\rightarrow$  DG pathway is important for overall performance. Furthermore, favoring of LTD over LTP in this pathway is beneficial as it forces DG to form sparse representations, suggesting that learning overall is helping with the DG pattern separation dynamics. Although our previous model without all of these improvements was sufficient for simulating smaller-scale one-off experiments, the significantly improved capacity of the present model opens up the potential to examine longer time-scale learning contributions of the hippocampus, and other larger-scale datasets as emphasized in Kowadlo et al. (2020).

The Theremin model retains the major tenets of the Hebb-Marr paradigm based on rapid episodic learning, while incorporating error-driven learning to optimize the learning capacity of the system relative to the predominant use of Hebbian learning in other models. In the following subsections, we consider other theoretical models of the hippocampus that can usefully be compared with the present model, including a number of widely-cited theories that postulate some form of error-signaling or error-driven learning. At the heart of many of these models is the idea that the hippocampus can generate predictions in order to then compute a novelty or error signal relative to such predictions, or to learn and predict sequences of future states. After briefly summarizing these models, we discuss what roles the hippocampus and the neocortex play in generating predictions according to the complementary learning systems (CLS) framework in which the current model is based (McClelland et al., 1995; O'Reilly et al., 2014; O'Reilly, Ranganath, & Russin, 2021).

### *Prediction-based Models of the Hippocampus*

One longstanding and influential set of theories suggests that the hippocampus acts as a *comparator*, generating predictions in order to detect and signal *novelty* or *surprise* (Gray, 1982; Lisman & Grace, 2005; Vinogradova, 2001). Specifically, the hippocampus in these models generates a global *scalar* signal as a function of the relative mismatch between a predicted state and the actual next state. For example Gray (1982) proposed that combining previous sensory information and the motor plan creates predictions about the current state, which are then compared with the actual current sensory information. The motor plan is maintained if the two states match, but it is interrupted in the case of incorrect predictions (i.e., surprise or novelty), so that the animal can attempt to solve the problem in a different way. In the Lisman and Grace (2005) model, the hippocampal novelty or surprise signal is hypothesized to drive phasic dopamine firing via its subcortical projection pathway through the subiculum. These different models vary in terms of the exact mechanisms and subfields proposed to compute the mismatch signal (CA3, CA1 or subiculum), but

they assume a similar overall functional role for the hippocampus in terms of synthesizing predictions.

In the present model, error signals are not summed, but rather used to optimize learning of specific associations. However, it is possible that the temporal-difference error signals present in our hippocampal model could play a role in generating a global novelty signal. For example, at different points in the theta phase cycle (Figure 1), area CA1 and E<sub>Cout</sub> are representing the current information as encoded in CA3 and its projections into CA1, versus the bottom-up cortical state present in E<sub>Cin</sub>. The difference between these two activation states could be converted into a global mismatch signal that would reflect the relative novelty of the current state compared to prior episodic memory learning in the CA3 of the hippocampus. Likewise, it is possible that a similar global error signal could be computed from the temporal differences over CA3 in our model, reflecting the extent to which CA3 has learned to encode the more pattern-separated DG-driven pattern, which is likely to also reflect the relative novelty of the current input state. We will investigate these possibilities in future research.

Prediction-based learning in the hippocampus is also central to another early computational model, which is based on error-driven backpropagation learning in the context of a predictive autoencoder (Myers & Gluck, 1995). In this model, the hippocampal network learns by simultaneously attempting to recreate the current input patterns, and also predict future reinforcement outcomes. The cortical network representations are then shaped by hippocampal training signals, similar in spirit to the scalar novelty / surprise signals. Simulations with this model and its hippocampus-lesioned variant have been shown to replicate a wide range of conditioned behaviors in rats and rabbits (Gluck & Myers, 1994), although it is notable that many of these same phenomena can also be accounted for using an earlier version of the episodic memory model presented here (O'Reilly & Rudy, 2001).

