Title: Memory in structurally unstable neural networks

Abbreviated Title: Memory in structurally unstable neural networks

Authors: Daniel Acker[1,a,b], Paul Miller[1,2], Suzanne Paradis[2]

[1] Brandeis University, Department of Neuroscience, 415 South St, Waltham, MA 02453
[2] Brandeis University, Department of Biology, 415 South St, Waltham, MA 02453

[a] Submitting author
[b] Corresponding author, dacker@brandeis.edu

Number of Pages: 43
Number of Figures: 5
Number of Tables: 1
Number of Multimedia and 3D Models: 0

Number of Words for Abstract: 154
Number of Words for Introduction: 647
Number of Words for Discussion: 1266

Acknowledgements: This work was supported by NIH grant R01NS065856 (S.P.) and NARSAD Independent Investigator grant number 23405 (S.P.). We thank Dr. Stephen Van Hooser for critical reading of the manuscript.

Dedication: To the memory of Dr. John Lisman who contributed to this work through stimulating discussion and insightful feedback.

Conflict of Interest: The authors declare no competing financial interests.

Abstract: Recent research suggests that spines in the adult mouse hippocampus are highly transient, with an average lifetime of 10.5 days. If this is true, ~95% of synapses onto a particular neuron will turn over within 30 days. Here, we use computational modeling to ask how memories can persist in the hippocampus and similarly unstable brain structures. We demonstrate that reactivation and Hebbian learning of weights are sufficient to drive newly formed synapses to approximate the collective role of lost synapses; we propose that this process can stabilize memories. Further, we find that in certain types of networks, such as those with winner-takes-all dynamics, it is possible to preserve multiple memories despite random synapse turnover. Finally, we turn to a maximum-excitation-modulated winner-takes-all model of hippocampal place cells and demonstrate that, consistent with recent *in vivo* observations, the place codes of individual cells may be stable given learning during reactivations.

Significance Statement: The impermanence of synaptic connections is an emerging theme in the study of hippocampal and cortical networks. We analyze the stability of memory traces in neural networks undergoing random rewiring. Our simulations indicate that Hebbian learning of synaptic strengths is sufficient to maintain memory under such conditions. Thus, we propose that memory is dependent not on individual synapses but on the propagation of network dynamics.

Introduction

Dendritic spines in adult mammals can be eliminated or grow in new locations (Trachtenberg et al., 2002; Grutzendler, Kasthuri, and Gan, 2002). The reported degree of spine turnover in adult mice varies considerably by brain area. In cortex, most spines appear to be

stable (Grutzendler, Kasthuri & Gan, 2002; Holtmaat et al., 2005; Attardo, Fitzgerald, & Schnitzer, 2015). In contrast, all spines in hippocampal area CA1 seem to be transient with a mean lifetime of ~10.5 days (Attardo, Fitzgerald, & Schnitzer, 2015). Thus, the hippocampus is expected to undergo dramatic structural remodeling with ~95% replacement of the CA1 spine population every 30 days (Attardo, Fitzgerald, & Schnitzer, 2015).

In light of such rapid and extensive restructuring of the hippocampal network, it is surprising that spatial memories can persist for at least one month in mice (Guskjolen, Josselyn, & Frankland, 2017). While these memories might not be entirely hippocampus dependent, Abraham et al. (2002) found that multi-afferent long-term potentiation (LTP) can persist for up to one year following stimulation in the rat dentate gyrus, suggesting that stable learning is possible within the hippocampus. Further, place fields of CA1 place cells can be stable in mice and rats (Thompson & Best, 1990; Agnihotri et al., 2004; Kentros et al., 2004; Ziv et al., 2013). While most place cells lose their location preference between exposures to an environment, some retain their place fields (Ziv et al., 2013; Rubin et al., 2015). The centroids of recurring place fields display little drift compared to their positions on earlier exposures (Ziv et al., 2013; Rubin et al., 2015). Most surprisingly, drift magnitude does not appear to increase over time, i.e. mean drift is no different after 30 days than after five days (Ziv et al., 2013). Similarly, the quality of a recalled spatial memory is not different between days one and 30 (Guskjolen, Josselyn, & Frankland, 2017).

Several proposed mechanisms might account for memory persistence in structurally unstable networks. One proposal is that a subset of spines are stable, and these are sufficient to encode memory (Mongillo, Rumpel, & Loewenstein, 2016; Chambers & Rumpel, 2017). This appears to be the case in some cortical regions, where a fraction of spines may persist throughout

47 the lifetime of the animal (Yang, Pan, & Gan, 2009). A second explanation is that spine stability
48 is regulated by activity dependent plasticity. Supporting this view, Hill and Zito (2013) found
49 that LTP stabilizes nascent spines in hippocampal slices from neonatal mice, although this
50 process might not be relevant in adults (see Discussion). A third suggestion is that network
51 phenomena correct the destabilization caused by turnover of individual synapses. Correction
52 mechanisms might include feedback error signals or reinforcement of attractor states (Mongillo,
53 Rumpel, & Loewenstein, 2016; Chambers & Rumpel, 2017)

54 Here, we use modeling to explore the hypothesis that memory can be embedded at the
55 network level (Mongillo, Rumpel, & Loewenstein, 2016; Chambers & Rumpel, 2017). Our view
56 is that synapse turnover-induced destabilization could be counteracted by network dynamics
57 during reactivation events that would train newly formed synapses to approximate the collective
58 role of removed synapses. We find that this approximation can be driven by Hebbian weight
59 plasticity and is sufficient to stabilize memory traces in sparsely connected attractor networks
60 despite synapse turnover, allowing a memory to survive complete rearrangement of the
61 connection matrix.

62 Storage of multiple memories is poor under turnover because the increased information
63 loss blurs the boundaries between memories. This deterioration can be offset by winner-takes-all
64 dynamics that prescribe a limited set of robust output states. Interestingly, hippocampal place
65 fields might be generated by a winner-takes-all process (de Almeida, Idiart, & Lisman, 2009a).
66 To explore the possibility that network-level memory could explain place field stability, we
67 turned to a maximum-excitation-modulated (E%-max) winner-takes-all model of the grid-cell-to-
68 place-cell transformation (de Almeida, Idiart, & Lisman, 2009a). In this model, we demonstrate
69 that reactivation stabilizes place fields over biologically relevant timescales.

Results

i. New synapses may collectively approximate lost synapses through Hebbian plasticity of synaptic strengths

In deep neural networks, information is propagated in steps from one layer of units to the next. Analogously layered architectures can be found in classical conceptions of both the mammalian visual system and the trisynaptic circuit of the hippocampus, where series of anatomical regions are linked by primarily feed-forward connections (Andersen, Bliss, & Skrede, 1971; Herzog & Clarke, 2014). In such networks, synapse turnover would result in a change in the way an input is projected across anatomical regions. We ask whether the projection of an input onto a downstream region can be preserved despite synapse turnover upstream.

Inspiration for an answer may be found in the overfitting problem of statistics and machine learning. An overfitted model can transform random noise into a nearly perfect prediction (Hawkins, 2004). Overfitting often occurs when the input is high dimensional, i.e. there are many covariates, and there are few outcomes. The overfitting problem demonstrates that a sufficiently high dimensional input can be projected to produce a close approximation of any desired output. In the case of a neural network, two units connected to different pre-synaptic populations (covariates in the overfitting analogy) can produce similar output if the pre-synaptic populations are large and the synaptic weights are tuned appropriately.

During synapse turnover, a post-synaptic neuron can become disconnected from a subset of its original pre-synaptic partners and form new connections with a set of different pre-synaptic

partners. The firing rates of the new synaptic partners might be uncorrelated with the rates of partners lost due to turnover. Nevertheless, careful tuning of synaptic weights can maintain the unit's firing rate dynamics (Fig. 1). This maintenance would come about because the sum of post-synaptic potentials (PSPs) over the stimulus at new synapses would resemble the sum of PSPs over the stimulus at lost synapses.

An important question is how the required careful tuning of synaptic weights could come about in the brain. We reasoned that, in high in-degree networks with biased response properties, the activity of a neuron would not be dramatically perturbed by replacement of a small fraction of inputs. If firing rates are not drastically altered, Hebbian learning might be sufficient to tune new synaptic weights such that the neuron's response properties are propagated through time.

