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

July 4, 2021

Abstract:

Now that's abstract..

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 large numbers of largely silent dentate gyrus (DG) granule cells, will drive the creation of cell assemblies that involve a distinct 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). Overall, the basic tenets established by Hebb and Marr account for a vast amount of behavioral and neural data on hippocampal function, and is as close to a "settled theory" as any in neuroscience (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, but

---

Draft Manuscript: Do not cite or quote without permission.

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 to CA3 pathway is not sufficient to make up for the interference caused by reliance on Hebbian learning.

According to this analysis, the logical solution 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):

$$dW = x(r - y), \tag{1}$$

where dW is the amount of synaptic weight change, x is the sending neuron activity level (e.g., average rate of firing 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:

$$y = \sum xW. \tag{2}$$

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 (LeCun, Bengio, & Hinton, 2015).

To be able to apply a similar self-limiting error-driven learning logic to the hippocampus, we need some kind of target signal that determines when the learning has accomplished what it needs to do, to achieve the overall memory functionality required. What kind of natural such target signal is there for the hippocampus? Hippocampal episodic encoding is widely regarded as automatically and continuously storing new memories rapidly as they occur in real time, which seems more like an automatic, Hebbian form of learning as compared to something driven toward something like a target signal. Nevertheless, we have previously shown that error-driven learning principles can be applied to the CA1 region of the hippocampus (Ketz, Morkonda, & O'Reilly, 2013), building on theta-phase dynamics discovered by (Hasselmo, Bodelon, & Wyble, 2002). These error-driven learning dynamics have been supported by studies of CA1 learning Schapiro, Turk-Browne, Botvinick, and Norman (2017); Schapiro, Turk-Browne, Norman, and Botvinick (2016).

Building upon this prior work, and a recent proposal by Kowadlo, Ahmed, and Rawlinson (2020), we extend the application of error-driven learning to the "heart" of the hippocampus: learning in the feedforward and recurrent synapses of area CA3. The key idea is that the dentate gyrus (DG) can serve as a teacher to
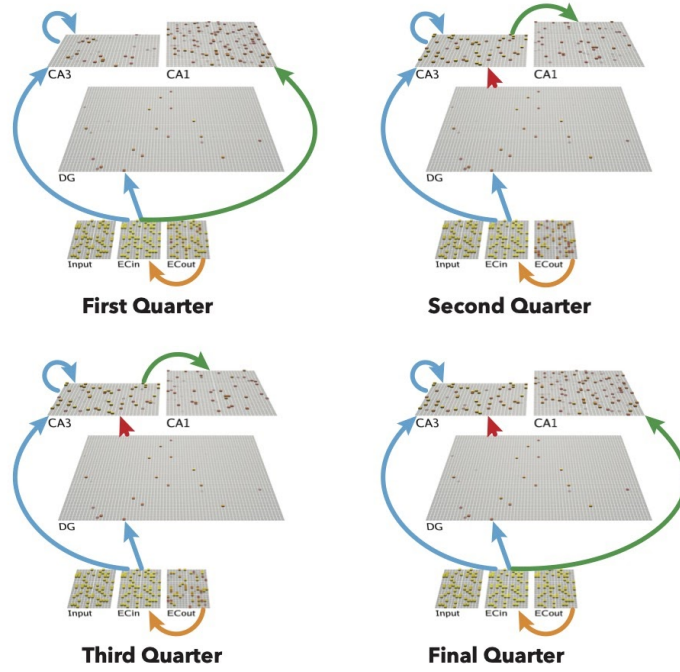
Figure 1: Architecture of the Theremin model. Visual depiction of one full alpha-cycle training in Theremin, separated by four different sequential time points (i.e., four quarters). Arrows depict theoretically important weight modifications in that quarter; on the contrary, connections not implied by arrows did not change their weights. Blue and red arrows indicate the new CA3 dynamics in CA3, with red arrows sourcing into CA3 after the first quarter and create temporal difference error signals. Green arrows indicate the theta-phase (Ketz et al, 2013) dynamics in CA1. Orange arrows indicate big-loop signals.

the CA3, driving error-driven 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 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 have broad implications for understanding important learning phenomena such as the testing effect. 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.

