Representational drift as the consequence of ongoing memory storage

Federico Devalle, Licheng Zou, Gloria Cecchini and Alex Roxin
Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra (Barcelona), Spain

Memory systems with biologically constrained synapses have been the topic of intense theoretical study for over thirty years. Perhaps the most fundamental and far-reaching finding from this work is that the storage of new memories implies the partial erasure of already-stored ones. This overwriting leads to a decorrelation of sensory-driven activity patterns over time, even if the input patterns remain similar. Representational drift (RD) should therefore be an expected and inevitable consequence of ongoing memory storage. We tested this hypothesis by fitting a network model to data from long-term chronic calcium imaging experiments in mouse hippocampus. Synaptic turnover in the model inputs, consistent with the ongoing encoding of new activity patterns, accounted for the observed statistics of RD. This mechanism also provides a parsimonious explanation for the recent finding that RD in CA1 place cells has two distinct components: one which depends only on the passage of time, and another which depends on the time spent exploring a given environment. Furthermore, in the context of ongoing learning, the drift rate of any one memory depends on its repetition rate, a mechanism which can reproduce the diverse effects of experience on drift found in experiment. Our results suggest that RD should be observed wherever neuronal circuits are involved in a process of ongoing learning or memory storage.

The synaptic hypothesis of learning and memory postulates that memories are stored in the structure of the synaptic weight matrix in neuronal circuits[1]. Theoretical studies have shown that network models built on this synaptic principle can store large numbers of patterns as fixed point attractors[2–4]. However, with biologically constrained synapses, new learning implies the overwriting of previously stored memories[5,6]. In the context of stimulus-driven neuronal activity, this overwriting would result in a distinct network response at two different points in time, even if the input pattern remained unchanged.

The advent of technologies allowing for stable, long-term recordings in awake behaving mice has indeed revealed that the neuronal activation underlying certain behaviors can change dramatically, even when sensory inputs do not. This phenomenon was first described in detail in the context of spatial memory in rodents[7]. It was already known that in area CA1 of the hippocampus a unique pattern of place-cell activity quickly emerged upon exploration of a novel space[8–12], and was reliably re-evoked when the animal was returned to the familiar environment[13–15]. However, the seeming stability of the hippocampal code only held true on relatively short time scales. Indeed, in familiar environments, both place-cell- and non place-cell activity slowly changed over days and weeks[7,16–23]. These long time-scale changes in the neuronal code, dubbed representational drift (RD), have also been seen in other cortical areas, such as parietal, piriform, visual and auditory cortex[24–28]. The phenomenology of RD is fundamentally similar in all of these cases: the identity of active neurons changes from session to session, although the sparseness of representation is stable. Furthermore, there is considerable heterogeneity in the stability of cells, or how often they take part in the code. It is not known what the mechanism generating the observed drift is, nor what its potential functional role might be.

Several computational studies have shown that ongoing plasticity can result in changes in neuronal dynamics reminiscent of RD at the population level[29–32]. In these studies the plasticity acted as a source of noise, driving changes in the representation of already stored patterns. In fact, it was previously hypothesized that such changes might provide the substrate for a time-stamp of a given memory[16]. Alternatively, it has also been hypothesized that RD may occur due to plasticity related to the encoding of new memories[25]. From this perspective RD would be the unavoidable signature of ongoing memory storage due to overwriting[5,33,34]. However, it remains unclear to what extent the mechanism of RD as ongoing memory storage is consistent with experimentally observed data. We therefore developed a biologically plausible network model which could be fit to data quantitatively, and which specifically allowed us to address the role of ongoing learning.

Through fitting of our network model to experimental data[16], we inferred that changes in neuronal activity from session-to-session are inherited from changes in the afferent inputs to cells in CA1. These

changes are consistent with the large degree of synaptic turnover observed through Ca2+ imaging of dendritic spines in CA1[35]. Interestingly, the inferred synaptic turnover was random and uncorrelated from cell-to-cell, and followed simple Gaussian statistics. Although the types of memories which the hippocampus stores are, in general, highly correlated, an efficient means of data compression is to store just the differences between like memories, which are uncorrelated[36–40]. Indeed, we show in our network model that the ongoing storage of such decorrelated patterns can account for the synaptic turnover, which, in turn, is responsible for RD.