To test this hypothesis, we simulated a simple network consisting of 100 pre-synaptic neurons that provided input to a single post-synaptic neuron. Activities of the pre-synaptic neurons over time were simulated by cosine functions with random phase and period. Activity of the post-synaptic neuron was defined as $\boldsymbol{y} = \boldsymbol{Xw}$, where $\boldsymbol{y}$ is a matrix with the firing rate of the post-synaptic neuron for each pattern (columns of $\boldsymbol{y}$ represent distinct cells; rows of $\boldsymbol{y}$ represent distinct input patterns), $\boldsymbol{X}$ is the input matrix (columns of $\boldsymbol{X}$ represent distinct cells; rows of $\boldsymbol{X}$ represent distinct levels of the stimulus in each pattern) and $\boldsymbol{w}$ is the synaptic weight matrix. Weights were initialized to random values between -1 and 1 and were updated based on Hebb's rule, $\boldsymbol{\Delta w = \eta X^T y}$, where $\boldsymbol{\eta}$ is the learning rate, set to $\boldsymbol{10^{-2}}$. The network was subjected to pre-training: one round of Hebbian learning to establish initial bias (see Materials and Methods for description). Next a random 50% of the input neurons were replaced with new neurons with activities characterized by new random cosine functions. The synaptic weights for these neurons were re-initialized to random values between -1 and 1. After replacement, the network was

subjected to reactivation: re-exposure to the inputs and a second round of learning. In control simulations, the pre-turnover and/or post-turnover training step was omitted.

We found that, following reactivation, the sum of PSPs at new synapses was consistently correlated with the sum of PSPs at lost synapses, as assessed following pre-turnover training (Fig. 2). However, if either training step was omitted, PSPs at new and lost synapses were predominantly uncorrelated (Fig. 2). These results suggest that Hebbian plasticity of weights in a biased network can lead new synapses to take on the role of lost synapses.

ii. Synaptic reentry reinforcement stabilizes memory in attractor networks despite turnover

To ask whether multidimensional memories can survive synapse turnover, we turned to the Hopfield attractor memory model (Cohen & Grossberg, 1983; Hopfield, 1983). In this model, synaptic reentry reinforcement (SRR) or periodic reactivation of the attractor coupled with learning is a mechanism that can maintain memories despite constant decay of synaptic weights such as would be caused by the turnover of synaptic receptors (Wittenberg, Sullivan, & Tsien, 2002). We hypothesized that SRR could also stabilize memories in networks undergoing turnover of whole synapses. To test this, we constructed sparsely connected Hopfield networks. The network activities and changes in synaptic weight were characterized by Wittenberg, Sullivan, and Tsien's (2002) equations 1 and 2 (see Materials and Methods), which increase the strength of connections between co-active neurons and otherwise reduce connection strengths. The networks were trained on a binary {-80, 80} pattern (Fig. 3A[inset]). Training was

performed as described by Wittenberg, Sullivan, & Tsien (2002) (see Materials and Methods for description).

To evaluate the effect of synapse turnover on the stability of memory in the attractor networks, we assigned specific turnover rates to the networks such that a fraction of random synapses (equivalent to the turnover rate) would be removed on each turnover iteration, and an equal number of new synapses would be formed at random, unoccupied locations in the connection matrix. We performed three experiments in which we alternately varied one of three parameters: the synapse turnover rate (Fig. 3A,D), the probability of connection ($\boldsymbol{p_{connection}}$) between any two neurons (Fig. 3B,E), and the number of units in the network (Fig. 3C,F), while holding the remaining two parameters constant. When varied, parameters were randomly sampled in the ranges shown in Table 1, except for number of units, which was randomly sampled from the set of perfect squares in the range.

After initial training, the networks were subjected to reactivation of the attractor with synapse turnover occurring at the beginning of each reactivation event. We defined the process of reactivation as follows: initialization of the network with random membrane potentials ($\boldsymbol{u}$) between -0.004 and 0.004, and updating $\boldsymbol{u}$ for 12 time steps. Synaptic weight updates were performed once per reactivation after 12 time steps. Reactivation proceeded for 100 iterations. At reactivations one and 100, we determined the coefficient of determination ($\boldsymbol{r^2}$) between the current network state and the training input.

We found that SRR-mediated stabilization depended on the turnover rate (Fig. 3A,D), $\boldsymbol{p_{connection}}$ (Fig. 3B,E), and the number of units in the network (Fig. 3C,F). When the turnover rate was low, the memory was stable: $\boldsymbol{r^2}$ values were high and did not differ between activation one and 100. However, at higher turnover rates the memory was unstable: $\boldsymbol{r^2}$ declined between

activations one and 100 (Fig. 3A). When $\boldsymbol{p_{connection}}$ was low, $\boldsymbol{r^2}$ declined, and when $\boldsymbol{p_{connection}}$ was high, $\boldsymbol{r^2}$ did not decline (Fig. 3B). When the number of units was low, the memory was unstable, and when the number of units was high, the memory was stable (Fig. 3C).

To explore the extent to which $\boldsymbol{p_{connection}}$ and unit number could protect against turnover, we performed the attractor network simulations as described previously while varying all three parameters simultaneously. For visualization, we used median in-degree (the number of inputs to a cell) as a stand-in for $\boldsymbol{p_{connection}}$ and number of units, as in-degree is linearly dependent on both of these parameters. We found that networks with higher in-degree were surprisingly resistant to memory degradation by synapse turnover, with some networks exhibiting stable memory despite greater than 80% turnover per reactivation (Fig. 3G).

To ask how SRR and synapse turnover affect the attractor dynamics, we examined three networks, one with zero synapse turnover, one with turnover but stable memory, and one with turnover and unstable memory. All networks consisted of 100 units with $p_{connection} = 0.2$. We found that the networks with zero turnover or stable memory consistently reached an attractor state (all changes in membrane potential $[dV_m(t) = V_m(t) - V_m(t-1)]$ equal zero) in six time steps, while the network with unstable memory failed to reach an attractor state within the 12 time steps allotted for an activation. Taken together, these data suggest that SRR can stabilize learned attractors in neural networks undergoing synapse turnover.

iii. Winner-takes-all dynamics stabilize multiple memories

While Hebbian plasticity of weights can preserve memory in neural networks undergoing synapse turnover, we reasoned that turnover would reduce the number of memories that could

successfully be stored. As more memories are reinforced during reactivation, the overall learning should tend toward the mean of pre and post-synaptic activities for a particular neuron. This mean need not correspond to any particular memory, thus when the number of memories is increased, all memories can be lost in a process called catastrophic forgetting (French, 1999). Synapse turnover should exacerbate catastrophic forgetting because the turnover-induced information-loss between training events can distort the boundaries between memories. One strategy to increase memory capacity would be to make the network output discrete and sparse. Discreteness would aggregate nearby network states, and sparsity would map memories to minimal ensemble representations, reducing the frequency of overlap. Winner-takes-all networks, in which only the most excited cell (or $\boldsymbol{k}$ cells) are able to fire, exhibit both of these properties.

We asked if winner-takes-all networks would retain a high memory capacity despite synapse turnover. To test this, we created three networks. One was a simple feed-forward (identity) network consisting of 1000 pre-synaptic neurons and 100 post-synaptic neurons. The $\boldsymbol{p_{connection}}$ between each pre and post-synaptic neuron was 0.2. The activities of the pre-synaptic neurons were described by random numbers sampled from the uniform distribution in the range (-1, 1). The activities of the post-synaptic neurons are described by $\boldsymbol{y = Xw}$, where $\boldsymbol{X}$ represents the input matrix (columns of $\boldsymbol{X}$ represent distinct cells; rows of $\boldsymbol{X}$ represent distinct input patterns or experiences) and $\boldsymbol{w}$ represents the synaptic weight matrix. Weights were initialized to random values between zero and one and updated based on the learning rule $\boldsymbol{\Delta w = -w + \tanh(w + \eta X^T y)}$, where $\boldsymbol{\eta = 0.1}$. The hyperbolic tangent function was included to limit the minimum and maximum synaptic weight. The second network was a winner-takes-all network. Construction was the same as for the identity network, except that the firing rates of output neurons that were

not in the top 10% were set to zero. The third network was an E%-max winner-takes-all network. In an E%-max winner-takes-all network, only cells excited to within a fixed percentage of the most excited cell are able to fire (de Almeida, Idiart, & Lisman, 2009b). In our experiment, the E%-max winner-takes-all network was constructed like the identity network; however, the firing rates of output neurons were set to zero when they were not within 10% of the firing rate of the output neuron with the maximum rate.

All networks were initially trained on the inputs with a single update of all synaptic weights based on the activities produced by each input independently. The number of presented input patterns ranged from 5 to 100. Next, the networks underwent turnover at rate of zero or 10% per update, with the update based on the activity due to presentation of each pattern once. In a separate set of simulations, the update was based on presentation of just a single pattern. In the latter case, patterns were cycled, interleaved with turnover, and the turnover rate was adjusted so that 10% turnover would be achieved by the end of each complete cycle (see Materials and Methods for description). This was repeated for 99 iterations. In the case of interleaved presentations, an iteration refers to a complete cycle of patterns. On a final, 100^{th} iteration, all inputs were presented again. To assess memory preservation, we calculated the correlation between each network's output in response to each input pattern at the final iteration and the output in response to the initial presentation of the same input pattern. We observed that the identity network was quickly overwhelmed by multiple memories such that preservation rapidly tended toward zero as the number of input patterns increased (Fig. 4A,C). In contrast, the winner-takes-all and E%-max winner-takes-all networks displayed enhanced memory capacity (Fig. 4A,C).