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 (Ketz et al., 2013). 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; Ketz et al., 2013). 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 extension

of the neocortex, where the principles of error-driven learning have been well-developed (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:

$$dW = \mathrm{CA1}(\mathrm{ECin} - \mathrm{ECout}), \tag{3}$$

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 2). 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 ECout activation between their CA1-driven vs. ECin-driven states produces the (ECin - ECout) difference in the above equation. Likewise, CA1 neurons directly experience a "reflection" of this temporal difference, conveyed via the projections from ECout $\rightarrow$ 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 (O'Reilly, 1996; Whittington & Bogacz, 2019), which is more properly the version of error-driven learning powering modern deep networks. 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).

In this paper, we extend error-driven learning principles to the remainder of the hippocampal system, showing how the main trisynaptic pathway involving the EC, DG, and CA3 can support error-driven learning, resulting in dramatic improvements in overall learning capacity. As a result of an extensive parameter search with these new mechanisms, we discovered a few other important principles that enhance episodic learning performance as well.

The primary challenge in extending error-driven learning to the trisynaptic pathway is in deriving an error signal that drives learning toward effective memory encoding. As captured in the delta rule equation, the essential ingredient is to have two different states, one that represents a better, target representation (e.g., the actual reward, or the actual ECin input in the examples considered previously) that can be compared against what the current synaptic weights produce on their own, so that these weights can be adjusted to better achieve the target representation. In this framing, one can consider the highly pattern-separated activity pattern in the DG, to provide a kind of target representation relative to what the CA3 can produce on its own. Thus, in effect, the DG 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:

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

where *ECin* is the sending activity into CA3 (via the perforant pathway (PP) projections), $\mathrm{CA3}_{dg}$ is the

activity of the CA3 neurons when being driven by the strong mossy fiber inputs from the DG, and CA3$_{nodg}$ is the CA3 activity when not being driven by those DG inputs, i.e., from only the perforant-path and CA3 recurrent inputs. In this way, the DG, which is the sparsest and most pattern-separated hippocampal layer, is serving as an error-driven teacher to CA3, teaching it to adopt its highly pattern-separated view of the input pattern. Critically, to the extent that CA3 is already matching the DG pattern separation, 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 $\rightarrow$ 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 may 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 activity signal through the DG to the CA3, compared to the more direct perforant pathway from ECin to CA3. Thus, as in our prior models, the minus phase term in the delta rule occurs first (i.e., 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. TODO: add specific numbers??

- Mossy fiber (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 it has historically been difficult to record from GCs, recent data from mice supports the idea that they drive significant levels of pattern separation. For example, in one study, 50% of GCs received spatially tuned inputs, but only 5% showed spatially turned spiking (Zhang, Schlögl, & Jonas, 2020), thereby producing sparse and specific (i.e., pattern separated) patterns.

- 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 PP-CA3 loop 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 showing that MF projections are not necessary in naturally recalling a memory (Bernier et al., 2017; Nakashiba et al., 2012). 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). Moreover, stimulating correct DG GCs assemblies alone can result in successful memory retrieval in CA3 (**?**). 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. 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. In effect, the DG serves as a teacher to the CA3, so what is left over to teach the teacher itself? 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 (colorfully capturing the concept of Total Hippocampal ERror MINimization), 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 entorhinal cortex (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 Dentate Gyrus (DG) and CA3 via broad, diffuse perforant pathway (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 CA3 receives strong mossy fiber (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 RR, 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.

