Reviewer's Responses to Questions

**Comments to the Authors:  
Please note here if the review is uploaded as an attachment.**

Reviewer #1: Review of PLoS Comp Biol article Dec 2021  
  
The paper shows how an alteration to connection-strength update rules in a multi-region connectionist model of hippocampal function leads to improved performance—in terms of faster learning and less interference—in associative learning tasks. The results are interesting as are the updates to the learning rule. My main concern is that the authors assume the reader will know or read a whole series of prior papers presenting models upon which this one is built – the methods presented here are simply inadequate for a reader to fully understand, yet alone reproduce the work. The authors send the reader to a githib repository, but even there basic features are unclear – for example, whether “Adaptive Exponential Leaky Integrate and Fire” model neurons are used, or firing-rate models based on the latter. These are the basic building blocks of any model, along with the impacts of connections – amplitudes and time constants, most importantly. The reader should not have to wade through all the prior papers to figure out what procedures might have been used in this one. I suggest providing full tables of the form of Nordlie et al, PLoS Computational Biol 2009.  
  
My second major thought is that to make small adaptations to a prior model in order to improve performance in a single task is not really newsworthy to any others than those who work on this somewhat niche sets of models. I think at least it should be shown that a second, more interesting phenomenon, such as the “testing effect” with partial information, which the authors say could arise from their model, should be validated. In this reviewer’s mind, that would make the paper stronger and a bit more general.  
  
Some examples of the need for better explanation follow:  
  
The equations shown are not mathematically rigorous. For example, equations 3 and 4 which indicate the rule used to update specific connection weights should indicate the two units (e.g. i and j) stating which is presynaptic, which is postsynaptic, and then how it depends on the firing rates r of either unit i or unit j.  
  
Equation 3 is particularly confusing as the difference in rates used is of vectors of different sets of units in entrorhinal cortex (layer 3 versus deep layers) which would not be the same size biologically speaking. Since it is impossible to take the difference of two vectors of different lengths, the authors must assuming identical numbers of units and then a one-to-one correspondence between units in deep layers with equivalent units in layer 3. Such important details and clear constraints/biological requirements for a model should be stated plainly so they can be tested.  
  
Since Equation 4 is the new one being used for this paper, again it would be better to see it written more accurately in terms of the time of firing rates. That is, is dW\_ij for a connection from unit i to unit j equal to the firing rate of unit i in EC to the rate of unit j in CA3 at time t¬-tau subtracted from its rate at time t for a value of tau equal to a quarter of a theta cycle, so about 30ms? If this is the case (my best guess given what is presented) then it is essential that it is stated plainly and either evidence for such a plasticity rule provided, or a clear statement that there must be a process which responds to the \*change\* in firing rate over a 30ms period (rather than the absolute rate) so that ideally experimental groups could look for it and find the corresponding biochemical process, or those who model synaptic plasticity could suggest a mechanism for it, using known processes. Without such clear statements of the requirements for a model showing how it can be disproven, there is little benefit in adding alteration to alteration of a complicated model that may or may not correspond to the underlying biology. I see there are suggestions, perhaps based on work by Hasselmo’s group on different signs of plasticity at different phases in the theta cycle, but the exact requirements and equations are not provided. The text at the top of p.8 (which is far from Figure 1 on p.4, may need to be expanded and connected better with that figure’s caption.  
  
On p.9 the authors state they use in their default setting, a pretraining process that involved “turning DG and CA3 off”. While it is reasonable to assume there are cortical representations of words in our vocabulary, a bit more justification is needed, given its limited capacity, and the ease of plasticity and interference in hippocampal areas, to (1) state evidence for long-term hippocampal representations of vocabulary and (2) justification for essentially switching off hippocampal structures while such representations are acquired. Perhaps (2) is to ensure cortical representations arise (in EC) without (1) in HC, but if that is the case more explanation is needed, as it looks like EC to CA1 synapses are “trained” which would result in CA1 representations of long-term semantics.  
  