In addition to consistent recapitulation, it is important that a memory be uniquely activated by a particular input or set of inputs. To evaluate the uniqueness of memories in our models, we compared (by subtraction) the preservation of a memory to that memory's correlation with the most correlated other memory. We found that memory uniqueness was greater in the winner-takes-all and E%-max winner-takes-all networks compared to the identity network (Fig. 4B,C).

iv. Place fields are stable despite synapse turnover in an E%-max winner-takes-all model

Winner-takes-all dynamics seem to be an effective strategy to preserve multiple memories despite synapse turnover. The hippocampus is both a site of high synapse turnover and a winner-takes-all network: place fields in the hippocampus have been proposed to arise from an E%-max winner-takes-all competitive process mediated by the gamma frequency oscillation (de Almeida, Idiart, & Lisman, 2009a). This dynamic manifests as the interneuron-mediated competition between place cells receiving excitatory input from grid cells in entorhinal cortex and/or place cells in upstream hippocampal subregions (de Almeida, Idiart, & Lisman, 2009a; de Almeida, Idiart, & Lisman, 2012). We hypothesized that this network architecture would enable place fields to survive in the high synapse turnover environment of the hippocampus.

We simulated the grid-to-place-cell transformation as described by de Almeida, Idiart, & Lisman (2009a). The network consisted of 10000 grid cells and 2000 place cells. Grid cells were modeled as a combination of 2d cosine gratings, as described by Blair et al. (2007) (see Materials and Methods). Each place cell received synaptic input from 1200 randomly selected grid cells. All grid-to-place-cell connections were monosynaptic. Synaptic weights were initialized to a

252 uniform value. Place cells' activities were calculated as described by de Almeida, Idiart, &
253 Lisman (2009a) such that only cells excited to within 8% of the most excited cell at a given
254 position were able to fire (see Materials and Methods for description). We added change in
255 weight formulae which enabled synaptic potentiation and depression (see Materials and
256 Methods).

257 We evaluated the activity of the network as on a 100-cm linear track by simulating the
258 grid cells at 100 equally spaced x-positions and one y-position. We set the synapse turnover rate
259 such that the mean synapse lifetime was 10 days, and we spaced reactivation events as if they
260 were occurring at one-day intervals. Thus 114 out of 1200 synapses per place cell turned over
261 between reactivations. After turnover, newly formed synapses were uniformly reset to the initial
262 weight. In total, we simulated 31 activations, representing the networks' responses on days zero
263 to 30.

264 Place fields were identified as unbroken regions of at least 5 cm where a cell was active
265 at greater than 80% of its maximum firing rate. We only considered cells with a single place
266 field. We assessed place field stability by measuring the shift in a place field's centroid from its
267 position on the first activation. When a place cell did not have a place field on a given activation,
268 it was not included in the analysis of that activation. We observed that with a learning rate of
269 zero, place fields remapped randomly over time; however, when the learning rate was increased,
270 place fields displayed more consistent centroid positions across activations (Fig. 5A,D ["No
271 learning" versus all other conditions]).

272 We next asked how variations in the learning rule would affect place field stability. To
273 test this, we made coarse adjustments to the learning process such that the network would
274 experience only potentiation (LTP-only), only depression (LTD-only), or both potentiation and

depression (LTP & LTD) (see Materials and Methods). We found that all conditions led to more stable place fields (Fig. 5A,D). The LTD-only condition displayed the greatest place cell retention, i.e. in total, more cells with a place preference were observed on reactivations, and more of the initial place cells remained as place cells on later activations (Fig. 5B,C).

To ask if new synapses were collectively approximating lost synapses, we examined the within position correlation between the sum of PSPs at new and lost synapses in the "LTP & LTD" condition. For this comparison, "lost synapses" refers to synapses removed between days 29 and 30. PSPs at lost synapses were assessed following the weight update on day 29. "New synapses" refers to synapses formed between days 29 and 30. PSPs at new synapses were assessed both prior to and following learning on day 30. Prior to the weight update, summed PSPs at new and lost synapses were positively correlated at only 48% of positions. In contrast, following the weight update, PSPs at new and lost synapses were positively correlated at all positions. (Figure 5E).

Finally, we asked whether the degree of place field stabilization in our model was sufficient to preserve consistent trajectory encodings. To test this, we trained a neural decoder (a gradient-boosted trees classifier) on the activities of place cells with place fields observed on the first activation. Gradient-boosted trees classifiers are ensemble decision-tree classifiers in which trees are iteratively fit to the residuals of earlier fits (Chen & Guestrin, 2016). Similar strategies have been used to decode animal position from neural activity recorded *in vivo*, e.g. Glaser et al. (2017) used the XGBoost algorithm (Chen & Guestrin, 2016). We then predicted the simulated mouse's location using the activities of these same cells on subsequent activations. We found that we could consistently decode the mouse's position, even after 30 days (Fig. 5G). Decoding appeared stable in the LTD-only condition; mean error did not increase across trials after day

three (Fig. 5G). These results suggest that the predicted degree of place code preservation is sufficient to represent a useful map of the environment.

Discussion

Recent research suggests that the neuronal connectomes of adult mammals are not as stable as once thought (DeBello & Zito, 2017). In regions such as the hippocampus, the majority of the connectivity structure might be rewired in a period as short as one month (Attardo, Fitzgerald, & Schnitzer, 2015). Under these conditions, it was not clear whether or how memories could be preserved in the long term. We sought to address this question using a modeling approach and found that memory reactivation and Hebbian plasticity of synaptic weights are sufficient to stably maintain memories in the midst of random synapse turnover.

Specifically, in a network with many varying inputs, our model predicts that the collective role of lost synapses will be approximated by newly formed synapses. The ability of the network to store multiple memories was enhanced by winner-takes-all or E%-max winner-takes-all dynamics. Further, in an E%-max winner-takes-all model of hippocampal place cells, reactivations and Hebbian plasticity led to stable place fields, consistent with what has been reported *in vivo* (Ziv et al., 2013). Together, these results suggest that stable synapses are not required for stable memories.

Several pre-existing models suggest that memory and computation can outlive individual synapses (Poirzai & Mel, 2001; Knoblauch et al., 2014; Fauth, Wörgötter, and Tetzlaff, 2015; Eppler et al., 2015; Gallinaro & Rotter, 2017). However, our model is unique in two ways. First, in our model synapse stability and the location of new synapses are random. This is in contrast to

models created by Poirzai and Mel (2001), Knoblauch et al. (2014), and Fauth, Wörgötter, and Tetzlaff (2015), all of which involve activity-dependent synapse formation and/or elimination. Because we leave the control of synapse turnover to chance, our model is generalizable to neural structures such as the adult hippocampus in which mechanisms for activity dependent wiring are poorly characterized. Second, our model requires only monosynaptic connections. This is in contrast to the model by Fauth, Wörgötter, and Tetzlaff (2015) in which information is represented by the number of realized synapses connecting neuron pairs. Because our model does not rely on multi-synaptic connections, it can be extended to represent a variety of primarily feed-forward neural circuits, as well as circuits with few potential connections per neuron pair.

An important open question is whether synapse turnover in the hippocampus is truly random. This is a difficult question to answer for several reasons. One reason is that some memory models require very few stable synapses, e.g. John Lisman favored a model in which only ~0.1% of spines must be stable (personal communication). Such a small stable population of dendritic spines would be below the threshold for detection by current live imaging technologies. Another reason is that spine stability might be related to spine size: LTP might lead to an expansion of the post-synaptic density and actin polymerization, resulting in an enlarged spine that is unlikely to collapse into the dendritic shaft (Bramham et al., 2010; Lüscher & Malenka, 2012). This hypothesis has not yet been tested in the adult hippocampus because the high density of dendritic spines in the region, as compared to cortex, can lead to optical merging of nearby spines, increasing uncertainty around volume estimates (Attardo, Fitzgerald, & Schnitzer, 2015).

Nevertheless, a number of studies begin to address this question. Hill and Zito (2013) found that glutamate uncaging at a single nascent spine was sufficient to induce stabilization of

the stimulated spine in organotypic hippocampal cultures. However, this work was performed in neonatal tissue, so its applicability to adult hippocampus is not clear, e.g. the half-life of spines in the stimulated preparation was only 14 hours, much shorter than that of spines in the adult hippocampus (Ziv et al., 2013). Further, given exponential decay of spines, the fraction of spines remaining 70 min after stimulation (~0.95) is inconsistent with the fraction remaining after 14 hours (0.50). One interpretation of this apparent discrepancy is that there are two categories of spines in the neonatal hippocampus: one with an adult-like half-life of ~7.3 days (derived from a mean lifetime of 10.5 days) and one with a much shorter half-life of about $\mathbf{7.4} \approx -\mathbf{14}/\mathbf{log_2}(\mathbf{0.27})$ hours (0.27 was the fraction of spines remaining after 14 hours in the unstimulated cultures). In this model, LTP would mediate the switch from the short-lived state to the adult-like state, but LTP would not indefinitely extend the lifetime of an adult-like spine.