Another class of models hypothesizes that the hippocampus learns *sequences* of events over time, such that, when a past state is encountered, the hippocampus can enable the prediction of potential outcomes of actions taken in novel situations, based on what has happened previously (Jensen & Lisman, 1996; Levy, 1996; Rolls, 2013; Schapiro et al., 2017; Stachenfeld, Botvinick, & Gershman, 2017; Tsodyks, Skaggs, Sejnowski, & McNaughton, 1996; Wallenstein & Hasselmo, 1997). In some of these models, the recurrent connections in area CA3 learn to associate prior time step representations with subsequent time step patterns, thus learning to predict the next state based on the current state. Other models suggest that the hippocampus learns systematic predictive representations (e.g., a successor map of subsequent states following from the current state in the case of Stachenfeld et al., 2017). Most of the models suggest that the hippocampus itself is capable of synthesizing novel predictions based on these learned sequences.

### *Neural Mechanisms of Prediction in Hippocampus and Cortex*

The models discussed above emphasize the idea that memory retrieval in the hippocampus is a form of prediction, and at a broader level, many researchers have embraced the idea that the hippocampus might be specialized for generating predictions in the service of navigation, reasoning, and imagination (Davachi & DuBrow, 2015; Kok & Turk-Browne, 2018; Lisman & Redish, 2009; Mack, Love, & Preston, 2018; Mizumori, 2013; Zeithamova, Schlichting, & Preston, 2012; Zeithamova et al., 2012). These theories, however, tend to describe prediction in broad strokes, and as such, we argue that they do not respect the computational limitations of the hippocampus.

In contrast to the above models, we do not believe that the hippocampus itself is well-suited for generating predictions in novel situations, and instead we think the relevant data can be better captured in terms of the simple episodic memory framework that the Hebb-Marr model embodies (as updated in the present paper). Here, the hippocampus is specialized for rapidly encoding memories of distinct events or episodes using highly pattern-separated representations, which can later be recalled through the process of pattern completion. Given the overwhelming empirical support for the idea that the hippocampus is specialized for rapidly learning new episodic memories, we believe that it also cannot support a semantic prediction system

capable of generating systematic predictions in novel situations.

Specifically, generating a novel prediction typically requires a cognitive process to synthesize prior experience and general principles (e.g., a scientific theory, or implicitly-learned regularities of the world, such as intuitive physics) to specify what will happen in the future. This kind of systematic generalization from prior experience to novel situations is precisely what the neocortex is thought to be optimized for according to the CLS theory (McClelland et al., 1995; O'Reilly et al., 2014; O'Reilly, Ranganath, & Russin, 2021). This is because the overlapping representations of cortical networks are optimized to slowly integrate statistical regularities across many different experiences to learn *semantic* representations capable of supporting systematic generalization in novel situations. Indeed there are various models of error-driven predictive learning in the neocortex capable of learning such systematic predictive abilities, including a biologically-detailed proposal based on thalamocortical loops (O'Reilly, Russin, et al., 2021).

Although the computational architecture of the hippocampus is not well-suited for generating predictions on its own, it can certainly provide relevant episodic memories as input to the cortical prediction generation process. For example, strategic recall of particular memories, followed by appropriate updating of the details to better match the current circumstances, could produce a more generative predictive system that can synthesize novel predictions for new situations. These kinds of complex interactions, however, go well beyond the capabilities of the hippocampal circuit by itself, as captured in any implemented computational model.

### *Nonmonotonic Plasticity vs. Error-driven Learning*

The temporal difference error signals that drive learning in our model can be related to the neural activation signals that drive nonmonotonic plasticity (NMP) learning dynamics as explored by Norman and colleagues (Ritvo, Turk-Browne, & Norman, 2019). Specifically, the nonmonotonic plasticity function drives LTD when activations are at a middling, above-zero level, while LTP occurs for more strongly activated neurons. This is the same underlying learning function that we use in our error-driven learning model (O'Reilly et al., 2012), and thus it can be difficult to strongly distinguish the predictions of these two models. In particular, the conditions under which errors drive LTD in our model can be construed as being within the LTD range of the nonmonotonic plasticity function, under various additional assumptions. However, the NMP models have not been implemented within the context of a full hippocampal circuit, and it is unclear how those models might actually perform in specific conditions. Thus, the difficulties are more at the level of abstract principles rather than detailed model comparisons at this point.