This same mechanism provides a parsimonious explanation for a recent finding on RD in CA1 place cells. While RD in the spatial tuning of these cells depended on the amount of time spent exploring the environment under study, RD in overall rates depended only on the absolute passage of time[41]. In the context of memory storage, when the animal is engaged in the task, plasticity can readily occur between the active CA1 place cells and their presynaptic inputs, affecting CA1 place cell tuning. On the other hand, episodes unrelated to experimental sessions will not engage the same subset of place cells and changes due to plasticity will therefore be largely spatially untuned (in the experimental environment). The number of encoded episodes is simply proportional to the total time elapsed. In this way, RD in neuronal firing rates occurs as a function of time, while RD in neuronal tuning is biased more strongly towards time spent exploring, when the relevant assemblies of tuned cells are active. Overall, we find that the statistics and phenomenology of RD are well predicted by theories of ongoing memory storage.

Results

Ongoing memory storage in a network model with biologically constrained synapses leads to the partial overwriting of previously encoded patterns, Fig.1. When such a network is driven by external input, the resulting patterns of activity will therefore change over time, even if the input is stable, Fig.1b-c. Such a drop in correlation of network activity over time, or representational drift (RD), has been observed in the hippocampus and other cortical areas. However, despite the broad qualitative similarity between plasticity-related overwriting and RD, it remains unclear if synaptic plasticity alone can provide a comprehensive account of the detailed statistical characteristics associated with RD. To answer this question, we fit a network model to experimental data from chronic Ca2+imaging in mouse hippocampus[16] and sought to reproduce the observed RD via changes in synaptic connectivity consistent with ongoing memory storage. We found differential effects of the spatially tuned versus untuned components of RD in the data, see Extended Fig.1, in line with recent findings[41]. Specifically, the large drop in the correlation of the population activity from one session to the next was due largely to rate effects alone, while drift in the spatial tuning occurred on a longer time scale.

A spiking network model with synaptic turnover reproduces drift statistics in CA1

We sought to reproduce the observed RD in a network model in which plasticity occurred at different rates at spatially tuned versus non-spatially tuned synapses. Specifically, we modelled a local circuit of CA1 pyramidal and interneurons as sparsely-connected leaky integrate-and-fire neurons. Neurons in CA1 received input from area CA3 of the hippocampus and the entorhinal cortex (EC), modeled as Poisson neurons, Fig.2a (top). For simplicity we chose a fraction of CA3 cells to be place cells, while the EC inputs were spatially untuned. Allowing for weak to moderate tuning in the EC inputs did not alter the results qualitatively, see Extended Fig.4. We simulated the movement of a virtual animal along a circular track during a given session, Fig.2a (bottom). Large excitatory and inhibitory input currents dynamically balanced in the model[42], leading to the emergence of heterogeneous spatial tuning, Fig.2b,c and Extended Fig.3a-g.

We modelled changes in the inputs to CA1 cells from session to session as a process of random synaptic turnover[35]. Later we will show that this turnover is also consistent with the ongoing storage of memories. Specifically, we rewired a random fraction of the inputs for each cell in CA1, independently for the EC and CA3 pathways, Fig.2d. The rewiring fraction for CA3 was chosen to be smaller than that