Other studies examined the effects of theta burst LTP (Bourne & Harris, 2011), fear (Giachero, Calfa, & Molina, 2015) or spatial (Moser, Trommald, & Andersen, 1994) learning, environmental enrichment (Attardo, Fitzgerald, & Schnitzer, 2015), and NMDA receptor blockade (Attardo, Fitzgerald, & Schnitzer, 2015) on spine density in the adult hippocampus. While these studies suggest that learning can regulate overall spine density, none address the question of whether individual spines can be selectively stabilized, as predicted by Hebbian structural plasticity models. In addition, these studies seem to tell conflicting stories: both LTP (Bourne & Harris, 2011) and loss of NDMA receptor activity (Attardo, Fitzgerald, & Schnitzer, 2015), which is required for LTP (Lüscher & Malenka, 2012), lead to a reduction in spine density. Thus, further research will be required to provide a more nuanced understanding of the effects of learning and LTP on spine structural plasticity in the adult hippocampus.

An additional open question is whether spine impermanence in the hippocampus would provide an advantage to the animal. One potential advantage is a mechanism for strong memory storage in the short-term combined with an option for either rapid forgetting or long-term preservation. In one theory of hippocampal/cortical memory, the hippocampus both learns and forgets quickly, while the cortex learns more slowly but retains information longer (Lisman & Morris, 2001; Roxin & Fusi, 2013). According to this view, the memories in the hippocampus are slowly transferred to the cortex for long-term storage. These memories can then be cleared from the hippocampus, making way for new learning (Lisman & Morris, 2001; Lisman & Grace, 2005; Richards & Frankland, 2017). The regional timescales of memory retention in this model are correlated with the difference in apparent turnover dynamics between hippocampus and cortex: most cortical spines are stable, while all hippocampal spines are transient.

However, in some cases it may be beneficial to maintain a memory in the hippocampus. One posited role for the hippocampus is as a relational structure (Konkel & Cohen, 2009). The hippocampal ensemble representation of a memory may act as a set of pointers (through specific axonal projections) to the locations of salient associations in other parts of the brain. Thus, ensemble stability within the hippocampus could enable faster, more holistic, or more robust memory of experiences that are often revisited or reactivated.

In sum, we describe a model of memory preservation amidst synapse turnover that accounts for an apparent contradiction emerging from recent research on connectome stability and memory and ensemble code persistence. We show that memory can be stabilized by Hebbian plasticity of weights, even as the network is randomly rewired. We also show that the place fields of individual place cells can be stable despite the ongoing, random replacement of grid cell inputs. This work highlights two open biological questions. First, are hippocampal synapses

independently stabilized in an activity dependent manner? Second, are a subset of hippocampal synapses stable? Our results suggest that neither activity-dependent synapse stabilization nor an inherently stable synapse population is necessary to explain observed phenomena. In the future, it will be interesting to use the methods presented here to ask whether newly generated neurons are likely to be assimilated into preexisting engrams, or if they are solely useful in creating new memories.

Materials and Methods

i. Experimental Design and Statistical Analysis

All simulations and analyses were performed using the R programing language. Required Python-language packages were imported into R sessions using the Reticulate R package. These included NumPy for fast matrix multiplication and scikit-learn for machine learning (Pedregosa et al., 2011). Statistical design for the experiment assessing collective approximation in a feed-forward network as plotted in Figure 2 can be found in Results section 1 and Materials and Methods section 2. Statistical design for experiments in Hopfield models as plotted in Figure 3 can be found in Results section 2 and Materials and Methods section 3. Statistical design for experiments comparing network architectures as plotted in figure 4 can be found in Results section 3 and Materials and Methods section 4. Statistical design for experiments in the E%-max winner-takes-all grid-cell-to-place-cell model can be found in Results section 4 and Materials and Methods section 5.

ii. Networks for assessing collective approximation

a. Architecture and dynamics

Networks consisted of two neuron layers. The pre-synaptic layer consisted of 100 neurons, and the post-synaptic contained just one neuron. The activities of neurons in the pre-synaptic layer were pre-specified and remained constant across activations. Activities of neurons in the first layer were simulated as cosine functions $\boldsymbol{f(x) = \cos(Bx + C)}$ of the sequence of decimal numbers from zero to $\boldsymbol{2\pi}$ in steps of $\boldsymbol{\pi/49.5}$. The period, $\boldsymbol{2\pi/B}$, was randomly selected from the uniform distribution in the range ($\boldsymbol{\pi/4}$, $\boldsymbol{2\pi}$), and the phase shift, $\boldsymbol{-C/B}$, was randomly selected from the uniform distribution in the range (0, $\boldsymbol{2\pi}$). Synaptic weights were all initialized to the value $\boldsymbol{10^{-4}}$ (arbitrary units). The firing rates of neurons in the second layer in response to each of the input patterns were determined by summing inputs as $\boldsymbol{y = Xw}$, where $\boldsymbol{y}$ is a matrix with the firing rate of the post-synaptic neuron for each pattern (columns of $\boldsymbol{y}$ represent distinct cells; rows of $\boldsymbol{y}$ represent distinct input patterns), $\boldsymbol{X}$ is the input matrix (columns of $\boldsymbol{X}$ represent distinct cells; rows of $\boldsymbol{X}$ represent distinct levels of the stimulus in each pattern) and $\boldsymbol{w}$ is the synaptic weight matrix.

A network activation corresponded to projecting each input pattern in the first layer into a corresponding output pattern in the second layer. Weight updates were performed once per activation, taking into account responses to all stimulus patterns. Weights were updated according to Hebb's rule, $\boldsymbol{\Delta w = \eta X^T y}$, where $\boldsymbol{\eta = 10^{-2}}$ is the learning rate. Weight updates were omitted following the first activation in the "Post-training only" condition, during the second activation in the "Pre-training only" condition, and during all activations in the "No training" condition.

Input turnover occurred after the first activation. During turnover, a random 50% of cosine functions in the first layer were replaced with new cosine functions with random phase shift and period. Weights for inputs with new cosine functions were reset to the initial value.

b. *Assessing collective approximation*

To ask whether the sum of PSPs for a given input pattern using new cells matched the sum of PSPs for the same pattern using old cells, we assessed the correlation between the sum of PSPs over new inputs, measured following the weight update on the second activation, with the sum of PSPs over the inputs that would be lost, measured following the weight update on the first activation.

iii. Attractor network models

a. Architecture and dynamics

We modeled sparsely connected attractor networks based on the classic Hopfield model (Hopfield, 1982). Networks were simulated according to Wittenburg, Sullivan, & Tsien's (2002) equations 1 and 2a, reproduced here as Eq. 1 & 2.

$$\tau_u \frac{du_i}{dt} = -u_i + \sum_j w_{ij} \tanh(\beta u_j) + I_i \quad \textbf{Eq. 1}$$

$$\Delta w_{ij} = -\gamma w_{ij} + \eta V_i V_j \quad \textbf{Eq. 2}$$

Eq. 1 describes how neuron $\boldsymbol{i}$'s membrane potential ($\boldsymbol{u_i}$) changes across iterations of the network during a single activation. The term $-\boldsymbol{u_i}$ causes the membrane potential to decay to zero in the absence of input. The term $\sum_{\boldsymbol{j}} \mathbf{w_{ij}}$ describes the change in membrane potential due to synaptic input. We set $\boldsymbol{\beta} = \mathbf{1}$. The term $\boldsymbol{I_i}$ represents external input and is used during training. The membrane time constant ($\boldsymbol{\tau_u}$), the learning rate ($\boldsymbol{\eta}$), and the synaptic weight decay constant ($\boldsymbol{\gamma}$) were set to 1. Eq. 2 describes how the synaptic weights change between activations of the network. The change in weight for the synapse from the $\boldsymbol{j^{th}}$ to the $\boldsymbol{i^{th}}$ neuron ($\boldsymbol{\Delta w_{ij}}$) is calculated on each activation after 12 time steps. The term $-\boldsymbol{\lambda w_{ij}}$ causes the synaptic weights to decay toward zero. This simulates the decay of synaptic receptors. The term $\boldsymbol{\eta V_i V_j}$ describes the Hebbian change in synaptic weight, where $\boldsymbol{\eta}$ is the learning rate, and $\boldsymbol{V_i} = \mathbf{tanh}(\boldsymbol{\beta u_i})$ is the firing rate of neuron $\boldsymbol{i}$.

b. Training

Training consisted of iterative activations of the network while supplying the input image unraveled into a vector as the input current $\boldsymbol{I}$. The training input for a given cell ($\boldsymbol{I_i}$) could take on the value -80 or 80. At the beginning of each training activation, the network was initialized with random membrane potentials in the range (-1, 1) and random synaptic weights in the range (-0.004, 0.004). The network was then updated for 12 iterations according to Eq. 1 (Wittenburg, Sullivan, & Tsien, 2002). Membrane potentials on the last iteration were supplied to Eq. 2 to update synaptic weights before the next activation (Wittenburg, Sullivan, & Tsien, 2002). Training proceeded for 3 activations.