Minor:  
  
The term “epoch” is not defined in the text or methods. Fig. 6 suggests 100 cycles per epoch? Is a cycle a theta cycle, about 100ms, so epochs are 10 sec long?  
  
p.16 “the difficulties are more at the level of abstract principles” – this statement is confusing, as it seems the models are so different at the biological level the “difficulties” are far from abstract. It is pretty easy to replace a plasticity rule in an architecture and test its consequences, or to find out what parameters are needed for the rule to work – nothing is abstract about that.  
  
The authors use in italics “ps” a lot when they I think mean p-values in significance tests? This is non-standard – I think just “p” is fine, but also state what sort of test is used.

Reviewer #2: My review is uploaded as an attachment.

**Review (PCOMPBIOL-D-21-02014): Zheng, Liu, Nishiyama, Ranganath, and O'Reilly**

Zheng et al. present a solid and well-written modeling study that follows up on a previous paper from this group (Ketz, 2013). This paper borrows the same hippocampal architecture but examines variations of the learning rules at critical perforant path and mossy fiber projections to CA3 synapses. Their findings demonstrate the benefits to memory capacity, training performance, and (arguably) biological fidelity of a certain class of error-driven learning (EDL) rules. Recent advances in AI and theoretical neuroscience have increasingly focused on the importance of predictions and error signals for learning, inference, and planning. This paper provides considerable exposition to motivate critical questions about error signals and error-driven plasticity in the hippocampus with respect to its role in episodic memory formation and retrieval. A key principle of episodic memory is that maximizing future retrieval requires minimizing potential representational interference between events with similar sensory and spatiotemporal properties. Thus, we need to better understand the dynamical interplay between pattern separation processes and the pattern completion that drives CA3-based autoassociative retrieval. This paper makes a clear contribution toward this understanding and should be of major interest to the field. My main concerns involve the authors' framing of Hebbian learning as a baseline for their results and whether the current set of modeling results as presented are sufficient to substantiate the broader claims. It was an enjoyable read and is appropriate for the journal; however, the presentation and results should be strengthened to both better connect with the literature and provide additional context for the audience.

*Major comments:*

1. The abstract and other places state the main problem with Hebbian learning as weight modification that continues "unnecessarily beyond" what is sufficient for retrieval. In some places, the authors refer to this naïve notion as the "simple Hebbian approach" (e.g., start of intro, para 4). While a simplistic ∆w=µxy rule encapsulates the old "fire together, wire together" dictum, I believe it is widely understood that such a learning rule is impractical, given its tendency to explode. That is, the foil in the authors' story is a straw man. The authors might have in mind some self-normalizing variant of generalized Hebbian plasticity (e.g., Oja's rule, BCM, etc.), but they don't mention it directly or discuss comparisons with other Hebb-like variations. The main issue with this is that implying "simple Hebb" to be an appropriate baseline for comparison is both unfair to Hebb and overgenerous to the authors' evaluation of their EDL models' performance. The paper's impact, and its connection to the comp/theory literature since the Ketz (2013) paper, would substantively benefit from clarifying throughout the manuscript precisely what is meant by "Hebbian learning", both in general discussion and when describing implementations in the 2013 model and this model (e.g., the ECin → DG pathway). Since the gap being studied here is Hebbian → EDL, any obscurity around "Hebbian" terminology detracts from the interpretability of the results.
2. Relatedly, the scientific logic motivating the study of RW-style EDL rules established in the introduction (para 4) is rather weak. That logic could be paraphrased: "To avoid simple Hebb we need self-limited learning; 'one well-established class' is EDL; therefore we incorporate EDL and show that it outperforms simple Hebb." This logic undercuts itself at 'one well-established class', both because it's an appeal to authority and it hints at other classes which the authors fail to mention or study. This latter point raises questions (in a reasonable reader) of why this particular formulation of EDL was chosen for study from among the wider universe of self-limiting learning rules, whether well-established or not. Its relationship to error backpropagation is potentially interesting, but it doesn't substantiate its support in biology (which is fine, it's just that there are presumably many other self-limiting rules with varying levels of biological evidence and plausibility). If this paper is pushing us toward an "error-driven hippocampus", then what we (the field) need to see is the comparison between EDL rules and some spanning set of non-EDL self-limiting rules. (I might be mistaken, but I think the authors' NoEDL variant here simply disabled CA3 synaptic modification and the ThetaPhase variant utilized "Hebbian learning"; there is no test of non-EDL self-limited learning at those synapses.) If self-limited learning is the logical alternative to the Hebbian "mistake", then why isn't self-limited learning studied and explicated more broadly (within the context of this hippocampal architecture/model)? If EDL emerges as the self-limiting "winner" (on grounds of capacity, performance, or biological fidelity), then that makes a much stronger case. To fully support the authors' claims, I would suggest that they expand their investigation, to a reasonable extent, to provide some baseline of comparison with non-EDL self-limiting rules (e.g., FORCE learning, synaptic resource allocation, etc.).