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 → CA3 until the 2nd theta phase cycle. 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, 1960; 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*).
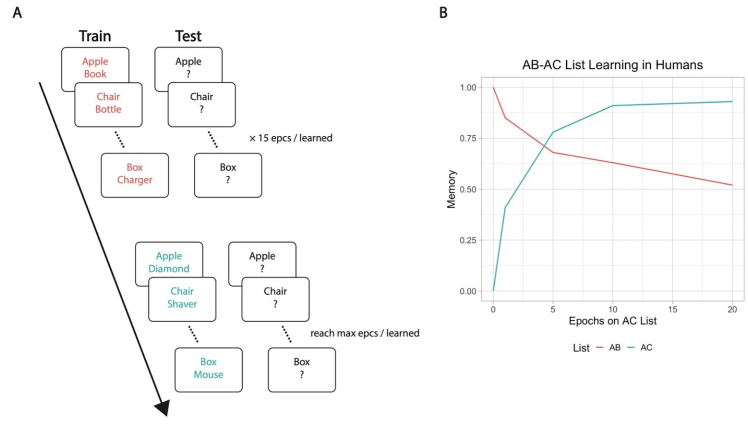
Figure 2: ABAC list learning paradigm diagram and human data reproduced from an empirical experiment (Barnes & Underwood, 1960). A) Depicted in the diagram is the ABAC list learning task used in the current study. The first "learned" means AB memory reaches 100%; the second "learned" means AC memory reaches 100%. The maximum number of epochs in the task is 30. Detailed procedure is described in Materials & Method. B) Human participants showed moderate interference of the AB list after learning the AC list. Note that there were 20 epochs following AC list learning in the original human experiment, unlike the current task design.

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 occured, 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 1, or 30 epochs have occured in total, that run was considered complete. We ran 30 different simulated "subjects" 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-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) to determine the relationship between size and capacity in each case. The list sizes
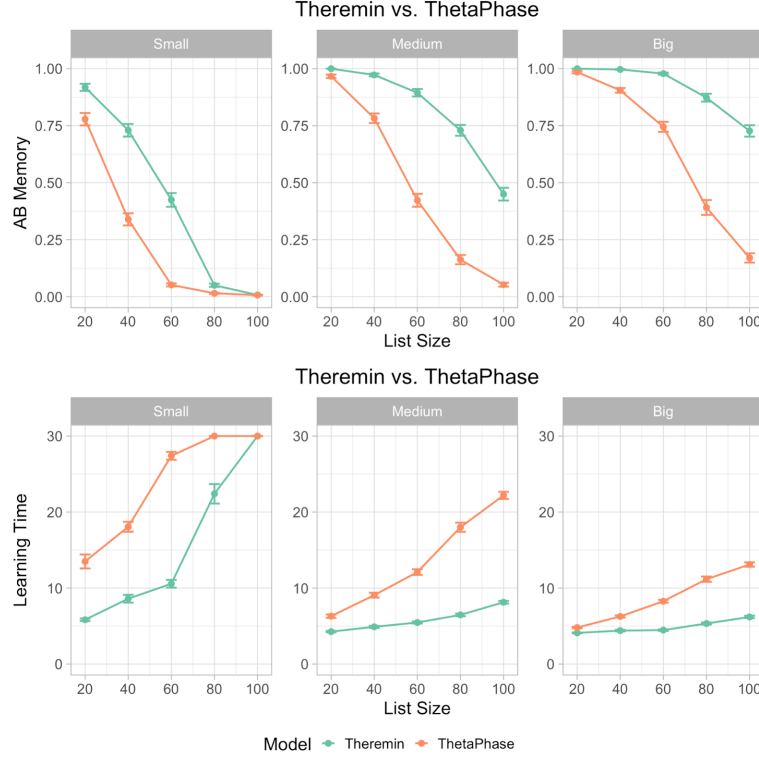
Figure 3: Theremin vs. ThetaPhase on AB memory and learning time.

ranged from 20 to 100 pairs of AB/AC associations (a typical human experiment uses only XX items – **TODO lookup number!**). It is worth noting that the number of units in DG is roughly 5 times to that in CA3, consistent with the theta-phase hippocampus model. Biological evidence for CA3:CA1 ratio?