c. Reactivation

Reactivations were performed by initializing the network with random membrane potentials in the range (-0.004, 0.004) and no external input. The network was then updated for 12 iterations according to Wittenburg, Sullivan, & Tsien's (2002) equation 1, reproduced above. Membrane potentials on the last iteration were supplied to Wittenburg, Sullivan, & Tsien's (2002) equation 2a, reproduced above, to update synaptic weights before the next activation. Synapse turnover occurred at the beginning of each reactivation. During turnover, a specified percentage of synapses was replaced. Synapse turnover resulted in the erasure of learned synaptic weights, with newly formed synapses taking on random values in the range (-1, 1).

d. Memory analysis

The quality of a memory at a given activation of the network was assessed as the coefficient of determination ($\boldsymbol{r^2}$) relating the vector of membrane potentials on the final iteration of that activation to the external input vector ($\boldsymbol{I}$) used for training that memory. To assess the change in memory quality over time, we compared the $\boldsymbol{r^2}$ values of individual networks on reactivations one and 100. To derive the confidence intervals shown in Figure 3A,B,C, we performed locally weighted regression using the "loess" function in the R Stats package with span set to 0.25 and otherwise default parameters.

iv. Networks for architectures comparisons

a. Architecture and dynamics

Networks consisted of two neuron layers. The pre-synaptic layer consisted of 100 neurons, and the post-synaptic contained 1000 neurons. The activities of neurons in the pre-synaptic layer were pre-specified and remained constant across activations or were cycled in the case of multiple memories presented sequentially. The firing rates of neurons in the first layer were represented as input pattern vectors of values sampled from the uniform distribution in the range (-1, 1). The first layer was sparsely connected to the second layer with a $\boldsymbol{p_{connection}}$ of 0.2. Synaptic weights were initialized to random values in the range (0, 1). For identity networks, firing rates of neurons in the second layer were determined by summing inputs as $\boldsymbol{y = Xw}$, where $\boldsymbol{y}$ is a matrix containing the firing rate of the post-synaptic neuron (columns of $\boldsymbol{y}$ represent distinct cells; rows of $\boldsymbol{y}$ represent distinct input patterns), $\boldsymbol{X}$ is the input matrix (columns of $\boldsymbol{X}$ represent distinct cells; rows of $\boldsymbol{X}$ represent distinct input patterns or experiences) and $\boldsymbol{w}$ is the synaptic weight matrix. For winner-takes-all networks, firing rates of neurons in the second layer were determined by summing inputs, then setting firing rates to zero when they not in the top 10%. For E%-max winner-takes-all networks, firing rates of neurons in the second layer were determined by summing inputs, then setting firing rates to zero when they within 10% of the most excited cell.

A network activation was defined as a projection of each input pattern in the first layer into an output response in the second layer. Weight updates were performed once per activation, corresponding to once following the experience of all patterns. In the case of interleaved inputs, weights were updated once for each input pattern, presented sequentially. Weights were updated

according the learning rule $\boldsymbol{\Delta w = -w + \tanh(w + \eta X^T y)}$, where $\boldsymbol{\eta = 0.1}$. Here, the term $\boldsymbol{-w}$ causes synaptic weights to decay toward zero in the absence of learning, and the hyperbolic tangent function limits the range of possible synaptic weights to between one and negative one.

Synapse turnover occurred at the beginning of each input presentation (excluding the final activation). During turnover, a random 10% of synapses were replaced. In the case of interleaved inputs, the turnover rate was adjusted such that 10% turnover would be achieved following each presentation of the complete sequence of inputs. The adjusted turnover rate ($\boldsymbol{T_{adjusted}}$) was calculated as

$$\boldsymbol{T_{adjusted} = 1 - \exp\left(\frac{\log(1-T_{base})}{n_{patterns}}\right)} \quad \textbf{Eq. 3},$$

where $\boldsymbol{n_{patterns}}$ is the number of input patterns and $\boldsymbol{T_{base}}$ is the unadjusted turnover rate of 10% per activation. Synapse turnover resulted in the erasure of learned synaptic weights, with newly formed synapses taking on random values in the original range.

b. Memory analysis

The activities of cells in the second layer were taken after reactivations one and 100. Correlation matrices were calculated to compare post-synaptic responses to each input pattern between these activations such that individual cells were covariates and input patterns were samples. To assess the quality of a preserved memory, we extracted response autocorrelations, i.e. the values on the diagonal of the correlation matrix. To assess memory discriminability, we calculated a uniqueness score as $\boldsymbol{v_i = r_{i,i} - \vec{r}_{i,\setminus i}^{max}}$, where $\boldsymbol{v_i}$ is the uniqueness of memory $\boldsymbol{i}$, $\boldsymbol{r_{i,i}}$ is

the autocorrelation of response $\boldsymbol{i}$, and $\vec{\mathbf{r}}_{\boldsymbol{i},\backslash\boldsymbol{i}}^{\boldsymbol{max}}$ is the maximum value of the vector of correlations between response $\boldsymbol{i}$ on reactivation 100 with all responses excluding $\boldsymbol{i}$ on reactivation one.

v. Grid-cell-to-place-cell model

a. Grid cells

Data were simulated by assuming a 1 m linear enclosure divided into 1 cm bins. The activity of each cell was characterized by its firing rate in each bin. We simulated a library of 10000 grid cell responses according to a method described by Blair et al. (2007).

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

Here, $\boldsymbol{G}$ is a grid cell's firing rate, r is the animal's position in 2-dimensional space, $\boldsymbol{\lambda}$ is the distance between grid vertices and ranged from 30 to 100 cm, $\boldsymbol{\theta}$ is the angular offset and ranged from $\mathbf{0}^\circ$ to $\mathbf{60}^\circ$, and $\boldsymbol{c}$ is the offset in 2-dimensional space and ranged from zero to 100 cm in both dimensions. $\boldsymbol{g}$ is a gain function $\boldsymbol{g(x) = \exp[a(x-b)] - 1}$, where $\boldsymbol{a}$ modulates the spatial decay and was set to 0.3, and $\boldsymbol{b}$ modulates the minimum firing rate and was set to $\boldsymbol{-3/2}$. The hexagonal grid is created by summing cosine gratings angled at $\boldsymbol{\theta_1 = -30^\circ}$, $\boldsymbol{\theta_2 = 30^\circ}$, and $\boldsymbol{\theta_3 = 90^\circ}$. $\boldsymbol{u}$ is the function $\boldsymbol{u(\theta_k) = (\cos(\theta_k), \sin(\theta_k))}$.

b. Place cells

We simulated 2000 place cells as described by de Almeida, Idiart, & Lisman (2009a). Each cell received and summed excitatory input from 1200 randomly selected grid cells. The activity of each place cell was determined in each 1 cm bin of the 1 m linear enclosure. The sum of input to a place cell was calculated as

$$\boldsymbol{I^{i}_{grid}(r) = \vec{G}(r) \cdot \vec{w}_{i}} \quad \textbf{Eq. 5},$$

where $\boldsymbol{I^{i}_{grid}(r)}$ is the input to the $\boldsymbol{i^{th}}$ place cell at position $\boldsymbol{r}$, $\boldsymbol{\vec{w}_{i}}$ is the weight vector representing grid cell synapses onto $\boldsymbol{i^{th}}$ place cell, and $\boldsymbol{\vec{G}(r)}$ is the vector of all grid cell firing rates at position $\boldsymbol{r}$. The firing rate of a place cell was calculated as

$$\boldsymbol{F(r) = I_{grid}(r) \cdot H\left(I_{grid}(r) - (1-k) \cdot I^{max}_{grid}(r)\right)} \quad \textbf{Eq. 6},$$

where $\boldsymbol{F(r)}$ is the place cell's firing rate at position *r*, $\boldsymbol{H}$ is the Heaviside function, and $\boldsymbol{I^{max}_{grid}(r)}$ is the sum of excitatory input received by the most excited place cell at position $\boldsymbol{i}$. E%-max parameter $\boldsymbol{k}$ is the fraction of $\boldsymbol{I^{max}_{grid}(r)}$ to which a place cell must be excited in order to fire. We set $\boldsymbol{k = 0.92}$.

c. Synaptic weights and learning

Synaptic weights were all initialized to the value $\boldsymbol{1/1200}$. Weight updates occurred once per activation. Weights were updated according to the following rule set.