### *Novel Predictions*

There are several novel, testable predictions from our model that can distinguish it from a more standard Hebbian-based model:

- As shown in Figure 5, the error-driven learning in area CA3 serves to drive pattern separation over time among otherwise similar representations, whereas the Hebbian version of the model showed increasing patterns similarity over learning. Thus, experiments that track the progression of representational similarity over the course of learning could distinguish these two patterns.
- Our model depends critically on the modulation of different pathways of connectivity within the hippocampus, organized according to the theta cycle in rodents according to Hasselmo et al. (2002), creating the temporal differences that drive learning. Thus, neural manipulations that selectively disrupt the theta cycle and / or these pathway-specific modulations should affect error-driven learning, but may not affect recall of previously-learned information to the same extent. By contrast, it is not clear why from the purely Hebbian learning framework that disrupting the theta cycle should impair



that form of learning. Intriguingly, a recent report appears to be consistent with the predictions of our model: Quirk et al. (2021) found that a highly selective disruption of the timing of the theta cycle produced selective deficits in learning. Even more selective tests of the specific temporal dynamics associated with the CA3 and CA1 could more specifically test our model.

- Our model also generates novel predictions about the functional characteristics of human memory. For instance, there is a large body of evidence about the *testing effect*, in which items that are tested with partial information (as compared to restudy of the complete original information) are better retained than items that are re-studied (Liu et al., 2021). The superiority of testing over restudy presents a challenge to models depending on Hebbian learning because learning a precise input pattern should be as good or better than learning from a partial cue. Theremin, however, provides a natural explanation for the testing effect, as the difference between an initial guess and subsequent correct answer provides an opportunity for error-driven learning, whereas restudy provides no opportunity to make the initial guess needed in order to optimize weights.

## Conclusions

In summary, results from our simulations show that error-driven learning mechanisms can dramatically improve both memory capacity and learning speed by reducing competition between learned representations. Furthermore, these mechanisms can potentially explain a wide range of learning and memory phenomena. Error-driven learning in CA3 can emerge naturally out of neurophysiological properties of the hippocampal circuits, building on the basic framework for error-driven learning in the monosynaptic EC  $\leftrightarrow$  CA1 pathway (Ketz et al., 2013). There are many further implications and applications of this work, and many important empirical tests needed to more fully establish its validity. Hopefully, the results presented here provide sufficient motivation to undertake this important future research.

## Appendix

In this appendix we provide some of the key parameters to understanding how the Theremin model is structured, and organized to do the AB-AC memory task. See table captions for detailed descriptions on parameters/diagrams. The best documentation for those interested in all the details is the fully-functioning Theremin model along with further detailed documentation, available at: <https://github.com/ccnlab/hip-edl>.

### Network Size Parameters

Parameter \ Network Size	Small	Medium	Large
Input Pool Size	7x7	7x7	7x7
Input Number of Pools	2x3	2x3	2x3
ECin Pool Size	7x7	7x7	7x7
ECin Number of Pools	2x3	2x3	2x3
ECout Pool Size	7x7	7x7	7x7
ECout Number of Pools	2x3	2x3	2x3
DG Size	44x44	67x67	89x89
CA3 Size	20x20	30x30	40x40
CA1 Pool Size	10x10	15x15	20x20
CA1 Number of Pools	2x3	2x3	2x3

Table 1: Parameters for network sizes. In neural networks, larger network size usually leads to higher capacity, when controlled for other settings. In the current study, we tested different variations of the hippocampus model for three different network sizes to show the benefit of error-driven learning for hippocampus regardless of sizes, meaning the mechanism is generalizable. For pool sizes, the numbers in the table refer to number of neurons in that specific pool. Note: DG size is around five times CA3 size as specified in our previous model (Ketz et al., 2013)

### Training Input Diagram

Context B3/C3	Context B4/C4
Context B1/C1	Context B2/C2
A	B/C

Table 2: Training phase patterns for network input. Each unit here represents a pool in the Input layer. Memory of AB and AC pairs are categorized into different experiences, with four different context pools for each experience.

### Testing Input Diagram

Context B3/C3	Context B4/C4
Context B1/C1	Context B2/C2
A	empty

Table 3: Testing phase patterns for network input. Each unit here represents a pool in the Input layer. Memory of AB and AC pairs are categorized into different experiences, with four different context pools for each experience. During testing, no input is given for the B/C pool, while A and context pools are used as partial cues for the hippocampus to do pattern completion (i.e., complete the pattern for the B/C pool).