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 more detailed analyses of the dynamical behavior of the different network layers over the course of processing individual inputs, and across iterations of learning. This enabled us to directly measure the temporal difference error signals that are driving learning in Theremin, and to see how these adapted 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. Figure 3 shows the results across all three network sizes and numbers of list items. For all three network sizes, the full Theremin model completed learning significantly
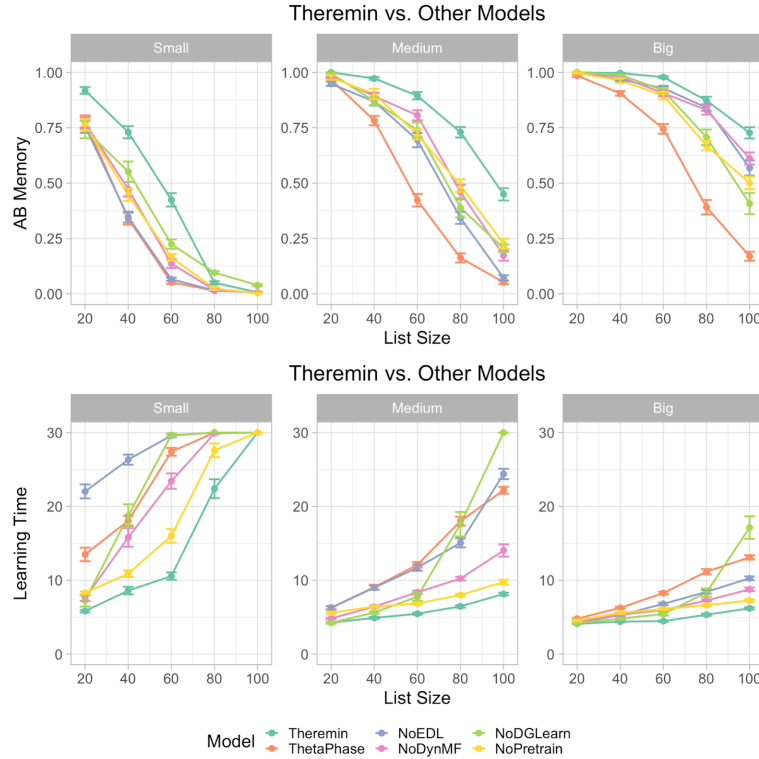
Figure 4: Theremin vs. other Models on AB memory and learning time.

faster (i.e., the NEpochs measure) than the theta-phase hippocampus model across all list sizes (ps ¡ .01, for detailed results for statistical analyses, see Supplementary material). Moreover, the ABmem results show that Theremin was better at counteracting forgetting 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)).

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 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 BigHip List20 (p = 0.321)) and learned faster (ps ¡ .01, except BigHip List20 (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 BigHip List20 (p = .013), ABMem ps ¡ .01, except BigHip List20 and BigHip List 80 (p = 0.127)..

To determine the importance of learning in ECin → 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 perforant

Figure 5: Testing statistics across epochs (List 100, Medium net). The number of epochs used in Theremin training was set to a fixed number (i.e., 10) that sufficed complete learning of AB and AC lists, while in NoEDL was set to the maximum amount used in the current paper. Learning of AB and AC lists were separated into two different halves.

pathway 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 NoStrongLTD 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.

Looking into parameters in ECin → DG pathway revealed that eliminating the higher rate of weight decreases (LTD = long-term depression in biological terms) significant 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 → DG pathway improved the ABMem score (reducing interference), but resulted in slower learning in the NEpoch measure, and vice-versa for higher learning rates. This tradeoff between resistance to interference and faster learning could potentially be something under neuromodulatory control, as neuromodulators such as dopamine are known to affect learning rates. The 2nd principle