$$\Delta w_{i,j} = \begin{cases} \Psi_{i,j} & \text{if } \frac{2}{1200} \geq w_{i,j} + \Psi_{i,j} \geq 0 \\ -w_{i,j} & \text{if } w_{i,j} + \Psi_{i,j} < 0 \\ -w_{i,j} + \frac{2}{1200} & \text{if } w_{i,j} + \Psi_{i,j} > \frac{2}{1200} \end{cases} \quad \textbf{Eq. 7}$$

$$\Psi_{i,j} = \begin{cases} \Lambda_{i,j} & \text{if } \rho = \text{LTP-LTD} \\ 0 & \text{if } \rho = \text{LTP-only and } \Lambda_{i,j} < 0 \\ 0 & \text{if } \rho = \text{LTD-only and } \Lambda_{i,j} > 0 \\ 0 & \text{if } \rho = \text{No learning} \end{cases} \quad \textbf{Eq. 8}$$

$$\Lambda_{i,j} = \begin{cases} \Upsilon_{i,j} & \text{if } G_j \geq \overline{G_j} \text{ or } F_i \geq \overline{F_i} \\ -\Upsilon_{i,j} & \text{if } G_j < \overline{G_j} \text{ and } F_i < \overline{F_i} \end{cases} \quad \textbf{Eq. 9}$$

$$\Upsilon_{i,j} = \sum_{r} \eta \left(G_j(r) - \overline{G_j}\right)\left(F_i(r) - \overline{F_i}\right) \quad \textbf{Eq. 10}$$

Eq. 7 prevents synaptic weights from falling below zero or growing beyond twice the original weight. Eq. 8 implements variations in the learning rule (ρ), a factor variable taking on the value "LTP-LTD" to allow both potentiation and depression, "LTP-only" to allow only potentiation, "LTD-only" to allow only depression, or "No learning" to allow no potentiation or depression. Eq. 9 and Eq. 10 are a modification of Hebb's rule. Potentiation occurs when the pre-synaptic firing rate at position r ($G_j(r)$) is greater than or equal to the mean pre-synaptic firing rate ($\overline{G_j}$) and the post-synaptic firing rate at position r ($F_i(r)$) is greater than or equal to the mean post-synaptic firing rate ($\overline{F_i}$), otherwise depression occurs. The learning rate (η) was set to 2×10^{-5}.

d. Synapse turnover

Synapse turnover occurred at the beginning of each trial. Of the 1200 active grid cell inputs, a random 114 were replaced. The number 114 corresponds to $\boldsymbol{N_0 - N(t)}$ in the exponential decay model (Eq. 11) assuming a mean synapse lifetime ($\boldsymbol{\tau}$) of 10 days in our paradigm with the inter-trial time ($\boldsymbol{t}$) taken to be one day. This is derived from the exponential decay model

$$\boldsymbol{N(t) = N_0 e^{\frac{-t}{\tau}}} \quad \textbf{Eq. 11},$$

where $\boldsymbol{N(t)}$ is the number of original synapses remaining after $\boldsymbol{t}$ days, $\boldsymbol{N_0 = 1200}$ is the number of synapses on day zero. Synapse turnover resulted in the erasure of learned synaptic weights, with newly formed synapses taking on the initial weight value.

e. Place field analysis

Place cells were said to have a place field on a given trial if there was exactly one continuous region in the enclosure, at least 5 cm in length, where their firing rate was within 80% of their maximum firing rate on the same trial. The place field shift on a given trial ($\boldsymbol{\Omega_t}$) was calculated for all place cells with place fields both on trial $\boldsymbol{t}$ and trial zero as the absolute centroid offset $\boldsymbol{\Omega_t = |C_t - C_0|}$, where $\boldsymbol{C_t}$ is the cell's place field centroid position on trial $\boldsymbol{t}$ and $\boldsymbol{C_0}$ is the position on day zero.

f. Trajectory decoding

Trajectory decoding was performed using the scikit-learn Python library implementation of gradient-boosted trees (Pedregosa et al., 2011). The gradient-boosted trees classifier (GBTC) was initialized with default parameters. The matrix of place cell firing rates by position was extracted from day zero simulations and filtered to remove non-place cells. This filtered matrix was scaled so that features had a mean of zero and a standard deviation of one, then used as training input to the GBTC such that cells were covariates and positions were samples. The true simulated position was used as the training outcome. For position prediction, the activities of the same cells used for training were extracted from each day, scaled as described, and separately supplied to the GBTC. A prediction for position was returned for each position sample on each day. Decoding error was calculated as the mean absolute value of the difference between the predicted position and the true position across all predictions for a given simulated trajectory.

References

Abraham WC, Logan B, Greenwood JM, Dragunow M (2002) Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. J Neurosci 22(21):9626-9634.

Agnihotri NT, Hawkins RD, Kandel ER, Kentros C (2004) The long-term stability of new hippocampal place fields requires new protein synthesis. Proc Natl Acad Sci U S A 101(10):3656-3661.

Andersen P, Bliss TV, Skrede KK (1971) Lamellar organization of hippocampal pathways. Exp Brain Res 13(2):222-238.

660 Attardo A, Fitzgerald JE, Schnitzer MJ (2015) Impermanence of dendritic spines in live
661 adult CA1 hippocampus. Nature 523(7562):592-596.
662 Blair HT, Welday AC, Zhang K (2007) Scale-invariant memory representations emerge
663 from moiré interference between grid fields that produce theta oscillations: a computational
664 model. J Neurosci 27(12):3211-3229.
665 Bourne JN, Harris KM (2011) Coordination of size and number of excitatory and
666 inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1
667 dendrites during LTP. Hippocampus 21(4):354-373.
668 Bramham CR, Alme MN, Bittins M, Kuipers SD, Nair RR, Pai B, Panja D, Schubert
669 M, Soule J, Tiron A, Wibrand K (2010) The Arc of synaptic memory. Exp Brain Res
670 200(2):125-140.
671 Chambers AR, Rumpel S (2017) A stable brain from unstable components: Emerging
672 concepts and implications for neural computation. Neuroscience 357:172-184.
673 Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of
674 the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
675 pp785-794.
676 Cohen M, Grossberg S (1983) Absolute stability of global pattern formation and parallel
677 memory storage by competitive neural networks. IEEE Trans Syst Man and Cyber 13(5):815-
678 826.
679 de Almeida L, Idiart M, Lisman JE (2009a) The input-output transformation of the
680 hippocampal granule cells: from grid cells to place fields. J Neurosci 29(23):7504-7512.

de Almeida L, Idiart M, Lisman JE (2009b) A second function of gamma frequency oscillations: an E%-max winner-take-all mechanism selects which cells fire. J Neurosci 29(23):7497-7503.

de Almeida L, Idiart M, Lisman JE (2012) The single place fields of CA3 cells: a two-stage transformation from grid cells. Hippocampus 22(2):200-208.

DeBello W, Zito K (2017) Within a Spine's Reach. In: The Rewiring Brain: A Computational Aproach to Structural Plasticity in the Adult Brain (van Ooyen A, Butz-Istendorf M, ed), pp295-317. Amsterdam, Netherlands: Elsevier Science.

Eppler B, Aschauer D, Rumpel S, Kaschube M (2015) Discrete cortical representations and their stability in the presence of synaptic turnover. BMC Neuroscience 16(S1):114.

Fauth M, Worgotter F, Tetzlaff C (2015) Formation and Maintenance of Robust Long-Term Information Storage in the Presence of Synaptic Turnover. PLoS Comput Biol 11(12):e1004684.

French RM (1999) Catastrophic forgetting in connectionist networks. Trends Cogn Sci 3(4):128-135.

Gallinaro JV, Rotter S (2017) Associative properties of structural plasticity based on firing rate homeostasis in recurrent neuronal networks. arXiv:1706.02912v1.

Giachero M, Calfa GD, Molina VA (2015) Hippocampal dendritic spines remodeling and fear memory are modulated by GABAergic signaling within the basolateral amygdala complex. Hippocampus 25(5):545-555.

Glaser JI, Chowdhury RH, Perich MG, Miller LE, Kording KP (2017) Machine learning for neural decoding. arXiv:1708.00909.

Grutzendler J, Kasthuri N, Gan WB (2002) Long-term dendritic spine stability in the adult cortex. Nature 420(6917):812-816.

Guskjolen A, Josselyn SA, Frankland PW (2017) Age-dependent changes in spatial memory retention and flexibility in mice. Neurobiol Learn Mem 143:59-66.

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44(1):1-12.

Herzog MH, Clarke AM (2014) Why vision is not both hierarchical and feedforward. Front Comput Neurosci 8:135.

Hill TC, Zito K (2013) LTP-induced long-term stabilization of individual nascent dendritic spines. J Neurosci 33(2):678-686.

Holtmaat AJ, Trachtenberg JT, Wilbrecht L, Shepherd GM, Zhang X, Knott GW, Svoboda K (2005) Transient and persistent dendritic spines in the neocortex in vivo. Neuron 45(2):279-291.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 79(8):2554-2558.

Kentros CG, Agnihotri NT, Streater S, Hawkins RD, Kandel ER (2004) Increased attention to spatial context increases both place field stability and spatial memory. Neuron 42(2):283-295.

Knoblauch A, Korner E, Korner U, Sommer FT (2014) Structural synaptic plasticity has high memory capacity and can explain graded amnesia, catastrophic forgetting, and the spacing effect. PLoS One 9(5):e96485.