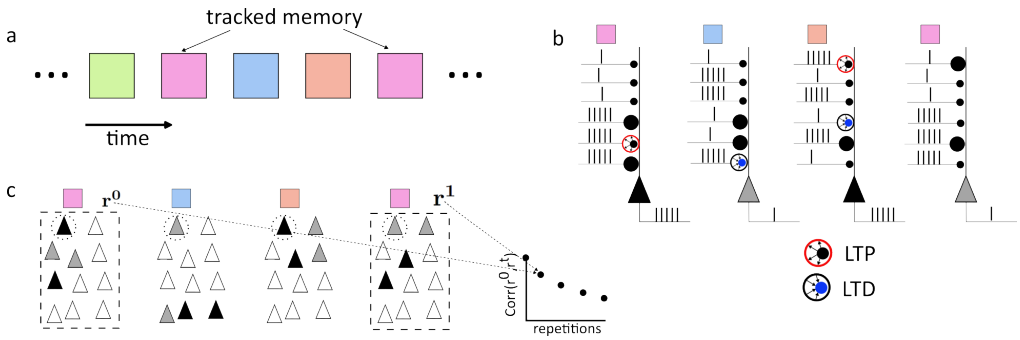


Fig. 1 |Ongoing memory storage generates representational drift. **a.** Memories are encoded in an ongoing fashion in time. Memory identity is indicated by color. We track and observe the neuronal activity corresponding to one memory in particular (pink square). **b.** Synaptic weights undergo plasticity as memories are stored. When pre- and post-synaptic activity is high, synapses may undergo potentiation (large red circle), while depression takes place if one cell fires strongly and the other only weakly (small blue circle). Active presynaptic cells are indicated by a train of action potentials. The postsynaptic cell fires strongly (black) or weakly (grey). Synaptic weights are strong (large circle) or weak (small circle). Note that due to the plasticity from the two intervening memories, the response of the post-synaptic cell to the same pre-synaptic pattern of activity, corresponding to the pink memory, has changed: its firing rate has decreased. **c.** The change in post-synaptic activity due to ongoing memory storage manifests itself at the population level as representational drift. Therefore, there is a drop in correlation between the initial pattern of neuronal activity given inputs corresponding to the pink memory r^0 and the first repetition r^1. If we assume that ongoing memory storage occurs between every repetition the correlation will continue to decrease.

for EC, resulting in more gradual changes in place-cell tuning, in accordance with the data, Extended Fig.1. This synaptic turnover could lead to significant changes in the mean drive to CA1 cells from session to session, which is the mechanism responsible for RD, Fig.2e,f.

We chose the network parameters in order to closely match the statistics of RD observed in experiment. To do this, we first considered a simple statistical model for CA1 pyramidal cells which could be quantitatively fit to the data using standard least-square optimization, see Methods and Extended Fig.2. Inputs in this model were Gaussian random variables with zero mean, and with variance and temporal correlation optimized to fit RD statistics from the data. Input statistics in the full spiking network model were also Gaussian in the balanced regime, and could be calculated analytically, allowing us to map network parameters onto the statistical model, and hence fit the network to the data, see Methods and Extended Fig.3h. As a result, we were able to reproduce the drop in the Population Vector (PV) correlation seen in the data quantitatively Fig.2g, which takes both rate and tuning effects into account, as well as several measures of RD related only to active versus inactive cells, see Extended Fig.3i. The network model also qualitatively reproduced the gradual diffusion of place field location, Fig.2h, although the fraction of place cells was always larger in the network than in the data.

Finally, the mechanism of RD through synaptic turnover predicts that the likelihood of a cell to remain active on a subsequent session should be positively correlated with its firing rate. This is because high (low) firing rates are due to a large (small) number of excitatory inputs. Therefore, the rewiring of the same number of inputs is more likely to cause a low-rate cell to become inactive, than a high-rate one. We reanalyzed the data and found such a positive correlation, see Extended Fig.5.