*Minor points:*

1. The paper's title is perhaps too strongly worded, since "Hebbian Mistake" implicates Hebb in an error and conjures up something of a straw man (cf.  comments 1&2 above). I can only suggest (sincerely suggest; I'm not demanding a title change here) that the authors reconsider whether the title strikes the right note given the study's claims.
2. Figure 1: Please clarify the relationship between the 4 "quarters" and the theta cycle. It would be helpful to have a simple diagram showing the correspondence of Q1-Q4 aligned with a sinusoidal theta wave. Also, it looks like Q2 and Q3 are identical? Do those quarters correspond to the cycle 25-75 interval (with the big-loop signals) in Figure 6?
3. Following on comment #3: It might be helpful to add a brief overview of the "nuts and bolts" of the model implementation (perhaps in the Methods section) that the authors could reference at this point, since many model details seem to be glossed over or presumed to be found in the 2013 paper. This should include at least brief descriptions of the units, activations, and layers an explanation of simulation timing relating time steps to cycles to quarters to epochs to train/test trials. These relationships won't be immediately obvious to most readers, and it shouldn't be necessary to go to the 2013 paper for basic information.
4. p.5, following Eq (4): "ECin is \_the\_ sending activity into CA3"; I'm assuming that 'the' is a typo?
5. p.7, para 1: "Figure 1 shows the \_standard\_ hippocampal architecture...". The word 'standard' shouldn't really be used to describe these models. This architecture might be standard in the authors' labs, but there are many ways to construct models. There are other instances of "standard" in the manuscript which should be similarly struck.
6. While "NEpochs" and "ABmem" work fine as variable names, they are awkward to read and difficult to remember. The authors could just as well use "N" and "C" (or a similarly readable refactoring) to refer to training epochs and memory capacity.
7. Are the Theremin and ThetaPhase curves shown in Figure 4 the same as those in Figure 3? If so, that should be clarified, preferably in one or both captions. Or, the authors could reconsider the presentation of this data to prevent reader confusion.
8. Figure 5: The text does mention the difference in training epochs between Theremin (10) and NoEDL (30), but these plots with differently scaled x-axes in the left and right columns are not intuitively interpretable. Because the epochs are discrete, having fewer in the Theremin plots means that the visual slope of the connecting lines has a different meaning regarding the learning dynamics of the model w.r.t. the NoEDL plots. Given the relative complexity of the data (esp. the similarity curves), the ease of interpretation of this figure would greatly benefit from an attempt to horizontally stretch (half of) the plots so that the x-axis scales are preserved over all of them.
9. Figure 6: It would be very helpful to add some dotted lines or arrows to indicate the cycles at which DG input turns on (25). Also, something clearly happens around cycle 75 at the end of Q3, but the text and caption don't make this very clear. This could be clarified in the Methods as I suggested above or in other ways as the authors see fit.
10. Discussion, p.15, para 3 "Another class...": This paragraph discusses the long-running class of sequence learning models, but it doesn't come to a very satisfying conclusion. How exactly are we supposed to interpret the current study's results in the context of sequence learning, which depends precisely on the kinds of learned representational overlaps that the authors' CA3 EDL rule and EC → DG plasticity work to remove? The authors' seem to dismiss this entire literature by stating that "most" of these models synthesize predictions. If it's only most that can be thusly dismissed, then the ones that remain need to be addressed, no? How would Chip Levy's local context units fare in an EDL regime?
11. Discussion, Novel Predictions, second bullet point: Since the promise of experimental predictions was prominent in the abstract and other places, it's not clear why the authors did not already undertake to investigate some of these pathway-specific modulations in their model. As written, the predictions are very vague: "should affect error-driven learning", "the specific temporal dynamics associated...". Given the effort and expense of experimental corroboration, it would be more impactful to more clearly and precisely demonstrate hypothesis-driven results from pathway lesion/modulation interventions in the Theremin model to both inspire and guide experimentalists who may be interested in these questions.