footnote: 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. Thus we compared two models 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 ↔ 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 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,
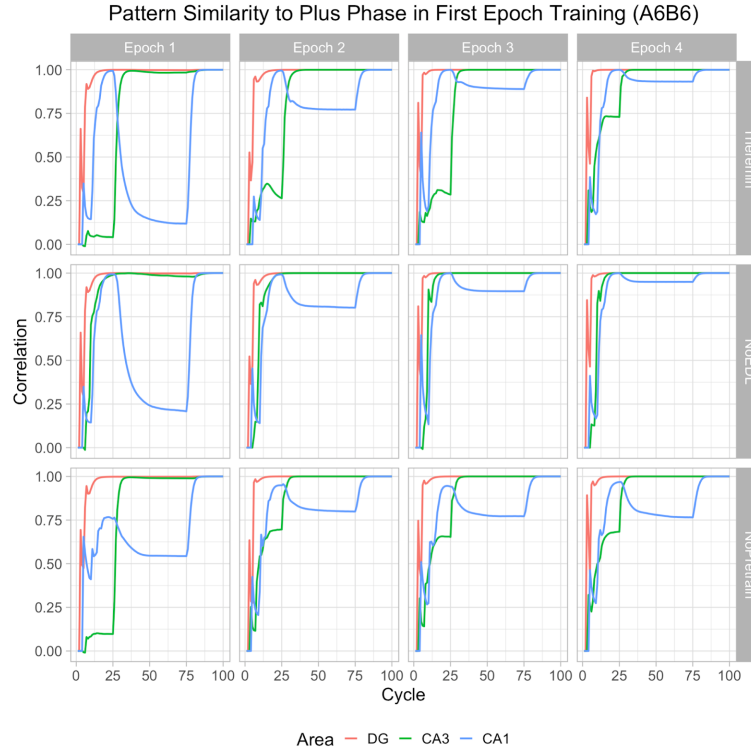
Figure 6: Hippocampal subregions' pattern similarities across a full alpha-cycle in different models for an example AB pair. Drops seen within the first few cycles were due to the settling of temporally different patterns and were not of interest to the current paper. The error-driven learning dynamics in CA3 and CA1 were captured in these plots, as well as the big-loop signals from ECout to ECin. Detailed interpretations of the figure are in the main text.

and 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 small 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 each layer 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 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 first quarter (cycle 1-25), representational similarities in each layer (except CA3 in Theremin and CA1 in NoPretrain because they

are not driven by their final, settling signals) increases due to strong ECin $\rightarrow$ CA1 learning described in the theta-phase model. In the second and third quarter (cycle 26-75), CA3 starts to source into CA1, then ECout and ECin, creating a big-loop error signal that finally reaches CA3, deviating its pattern from the stabilized one, as depicted by the slightly curved green line in the first epoch. In particular, DG also starts to impact CA3 from the second quarter in the Theremin model, and their strong connections drive the CA3 pattern to settle into a DG-dominated one. In the final quarter (cycle 76-100), ECin $\rightarrow$ CA1 opens again and drives the CA1 pattern to settle into its final state.

It is worth noting that in the Theremin model, as learning progresses, the CA3 pattern in the first quarter becomes increasingly similar to its final pattern (Figure 7). In effect, this similarity signal acts as a prediction to how close CA3 pattern is to its final DG-corrected pattern, before DG starts to have an effect on CA3. Since in the first epoch, the prediction is driven only by ECin $\rightarrow$ CA3 and CA3 $\rightarrow$ CA3 inputs, a huge temporal difference (error signal) is created to modify these connections (i.e., heterosynaptic plasticity). This error becomes smaller fast and training is stopped immediately when there is no error anymore. On the other hand, the NoEDL model increases synaptic weights between CA3 and other regions whenever two neurons are active together, taking a longer time to finish learning but does not effectively increase retrieval performance. Similarly, CA1 similarity during the second and third quarters also increases across epochs, meaning the CA3 $\rightarrow$ CA1 connections are also learning to predict the final CA1 pattern driven by only ECin.

Discussion

By incorporating error-driven learning mechanisms at the critical CA3 synapses in our computational model of the hippocampus, along with a few other important optimizations, we have been able to significantly improve learning rate 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.

In summary, consistent with these computational principles, we argue that we should break with the long historical tradition of Hebbian learning in the hippocampus, and begin exploring the implications of error-driven learning in this system. For example, Schapiro and colleagues have recently discovered that the error-driven learning dynamics of the theta-phase based monosynaptic EC – CA1 pathway support a form of statistical learning, that was previously thought to only exist in more purely cortical areas (cites). Likewise, adopting an error-driven learning framework for the trisynaptic pathway opens up further opportunities for powerful error-driven learning phenomena in this pathway, and the hippocampal system as a whole.

For example, we have recently been able to use the error-driven learning mechanisms of this hippocampal model to account for intriguing data on the ongoing learning repercussions that emerge via an interplay of hippocampal and cortical representations and learning dynamics (Xiaonan model). Specifically, we found that hippocampal context-sensitive episodic encoding can shape subsequent "testing effect" cortical learning to increase the likelihood of recalling items that were seen within the same temporal context, while decreasing that of items seen within different temporal contexts. Over the course of a 24hr period including sleep, however, cortical semantic associations were able to penetrate across this temporal context barrier and entrain a positive association for semantically-related items. In our model of this case, the cortical component of the model trained the hippocampal system via error-driven learning, while the opposite was the case for the earlier testing-effect learning.