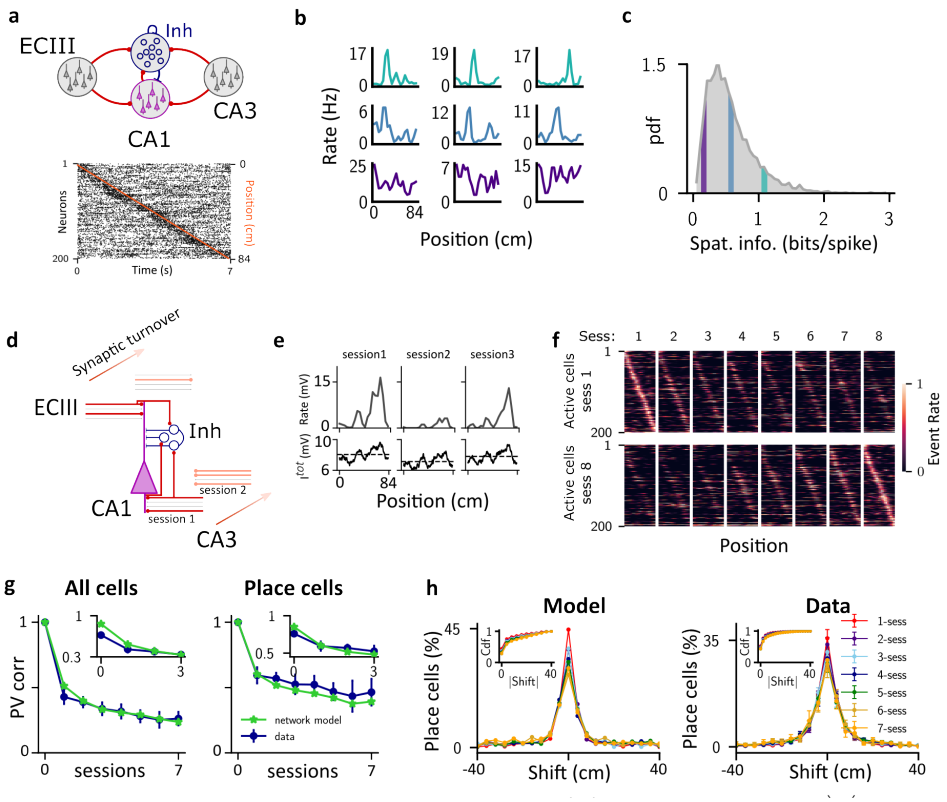


Fig. 2 |A spiking network model with random synaptic turnover reproduces drift dynamics a. Network architecture and sample raster plot of CA1 pyramidal cells. **b.** Heterogeneous response profiles for sample CA1 cells. The color indicates the value of spatial information from c. **c.** Histogram of spatial information for all CA1 cells over one session. **d.** Synaptic turnover from session to session is modeled by randomly rewiring a fraction of inputs to each CA1 pyramidal cell, independently for EC versus CA3 pathways. **e.** Tuning curve (top row) and total input (bottom row) for one example cell over three sessions. The dashed line in the bottom panel indicates the average total input along the track for each session. **f.** Place field maps for 200 randomly selected active cells found on session 1 (top) or session 8 (bottom), ordered according to their place field positions. **g.** PV correlation of all cells (left), and only place cells (cells significantly spatially tuned in both sessions) (right). The insets show the first four points of the respective curves, where the initial point (within-session correlation) is computed considering odd vs even trials. **h.** Distribution of the centroid shift for different number of elapsed sessions (color-coded). Inset: cumulative distribution of the absolute shift. See Methods for model details and parameter values.

The synaptic turnover responsible for RD is consistent with ongoing memory storage

In the previous section we showed that synaptic turnover, modelled as a process of random rewiring of the synaptic inputs to CA1 pyramidal cells, could account for the statistics of RD observed in mice. At such a level of description, the mechanism behind the turnover itself remains unclear. We hypothesize that this rewiring process may, in fact, reflect the encoding of episodes or memories. In this scenario, much of the synaptic turnover, and hence RD observed from session to session, would be due to the storage of memories unrelated to the environment in which the recordings are made, Fig.3a-b. Specifically, instead of rewiring synapses randomly, we modified them according to a simple Hebbian plasticity rule. We imposed random binary patterns of activity for the CA3 and EC inputs, and the CA1 outputs, with sparseness f. We then potentiated synapses between co-active cells with probability p_+ and depressed synapses between cells with differing activities with probability p_-. These probabilities were different for the EC and CA3 inputs. We repeated this process many times, until the synaptic weight matrices from

EC and CA3 to CA1 reached a statistical steady state, see Methods and Supplementary Information for more details. We then tracked one of the random patterns in particular, as if it were the experimentally observed activity pattern. In between observations of this pattern, we encoded other random patterns, the number of which we call the inter-session interval (ISI), Fig.3b. With this simple model we could calculate the temporal correlation of the inputs to CA1 cells analytically as a function of the plasticity parameters, and match them to the synaptic turnover process from the network, Fig.3c.

In our model, both synaptic turnover and RD are reflections of the changes in network structure due to the storage of memories, most of which are episodes unrelated to the ones observed during an experimental session. Nonetheless, given a Hebbian rule, plasticity occurring during a session, namely a repetition of the tracked pattern, can have an outsized effect on RD. This is because all of the neurons being tracked are co-active by definition, whereas other sparse, random patterns will have a small overlap with the tracked pattern. The repetition rate of a pattern therefore affects the degree of RD observed, although whether drift is increased or decreased depends on the details of the input statistics. In the following sections we show how this phenomenon can account for recent findings from both hippocampus[41] and piriform cortex[25].