\*

### References

Abu-Mostafa, Y., & St. Jacques, J. (1985, July). Information capacity of the Hopfield model. *IEEE Transactions on Information Theory*, 31(4), 461–464. doi: 10.1109/TIT.1985.1057069

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985, December). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1), 147–169.
- Barnes, J. M., & Underwood, B. J. (1959, August). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58, 97–105. doi: 10.1037/h0047507
- Bernier, B. E., Lacagnina, A. F., Ayoub, A., Shue, F., Zemelman, B. V., Krasne, F. B., & Drew, M. R. (2017, June). Dentate Gyrus Contributes to Retrieval as well as Encoding: Evidence from Context Fear Conditioning, Recall, and Extinction. *The Journal of Neuroscience*, 37(26), 6359–6371. doi: 10.1523/JNEUROSCI.3029-16.2017
- Davachi, L., & DuBrow, S. (2015, February). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, 19(2), 92–99. doi: 10.1016/j.tics.2014.12.004
- Do, V. H., Martinez, C. O., Martinez, J. L., & Derrick, B. E. (2002, February). Long-Term Potentiation in Direct Perforant Path Projections to the Hippocampal CA3 Region In Vivo. *Journal of Neurophysiology*, 87(2), 669–678. doi: 10.1152/jn.00938.2000
- Eichenbaum, H. (2016, September). Still searching for the engram. *Learning & Behavior*, 44(3), 209–222. doi: 10.3758/s13420-016-0218-1
- Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007, July). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, 30(1), 123–152.
- Gluck, M. A., & Myers, C. E. (1994, January). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3, 491–516. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8269040>
- Gray, J. A. (1982). *The neuropsychology of anxiety: An inquiry into the functions of the septo-hippocampal systems*. Oxford, England: Oxford University Press.
- Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002, April). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, 14(4), 793–818. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11936962>
- Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.
- Jensen, O., & Lisman, J. E. (1996, January). Hippocampal CA3 region predicts memory sequences: Accounting for the phase precession of place cells. *Learning & Memory*, 3, 279–287.
- Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013, June). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, 9, e1003067. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23762019>
- Kitamura, T., Sun, C., Martin, J., Kitch, L. J., Schnitzer, M. J., & Tonegawa, S. (2015, September). Entorhinal Cortical Ocean Cells Encode Specific Contexts and Drive Context-Specific Fear Memory. *Neuron*, 87(6), 1317–1331. doi: 10.1016/j.neuron.2015.08.036
- Kobayashi, K., & Poo, M.-m. (2004, February). Spike Train Timing-Dependent Associative Modification of Hippocampal CA3 Recurrent Synapses by Mossy Fibers. *Neuron*, 41(3), 445–454. doi: 10.1016/S0896-6273(03)00873-0
- Kok, P., & Turk-Browne, N. B. (2018, August). Associative prediction of visual shape in the hippocampus. *Journal of Neuroscience*, 38(31), 6888–6899. doi: 10.1523/JNEUROSCI.0163-18.2018
- Kowadlo, G., Ahmed, A., & Rawlinson, D. (2020). Unsupervised One-Shot Learning of Both Specific Instances and Generalised Classes with a Hippocampal Architecture. In M. Gallagher, N. Moustafa, & E. Lakshika (Eds.), *AI 2020: Advances in Artificial Intelligence* (pp. 395–406). Cham: Springer International Publishing. doi: 10.1007/978-3-030-64984-5\_31
- Kumaran, D., & McClelland, J. L. (2012, July). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, 119(3), 573–616. doi: 10.1037/a0028681
- LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, 521(7553), 436–444. doi:

10.1038/nature14539

- Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, 6(6), 579–590. doi: 10.1002/(SICI)1098-1063(1996)6:6<579::AID-HIPO3>3.0.CO;2-C
- Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020, June). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6), 335–346. doi: 10.1038/s41583-020-0277-3
- Lisman, J. E., & Grace, A. A. (2005, May). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron*, 46(5), 703–713. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15924857>
- Lisman, J. E., & Redish, A. D. (2009, May). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1193–1201. doi: 10.1098/rstb.2008.0316
- Liu, X., O'Reilly, R. C., & Ranganath, C. (2021). Effects of retrieval practice on tested and untested information: Cortico-hippocampal interactions and error-driving learning. In *The Context of Cognition: Emerging Perspectives*. Elsevier.
- Mack, M. L., Love, B. C., & Preston, A. R. (2018, July). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, 680, 31–38. doi: 10.1016/j.neulet.2017.07.061
- Marr, D. (1971, July). Simple Memory: A Theory for Archicortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 262(841), 23–81. doi: 10.1098/rstb.1971.0078
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995, August). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, 102(3), 419–457. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7624455>
- McCloskey, M., & Cohen, N. J. (1989, January). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 24 (pp. 109–164). San Diego, CA: Academic Press.
- McHugh, T. J., Jones, M. W., Quinn, J. J., Balthasar, N., Coppari, R., Elmquist, J. K., ... Tonegawa, S. (2007, July). Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network. *Science*, 317(5834), 94–99. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17556551>
- McMahon, D. B., & Barrionuevo, G. (2002, July). Short- and Long-Term Plasticity of the Perforant Path Synapse in Hippocampal Area CA3. *Journal of Neurophysiology*, 88(1), 528–533. doi: 10.1152/jn.2002.88.1.528
- McNaughton, B. L., & Nadel, L. (1990, January). Hebb-Marr Networks and the Neurobiological Representation of Action in Space. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and Connectionist Theory* (pp. 1–63). Hillsdale, NJ: Erlbaum.
- Milner, B., Squire, L. R., & Kandel, E. R. (1998, January). Cognitive Neuroscience and the Study of Memory. *Neuron*, 20, 445.
- Mizumori, S. (2013). Context Prediction Analysis and Episodic Memory. *Frontiers in Behavioral Neuroscience*, 7, 132. doi: 10.3389/fnbeh.2013.00132
- Myers, C. E., & Gluck, M. A. (1995, February). Context, conditioning, and hippocampal rerepresentation in animal learning. *Behavioral neuroscience*, 108, 835–847. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7826508>
- Nakashiba, T., Cushman, J. D., Pelkey, K. A., Renaudineau, S., Buhl, D. L., McHugh, T. J., ... Tonegawa, S. (2012, March). Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. *Cell*, 149. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22365813>

- Norman, K. A., & O'Reilly, R. C. (2003, November). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14599236>
- O'Reilly, R. C. (1996, January). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8(5), 895–938. doi: 10.1162/neco.1996.8.5.895
- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014, December). Complementary Learning Systems. *Cognitive Science*, 38(6), 1229–1248. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22141588>
- O'Reilly, R. C., & McClelland, J. L. (1994, January). Hippocampal Conjunctive Encoding, Storage, and Recall: Avoiding a Tradeoff. *Hippocampus*, 4(6), 661–682.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: <http://ccnbook.colorado.edu>. Retrieved from <http://ccnbook.colorado.edu>
- O'Reilly, R. C., Ranganath, C., & Russin, J. L. (2021, August). The Structure of Systematicity in the Brain. *arXiv:2108.03387 [q-bio]*. Retrieved from <http://arxiv.org/abs/2108.03387>
- O'Reilly, R. C., & Rudy, J. W. (2001, May). Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function. *Psychological Review*, 108(2), 311–345. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11381832>
- O'Reilly, R. C., Russin, J. L., Zolfaghar, M., & Rohrlach, J. (2021). Deep predictive learning in neocortex and pulvinar. *Journal of Cognitive Neuroscience*, 33(6), 1158–1196. doi: 10.1162/jocn.a\_01708
- Pignatelli, M., Ryan, T. J., Roy, D. S., Lovett, C., Smith, L. M., Muralidhar, S., & Tonegawa, S. (2019, January). Engram Cell Excitability State Determines the Efficacy of Memory Retrieval. *Neuron*, 101(2), 274–284.e5. doi: 10.1016/j.neuron.2018.11.029
- Quirk, C. R., Zutshi, I., Srikanth, S., Fu, M. L., Devico Marciano, N., Wright, M. K., ... Leutgeb, S. (2021, October). Precisely timed theta oscillations are selectively required during the encoding phase of memory. *Nature Neuroscience*, 1–14. doi: 10.1038/s41593-021-00919-0
- Rebola, N., Carta, M., & Mulle, C. (2017, April). Operation and plasticity of hippocampal CA3 circuits: Implications for memory encoding. *Nature Reviews Neuroscience*, 18(4), 208–220. doi: 10.1038/nrn.2017.10
- Rescorla, R. A., & Wagner, A. R. (1972, January). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Theory and Research* (pp. 64–99). New York: Appleton-Century-Crofts.
- Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019, September). Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends in Cognitive Sciences*, 23(9), 726–742. doi: 10.1016/j.tics.2019.06.007
- Rolls, E. (2013). A quantitative theory of the functions of the hippocampal CA3 network in memory. *Frontiers in Cellular Neuroscience*, 7, 98. doi: 10.3389/fncel.2013.00098
- Ruediger, S., Vittori, C., Bednarek, E., Genoud, C., Strata, P., Sacchetti, B., & Caroni, P. (2011, May). Learning-related feedforward inhibitory connectivity growth required for memory precision. *Nature, advance online publication*. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/21532590>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, January). Learning representations by back-propagating errors. *Nature*, 323(9), 533–536.
- Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017, January). Complementary learning systems within the hippocampus: A neural network modelling approach to recon-

- ciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B*, 372(1711), 20160049. doi: 10.1098/rstb.2016.0049
- Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016, January). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, 26(1), 3–8. doi: 10.1002/hipo.22523
- Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017, November). The hippocampus as a predictive map. *Nature Neuroscience*, 20(11), 1643–1653. doi: 10.1038/nn.4650
- Treves, A., & Rolls, E. T. (1991, January). What Determines the Capacity of Autoassociative Memories in the Brain. *Network: Computation in Neural Systems*, 2, 371–397.
- Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., & McNaughton, B. L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: A spiking neuron model. *Hippocampus*, 6(3), 271–280. doi: 10.1002/(SICI)1098-1063(1996)6:3<271::AID-HIPO5>3.0.CO;2-Q
- Tsukamoto, M., Yasui, T., Yamada, M. K., Nishiyama, N., Matsuki, N., & Ikegaya, Y. (2003, February). Mossy fibre synaptic NMDA receptors trigger non-hebbian long-term potentiation at entorhino-CA3 synapses in the rat. *The Journal of Physiology*, 546(3), 665–675. doi: 10.1113/jphysiol.2002.033803
- Vinogradova, O. S. (2001, December). Hippocampus as comparator: Role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus*, 11(5), 578–598. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11732710>
- Wallenstein, G. V., & Hasselmo, M. E. (1997, December). GABAergic modulation of hippocampal population activity: Sequence learning, place field development, and the phase precession effect. *Journal of neurophysiology*, 78, 393. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9242288>
- Whittington, J. C. R., & Bogacz, R. (2019, March). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 23(3), 235–250. doi: 10.1016/j.tics.2018.12.005
- Widrow, B., & Hoff, M. E. (1960). Adaptive Switching Circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4* (pp. 96–104).
- Witter, M. P., Doan, T. P., Jacobsen, B., Nilssen, E. S., & Ohara, S. (2017). Architecture of the entorhinal cortex A review of entorhinal anatomy in rodents with some comparative notes. *Frontiers in Systems Neuroscience*, 11. doi: 10.3389/fnsys.2017.00046
- Yassa, M. A., & Stark, C. E. L. (2011, October). Pattern separation in the hippocampus. *Trends in Neurosciences*, 34(10), 515–525. doi: 10.1016/j.tins.2011.06.006
- Yeckel, M. F., & Berger, T. W. (1990, September). Feedforward excitation of the hippocampus by afferents from the entorhinal cortex: Redefinition of the role of the trisynaptic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 87, 5832–5836. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2377621>
- Yonelinas, A. P., Ranganath, C., Ekstrom, A. D., & Wiltgen, B. J. (2019, June). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*, 20(6), 364–375. doi: 10.1038/s41583-019-0150-4
- Zeithamova, D., Schlichting, M., & Preston, A. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience*, 6, 70. doi: 10.3389/fnhum.2012.00070