More generally, error-driven learning produces similar effects as those associated with the non-monotonic learning dynamics explored by Norman and colleagues (cites). Indeed, these models share the same fundamental temporal-difference driven learning rule, and mainly differ in the sources of the temporal

differences that drive it.

The idea that the cortex learns through a process of predictive error-driven learning has long been appealing for many reasons (cites). It may be possible to extend the above mechanisms to include a predictive element as well, such that hippocampal episodic learning can be seen as the highest extension of a general principle of predictive learning that operates across multiple time scales throughout the hierarchy of cortical areas. This would be consistent with the ideas initially developed by Chip Levy and colleagues (Levy, 1985.. etc)..

There are also a few directions for further modification of the hippocampal learning mechanisms based on some intriguing aspects of the underlying neuroscience. For example, there is evidence that CA3 pyramidal cells might also backpropagate to mossy cells in the polymorphic layer of DG, and potentially these signals can act as modulators to DG representations (cites)

Neurophysiologically, we know that GCs send efferent signals to both mossy cells and GABAergic interneurons in the hilus, these neurons then send excitatory and inhibitory signals back to some GCs, respectively (cites). Furthermore, mossy cells also send excitatory signals to GABAergic interneurons, which then circle back. These circuits were described as another kind of recurrent network in the hippocampus besides the CA3 recurrent network (Lisman, 1999). Potentially, this excitatory-inhibitory dynamic amplifies/filters GCs' activities to achieve strong pattern separation in DG.

The big loop recurrence plays a critical role in the hippocampal system, supporting integration of information from neocortex across different episodes (Koster et al., 2018; Kumaran, Hassabis, McClelland, 2016). Although not used in the current task, it could still be observed in the model (Figure 7) and might potentially be another source of temporal difference error that casts different training signals to the hippocampus.

TODO: The most valid error signal is: given a partial cue, do you correctly recall the original episodic memory. This is hard to use computationally: only available at test, and how do you know if you got the *right* episodic memory!? There might be some kind of reinforcement learning dynamic available in this case, where if you recall a *useful* memory, it is reinforced, even if in some sense it isn't known to be the correct original memory. In general, need a learning mechanism that can work during initial learning, when the full pattern is present, which limits options. In the cortex, we use predictive learning to get around this issue: attempt to predict the current input based on prior information. But episodic memories are not predictable in the same way. Could try to randomly turn off different parts of input, then learn to reactivate the current full input, but not clear there is time in the moment to do this. In short, the current proposal of learning to produce the DG-driven pattern separated pattern via the shortcut EC→CA3 weights provides a reasonable proxy for these more direct forms of error-driven learning. Furthermore, the models demonstrate that it is indeed effective in supporting subsequent recall from partial cues.

Further issue: monosynaptic learning is based on CA3→CA1→ECout vs. ECin – that trains CA3→CA1 projections. Why can't that also grain CA3 itself! The problem is that there is only a very indirect way for ECout to feed back and affect ECin, and this should mainly arise with partial retrieval cues as noted above, so it is not clear how this error signal might work. TODO: update in light of big loop tests, and probably move to results section!

TODO: Jeffrey Gray, etc

The longstanding idea that the hippocampus acts as a comparator has been historically intriguing for many reasons (Gray, 1982; Lisman & Grace, 2005; Vinogradova, 2001). Although the specific hippocampal subregions responsible for this computational purpose is still under debate, we here suggest that, using temporal difference as the error function, powerful learning could emerge naturally in CA1 (Ketz et al., 2013) and CA3 in the current model, and potentially expand to other areas (e.g., DG) as suggested before. While previous theories mostly focused on a scaler error signal (i.e., match/mismatch) of the hippocampal com-

parator that guides actions, the current model suggests that vector-based error signals in the hippocampus could emerge from temporally different patterns, which drive fast learning and prevent interference.