Reviewer #3: Attached.

The manuscript submitted by Zheng et al. details a compelling model in which synaptic strengths of connections in the medial temporal lobe are updated during specific phases of a theta cycle according to error-driven learning rules. The model focuses on key regions of the medial temporal lobe in which hippocampal synaptic weights can be adjusted (CA1 inputs to EC, EC inputs to CA3, or CA3 recurrent collateral inputs) to yield activity in CA1 or CA3 that matches target activity patterns (driven by EC or DG, respectively). The authors note that such error-driven learning rules provide the advantage of limiting synaptic weight changes to only those necessary to faithfully represent input patterns, mitigating interference created by excessive weight changes and increasing overall memory capacity. I have several broad questions regarding the goal of achieving specific target activity patterns in this system and the necessity of the theta-constrained timing to achieve the temporal differences in activation state. I also have a number of recommendations that I believe will improve the clarity and impact of this work.

Major Comments:

1. One of the central goals of this model (and a previous model published by this group) is to adjust synaptic weights over the course of learning until EC­OUT matches ECIN, with the specific addition of adjustments in this model until CA3 DG − CA3 OUT. I am familiar with the history of the CLS model and the advantage that the new adjustments in this model provide for reducing interference. The adjustments in the original version of this model (EC­OUT matches ECIN) require a bit more explanation. What are the goals/advantages of having EC­OUT match that of ECIN? This seems to imply that CA1 is being optimized to provide faithful representations of original EC input, rather than performing any unique transformations. The *dW* = CA1(EC­OUT - ECIN) seems to negate somewhat the advantages afforded by *dW* = EC(CA3­DG – CA3NODGE). In other words, the structure of *dW* = CA1(EC­OUT - ECIN) seems like it disrupts the ability of CA1 to output unique separations, associations, or other transformations formed *within* the hippocampus. Moreover, it makes it challenging to incorporate strong recent evidence that entorhinal activity is heavily influenced by hippocampal activity. It might be interesting to probe CA1 activity patterns of EC OUT without *dW* = CA1(EC­OUT - ECIN) while still retaining *dW* = EC(CA3­DG – CA3NODGE) to provide insight into how *dW* = CA1(EC­OUT - ECIN) influences hippocampal computations/interference (and then change the name of the noEDL condition).
2. The authors attempt to leverage the unique timing of different inputs within a theta cycle in order to allow different patterns of activity to manifest across a region at distinct times. While I think it is a great idea to consider a major temporal organizing principle of the region, I am not sure the model needs to be nested within a theta timescale. Attempting to do so makes it fragile to scrutiny of the feasibility of the timescale and potentially less applicable to primate models where the theta rhythm is less prominent. The authors include some citations indicating that plasticity within a theta timescale may work (e.g. Hasselmo et al. 2002), but the work was targeting projections to CA1, which are not the target of their error-driven learning models. Synaptic plasticity of other synapses may obey very different temporal rules, and a more recent study they cite from Quirk et al. suggests that rhythms facilitating learning can be quite broad (i.e. <10Hz). Synaptic plasticity of other synapses may obey very different temporal rules. I think that the authors can maintain the theta cycle structure of the current model, but perhaps provide context that it is a valuable framework here for achieving their goals rather than *the* only hypothesized mechanism for error-driven learning. They can provide context in the introduction and discussion with any additional references justifying their timing and other possible timing mechanisms.
3. The authors do a great job explaining the relevant connectivity. However, there are occasions where it is not clear which connections are undergoing Hebbian learning and which connections are undergoing error-driven learning. For example, there are places in the text where the language can read as if there is EDL at CA3 inputs to CA1. Figure 1 is extremely helpful in detailing the connections that are giving rise to expected patterns of activity at different phases of the theta cycle. Perhaps this figure can be adapted to indicate (perhaps through dotted or dashed lines versus solid lines) which connections have EDL, which ones have Hebbian learning, and which (if any) have no plasticity. In addition, perhaps patterns of activity can be labeled as CA3DG (or “target”) versus CA3NODGE (or “expected”) when they occur in the upper panels, and similarly for other patterns utilized for EDL. Adding the equations next to the connections undergoing EDL would also be tremendously helpful.