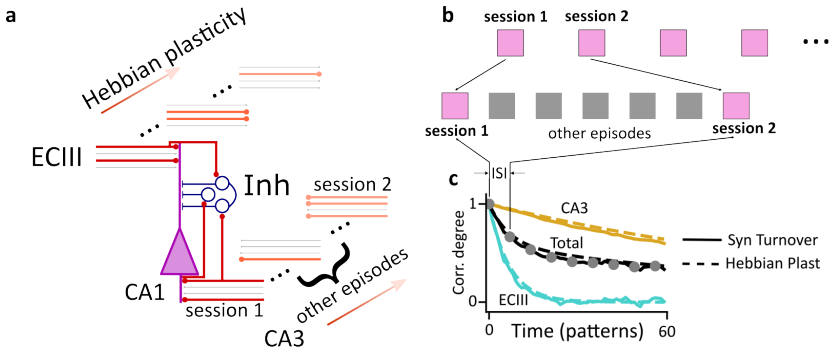


Fig. 3 |Synaptic turnover is consistent with the ongoing storage of random patterns. a. The synaptic turnover used to fit RD from the data is now generated through the encoding of random patterns with a Hebbian plasticity rule. **b.** We assume that in between sessions in which the activity patterns are tracked, there are a number of random patterns encoded. **c.** The learning process is fit to the synaptic turnover by matching the drop in correlation in the number of inputs to CA1 cells over sessions, ρ_d. The grey circles indicate the values of correlation which correspond to the snapshots of activity shown in Fig.2f, see Methods for details and parameter values.

Ongoing memory storage accounts for the differential roles of time and experience on RD in place cells in CA1

Recent experimental work has shown that the spatially untuned component of RD in CA1 pyramidal cells occurs at a rate proportional to the absolute time elapsed between sessions on a linear track, while RD in the tuned component depends on the time spent exploring the track itself[41]. The mechanism we propose here for RD suggests a potential explanation for this dissociation. Namely, we would expect that the ongoing storage of memories unrelated to the linear track itself would simply occur at a rate proportional to time. Such memories would naturally overwrite some synapses involved in the representation of the track, leading to RD. In principle, these synapses could be from spatially tuned or untuned presynatic inputs. However, it is reasonable to assume that it is the spatially untuned, contextual inputs which are more strongly shared between temporally proximate memories[16]. On the other hand, it is only when the animal is actively exploring the track itself that the entirety of the spatially tuned inputs are active. Given an activity-dependent plasticity rule, we should thus expect maximal RD due to changes in the spatially tuned inputs to occur precisely during exploration.

In the framework of our network model, in which a series of patterns with spatial and non-spatial component are stored, the dissociation between time and experience is, in general, approximate. It is exact only in the limit in which the overlap between spatially tuned presynaptic inputs vanishes between patterns, Fig.4a. In this case, if we consider two sets of simulations in which one particular pattern (e.g. corresponding to the linear track) is repeated more often in one set than in the other, Fig.4b, then the RD in firing rates will depend only on the number of patterns presented, i.e. time, while the RD in tuning will depend only on the number of repetitions of the pattern of interest, i.e. experience. We ran simulations in which the encoded patterns consisted of a random fraction of spatial and non-spatial inputs, and hence the overlap in spatially tuned presynaptic inputs between pattern was not zero, but rather equal to the fraction of active spatially tuned cells $f_s = 0.1$. In this case, the RD in firing rates robustly depended on absolute time, but the RD in tuning depended both on time and experience, Fig.4c. However, this result is true only when the patterns of presynaptic activity are stable over time. In fact, place cell activity in CA3 itself undergoes RD[43], suggesting that those presynaptic inputs should themselves vary from one repetition to the next, Fig.4d. Including this effect tips the balance of spatial RD from time to experience through a combination of presynaptic RD and plasticity, Fig.4e.

Ongoing memory storage accounts for the effect of repetition rate on RD

Recent experiments have revealed that the degree of RD in the piriform cortex of mice is reduced by increased repetition of familiar odors[25]. This was shown by familiarizing two groups of animals with a particular odor over many sessions, and then repeating that same odor each subsequent session for cohort A, while cohort B was exposed to the familiar odor only after a number of sessions without odor, Fig.5a.

As was the case in the previous section, we again assume that RD for the familiar odor is largely due to the storage of other patterns in piriform cortex, which occur in between sessions, see "random pattern" in Fig.5a. This assumption is enough to provide a simple explanation for the experimental finding, Fig.5b. Namely, if we consider the synaptic weight matrix for cohorts A and B, they are both initially identical due to the familiarization process. The storage of other patterns, which are random and uncorrelated with the familiar one, degrade the structure in the matrix which is correlated with the familiar pattern. However, in cohort A, the familiar pattern is repeated at the subsequent session, once again boosting the structure in the matrix, while in cohort B this is not the case.