Konkel A, Cohen NJ (2009) Relational Memory and the Hippocampus: Representations and Methods. Front Neurosci 3(2):166-174.

Lisman J, Morris RG (2001) Memory: Why is the cortex a slow learner? Nature 411(6835):248-249.

Lisman JE, Grace AA (2005) The hippocampal-VTA loop: controlling the entry of information into long-term memory. Neuron 46(5):703-713.

Lüscher C, Malenka RC (2012) NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD). Cold Spring Harb Perspect Biol 4(6):a005710.

Mongillo G, Rumpel S, Loewenstein Y (2017) Intrinsic volatility of synaptic connections - a challenge to the synaptic trace theory of memory. Curr Opin Neurobiol 46:7-13.

Moser MB, Trommald M, Andersen P (1994) An increase in dendritic spine density on hippocampal CA1 pyramidal cells following spatial learning in adult rats suggests the formation of new synapses. Proc Natl Acad Sci U S A 91(26):12673-12675.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. JMLR, 12:2825-2830.

Poirazi P, Mel BW (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron 29(3):779-796.

Richards B, Frankland P (2017) The Persistence and Transience of Memory. Neuron 94(6):1071-1084.

Roxin A, Fusi S (2013) Efficient partitioning of memory systems and its importance for memory consolidation. PLoS Comput Biol 9(7):e1003146.

Rubin A, Geva N, Sheintuch L, Ziv Y (2015). Hippocampal ensemble dynamics timestamp events in long-term memory. eLife 4:e12247.

Thompson LT, Best PJ (1990) Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats. Brain Res 509(2):299-308.

Trachtenberg JT, Chen BE, Knott GW, Feng G, Sanes JR, WelkerE, Svoboda K (2002) Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. Nature 420(6917):788-794.

Wittenberg GM, Sullivan MR, Tsien JZ (2002) Synaptic reentry reinforcement based network model for long-term memory consolidation. Hippocampus 12(5):637-647.

Yang G, Pan F, Gan WB (2009) Stably maintained dendritic spines are associated with lifelong memories. Nature 462(7275):920-924.

Ziv Y, Burns LD, Cocker ED, Hamel EO, Ghosh KK, Kitch LJ, El Gamal A, Schnitzer MJ (2013) Long-term dynamics of CA1 hippocampal place codes. Nat Neurosci 16(3):264-266.

Legends

Figure 1. Synaptic weight adjustments permit deep layer neurons to maintain stimulus-response tuning despite turnover in superficial layers. A two-layer neural network is shown receiving sensory input. Individual curves represent a unit's firing rate over time. (**A**) The network prior to turnover. (**B**) The network following turnover. (**A,B**) During turnover, the layer 1 units in red are disconnected from the network, while new layer 1 units, in blue, are connected. Synaptic weights are adjusted after turnover. The layer 2 unit's firing rate retains its relationship to the sensory input.

Figure 2. Hebbian plasticity drives new synapses to approximate lost synapses. Correlation of summed PSPs across lost inputs after pre-training with summed PSPs across new inputs after post-training. N=500 simulations per condition.

Figure 3. SRR maintains memory in attractor networks despite synapse turnover. Sparsely connected attractor networks were trained on an image (**A: inset**) and subjected to 100 iterations of reactivation and synapse turnover. The coefficients of determination (r^2) between the attractors and the input were calculated on the first and final reactivation. (**A**) Fraction of synapses replaced per activation (turnover fraction) versus r^2 when $p_{connection}$ and the number of units were constant. (**B**) $p_{connection}$ versus r^2 when turnover rate and number of units were constant. (**C**) Number of units versus r^2 when $p_{connection}$ and turnover rate were constant. (**A,B,C**) Shaded areas represent 95% confidence intervals of locally weighted regression models. N=100 networks per panel. (**D,E,F**) Representative firing rates of networks shown in **A,B,C**. (**G**) r^2 on

reactivation 100 versus turnover fraction and median in-degree when number of units, $p_{connection}$, and turnover rate varied. (**H**) The change in membrane potential (dV_m) versus time step for units in three networks, one with turnover but stable memory, and one with turnover and unstable memory, and one with no turnover. Black triangles mark the onset of reactivation events.

Figure 4. Winner-takes-all network dynamics enhance the stability of multiple memories during turnover. Feed-forward (Identity), winner-takes-all, and E%-max winner-takes-all (E%-max) networks were simulated and underwent 100 iterations of learning in response to 5 to 100 input patterns and synapse turnover at a rate of 0% or 10% per iteration. (**A**) Memory preservation (autocorrelation) versus number of input patterns by network architecture when weights were updated simultaneously after presentation of all input patterns or inputs were presented successively, weights were updated once for each pattern, and updates were interleaved with turnover. (**B**) Memory uniqueness versus number of input patterns by network architecture when weights were updated simultaneously after presentation of all input patterns or inputs were presented successively, weights were updated once for each pattern, and updates were interleaved with turnover. (**A,B**) N=10 independent simulations and 50 to 1000 units per point. (**C**) Representative correlation (r) matrices comparing responses to five input patterns on reactivations one and 100 when weights were updated simultaneously after presentation of all input patterns or weights were updated once for each pattern and updates were interleaved with turnover. The turnover rate was 10%.

Figure 5. Stable place fields amidst synapse turnover in an E%-max winner-takes-all model. Activity in an E%-max winner-takes-all grid-cell-to-place-cell model was simulated over

a period of 30 days with both learning and synapse turnover. (**A**) Time versus mean place field shift (|Offset|) by learning method. (**B**) Time versus total number of place cells with valid place fields (PCs). (**C**) Time versus total number of PCs that also had place fields on day zero. (**D**) Representative place cell firing rates sorted by place field centroid on day zero. (**E**) Cumulative density plot of the correlation between the within position sums of PSPs at new and lost synapses between days 29 and 30 in the "LTP & LTD" condition. (**F**) Time versus mean error in decoded position. N=10 simulations per condition.

Table 1. Parameters for SRR experiments. For experiments involving Hopfield networks and synaptic reentry reinforcement (SRR), as plotted in figure 3, network parameters were sampled from the ranges shown.

Illustrations and Tables

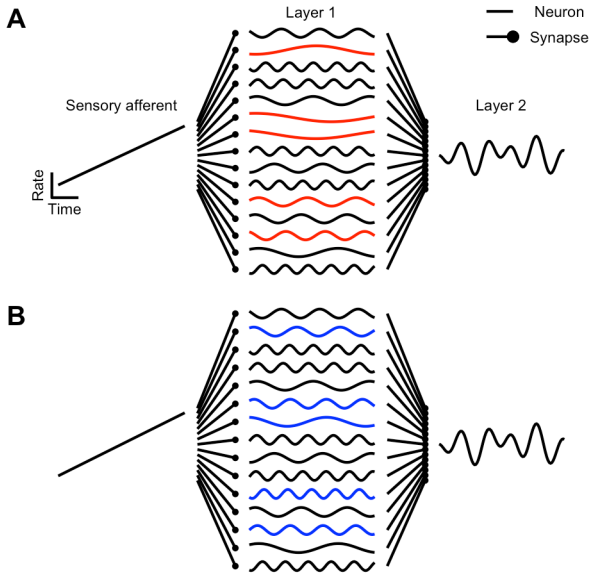


Figure 1. Synaptic weight adjustments permit deep layer neurons to maintain stimulus-response tuning despite turnover in superficial layers.