Minor Comments

1. The acronym Theremin, while fun, is an odd combination of letters from the model name, has an extra E, and perhaps doesn’t capture the description or goals of the model well. I understand that model names are often somewhat personal, but I think that a different name would increase the impact and accessibility long-term. In addition, as there are only two synapses with EDL, I suggest removing the word Total from the model name.
2. I find it hard to interpret the tables at the end of the manuscript, particularly the testing table. I am not exactly sure what is to be learned from these table. If no input is given for the B/C pool, then why do several pools have Context B#/C# labels? In addition, B#/C# goes to 4, when this notation is used in the test to indicate a single pair of a # of possible pairs much greater than 4. Perhaps the two tables can be combined into a figure detailing how the inputs to the a subset or all pools evolve across training and test.
3. The authors experiment with plasticity at EC-DG synapses, and their results suggest that favoring LTD increases separation/reduces interference. Perhaps the authors could relate this finding to what is known about homeostatic synaptic plasticity at these synapses (Hsu et al., *The dentate gyrus as a filter or gate: a look back and a look ahead*, 2007)
4. The authors relate there model to several other models and comment on the ability of the hippocampus to generate predictions. It would be interesting if they could also relate their model to Jung et al. 2018 (*Remembering rewarding futures: A simulation‐selection model of the hippocampus*).
5. There is a typo in the Figure 2A description “traind.”

**Have the authors made all data and (if applicable) computational code underlying the findings in their manuscript fully available?**  
The [PLOS Data policy](https://journals.plos.org/ploscompbiol/s/materials-and-software-sharing) requires authors to make all data and code underlying the findings described in their manuscript fully available without restriction, with rare exception (please refer to the Data Availability Statement in the manuscript PDF file). The data and code should be provided as part of the manuscript or its supporting information, or deposited to a public repository. For example, in addition to summary statistics, the data points behind means, medians and variance measures should be available. If there are restrictions on publicly sharing data or code —e.g. participant privacy or use of data from a third party—those must be specified.

Reviewer #1: Yes

Reviewer #2: Yes

Reviewer #3: Yes

PLOS authors have the option to publish the peer review history of their article ([what does this mean?](https://journals.plos.org/ploscompbiol/s/editorial-and-peer-review-process#loc-peer-review-history)). If published, this will include your full peer review and any attached files.  
  
  
If you choose “no”, your identity will remain anonymous but your review may still be made public.  
  
  
**Do you want your identity to be public for this peer review?** For information about this choice, including consent withdrawal, please see our [Privacy Policy](https://www.plos.org/privacy-policy).

Reviewer #1: No

Reviewer #2: No

Reviewer #3: No

Figure Files:

While revising your submission, please upload your figure files to the Preflight Analysis and Conversion Engine (PACE) digital diagnostic tool, [https://pacev2.apexcovantage.com](https://pacev2.apexcovantage.com/). PACE helps ensure that figures meet PLOS requirements. To use PACE, you must first register as a user. Then, login and navigate to the UPLOAD tab, where you will find detailed instructions on how to use the tool. If you encounter any issues or have any questions when using PACE, please email us at [figures@plos.org](mailto:figures@plos.org).

Data Requirements:

Please note that, as a condition of publication, PLOS' data policy requires that you make available all data used to draw the conclusions outlined in your manuscript. Data must be deposited in an appropriate repository, included within the body of the manuscript, or uploaded as supporting information. This includes all numerical values that were used to generate graphs, histograms etc.. For an example in PLOS Biology see here: <http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001908#s5>.

Reproducibility:

To enhance the reproducibility of your results, we recommend that you deposit your laboratory protocols in [protocols.io](http://protocols.io), where a protocol can be assigned its own identifier (DOI) such that it can be cited independently in the future. Additionally, PLOS ONE offers an option to publish peer-reviewed clinical study protocols. Read more information on sharing protocols at <https://plos.org/protocols?utm_medium=editorial-email&utm_source=authorletters&utm_campaign=protocols>