TODO: Mark Gluck http://www.ncbi.nlm.nih.gov/pubmed/7826508 MyersGluck95 Stimuli representations in the cortex are mediated by hippocampus representations Don't see a strong link to the current paper..? They're using EDL in hip Hip - PS always – Marr Unclear how backprop would work in hip

The use of backpropagation in the hippocampus or neocortex is not a new idea and could simulate many empirical data (Myers & Gluck, 1995). However, it is still unclear how backpropagation could work in the brain biologically. Using neural activities from different time points or spatial compartments, local information is available for biologically plausible error-driven learning, which in some cases follow the gradient produced by backpropagation (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020). In the current model, we broaden the use of this kind of error-driven learning mechanism in the hippocampus as suggested by our previous version of the model (Ketz et al., 2013) and suggests that the hippocampus should break its historical affiliation with Hebbian learning.

TODO: make predictions!!! Something like when lesioning X, Y will get worse? Or RSA patterns for unknown questions? Testing effect – general EDL in hip/cortex

The use of the error-driven learning mechanism in the hippocampus might be critical to account for behavioral results of cognitive tasks that require hippocampal-cortical interaction. For example, the testing effect has been associated with superior learning in both the hippocampus and the cortex, and unpublished results suggest that lack of the error-driven component in either the hippocampus or cortex may diminish the testing effect.

Behaviorally, the use of the error-driven learning mechanism in the hippocampus and its interactions with the cortex might be helpful when performing cognitive tasks that require computing differences between different task states. For example, testing effects have been shown to rely (?) on the interaction between hippocampus and the neocortex, and unpublished results suggest that lack of the error-driven component in the hippocampus may diminish the testing effect.

One consideration is whether the improved learning exhibited by the Theremin model is now somehow at a superhuman level, and thus renders our model less plausible as a realistic model of hippocampal function. However, there are several factors mitigating against this concern. First, our models are trained and tested in a "blank slate" manner, with no prior episodic memories that would likely drive significant proactive interference (note that the pretraining is only done on the EC $\leftrightarrow$ CA1 pathway, and not on the overall episodic memory circuitry involving CA3).

In addition, these granule cells have been identified as a rare locus of adult neurogenesis (?), which has been shown to enhance the pattern separation capacity of the DG in computational models (?). A future version of the model will explore the effects of neurogenesis, and other speculative possible feedback circuits within the DG as discussed later.

Appendix

Learning Algorithm

Can we reference Ketz et at. for algorithms except for new ones we need in the paper? YES I THINK IT SHOULD BE FINE. I don't even think we need this supplement – just stick everything into the main materials and methods.

Model parameters Size params Key params Number of units in each subfield Results for Statistical Analyses See the RMarkdown doc for stats

TODO: Shortening principles, strengthening EDL! RSA across cycles to explain HOW EDL helps pre-

vent interference Third, the paper is written with a sense that the reader already is up to speed on our model and on the AB-AC paradigm and the key issues in model implementation (e.g., pretraining). Other key pieces of information, such as the # units dedicated to A, B, and context, and the relative sizes of the different subfields isn't there (though this may be in the figures). Take photos of the model Make a table of unit numbers etc in every subfield Adding info about CURRENT exp using the model in Discussion paragraph 2 Check whether intro and discussion is coherent Discuss new figures of RSA results!

<div align="center">*</div>

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. (1960, December). Fate of first-list associations in transfer theory. *Journal of experimental psychology*, *58*, 97–105. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/13796886`

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

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., Yonelinas, A. P., & Ranganath, C. (2007, July). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*(1), 123–152.

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.

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

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

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`

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.

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., & 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., & Rohrlich, J. (2021). Deep predictive learning in neocortex and pulvinar. *Journal of Cognitive Neuroscience*, *33*(6), 1158–1196.

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

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.

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 reconciling 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

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.

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`

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

Zhang, X., Schlögl, A., & Jonas, P. (2020, August). Selective Routing of Spatial Information Flow from Input to Output in Hippocampal Granule Cells. *Neuron*, S0896627320305237. doi: 10.1016/j.neuron .2020.07.006