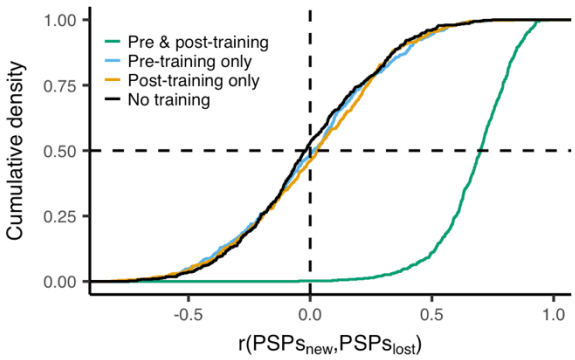
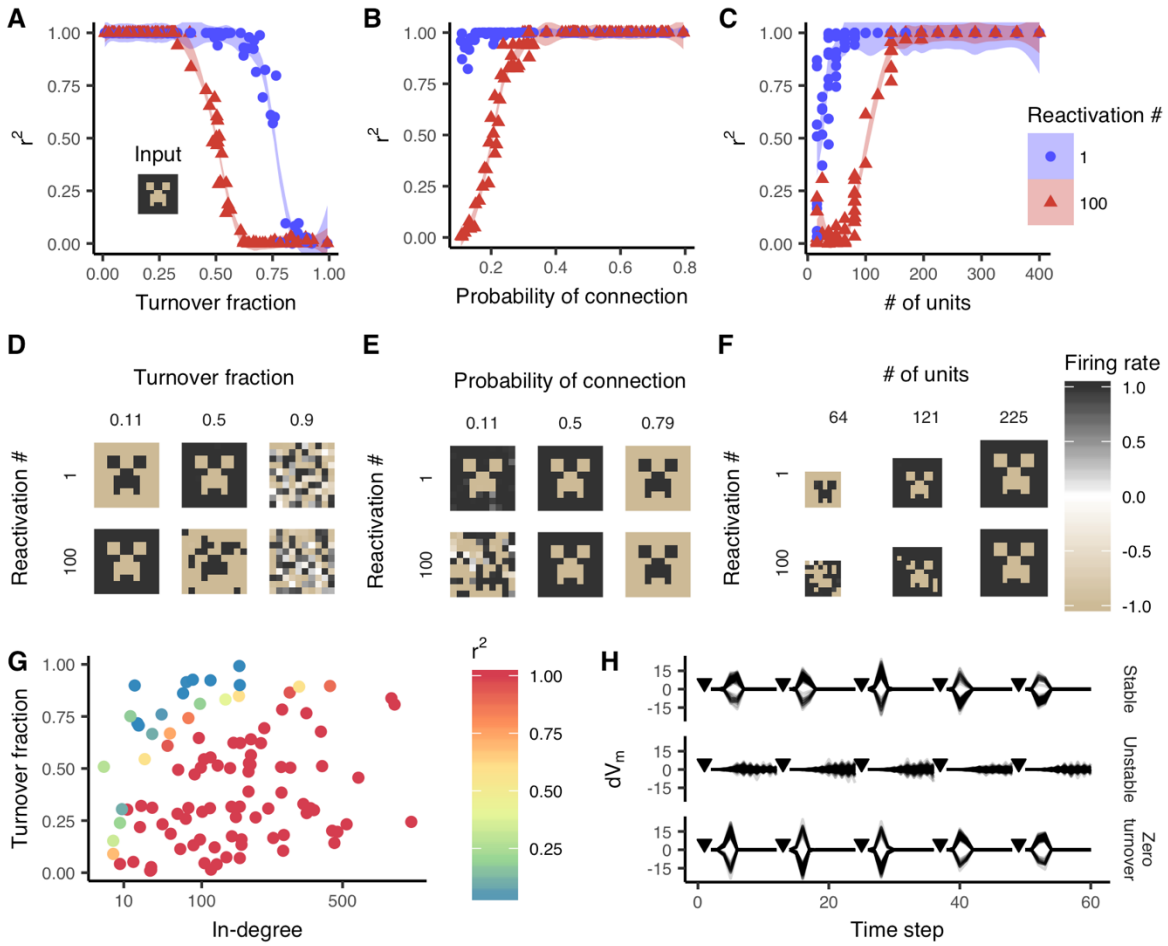


Figure 2. Hebbian plasticity drives new synapses to approximate lost synapses.



856

857 **Figure 3. SRR maintains memory in attractor networks despite synapse turnover.**

858

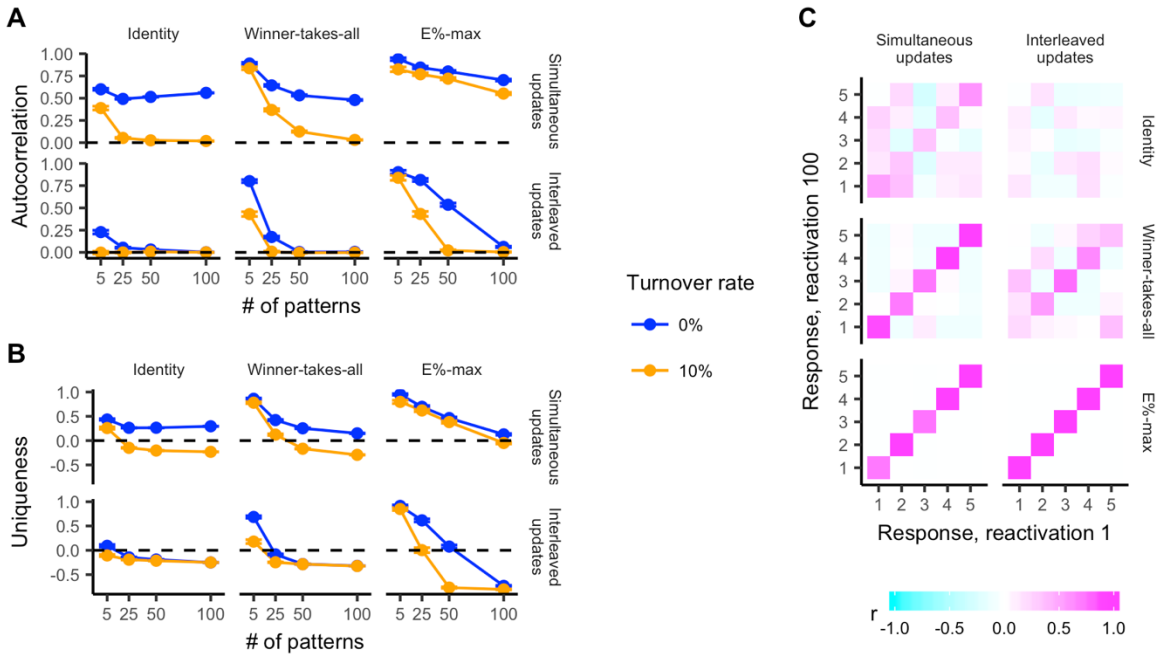
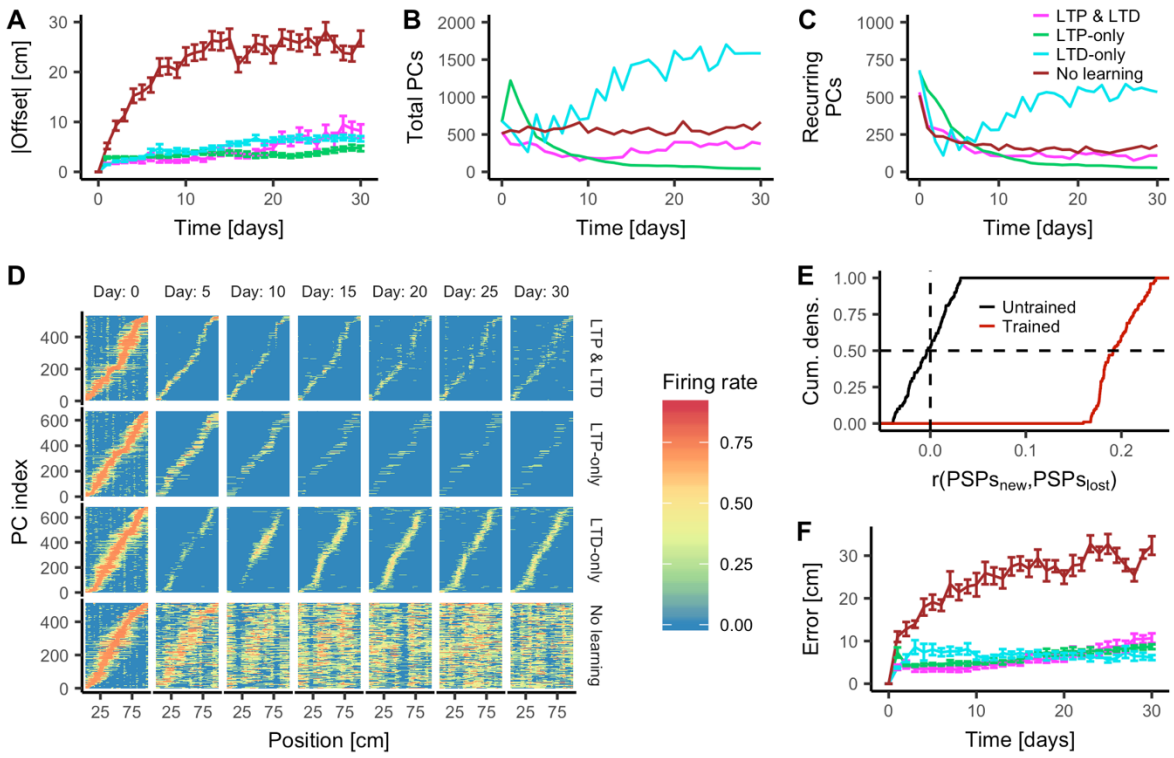


Figure 4. Winner-takes-all network dynamics enhance the stability of multiple memories during turnover.



867

868 **Figure 5. Stable place fields amidst synapse turnover in an E%-max winner-takes-all**

869 **model.**

870

871

872

873

874

875

876

877

878 **Table 1. Parameters for SRR experiments.**

Experiment	Figure	Turnover rate	Probability of	# of units
1	A,D	[0%, 100%]	0.2	100
2	B,E	50%	(0.1, 0.8)	100
3	C,F	50%	0.2	Perfect squares in (16, 225)
4	G	[0%, 100%]	(0.1, 0.8)	Perfect squares in (16, 1024)

879