We simulated this process using the same Hebbian plasticity rule as in previous sections, but with only a single input layer. We presented a pair of sparse, binary input and output vectors repeatedly until the synaptic weight matrix reached a steady state. We then repeatedly encoded the pattern to be tracked and considered this to be the familiarized state, see Methods for details. To simulate cohort A, we then encoded a number of random patterns in the network equal to the inter-session interval (ISI) after which time we again presented the familiar pattern, Fig.5c (black line). For cohort B we presented repetitions of the familiar patterns only after every eight repetitions in cohort A, leading to significantly reduced output correlations, Fig.5c (green line). The size of the difference in output correlations (RD) is strongly affected by the ISI, Figs.5d,e.

In the same set of experiments, an unfamiliar odor was presented to both cohorts with the same repetition rate and no difference in drift rate was found. We can account for this by running the same simulation as before, but now tracking both the familiar pattern, and one pattern uncorrelated with the familiar one, Fig.5f. This amounts to considering the structure of the weight matrix, and the output vector, with two distinct orderings. As long as the familiar and unfamiliar patterns are uncorrelated, the drift rate of the unfamiliar pattern depends only on the repetition rate and ISI, which are the same for both cohorts, Fig.5g-h. Finally, the familiar pattern with high repetition rate has a significantly reduced drift rate compared to the other three cases, in agreement with experiment, Fig.5i.

Ongoing memory storage can explain diverse effects of experience

It may at first seem paradoxical that increased exposure to a spatial environment would lead to more RD in CA1 of hippocampus[41], while increased exposure to a familiar odor would actually reduce RD in

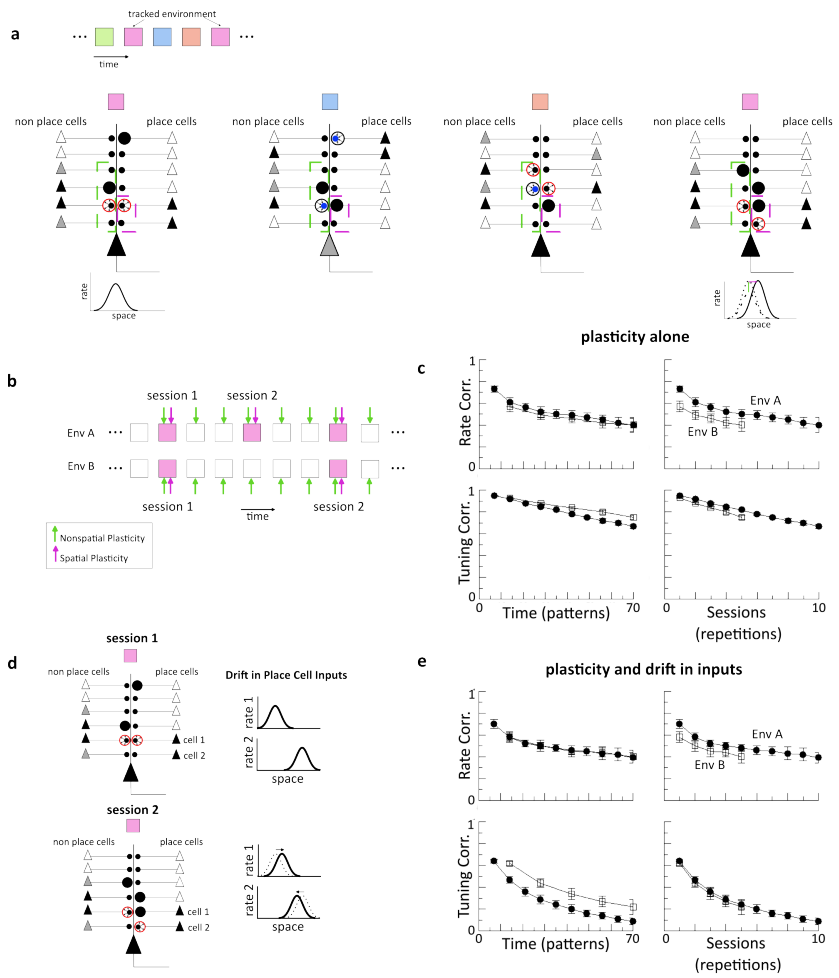


Fig. 4 |Ongoing memory storage is consistent with differential effects of drift on rate and tuning **a.** We consider the storage of many memories, indicated here by color, all with spatial and non-spatial features, and one of which (pink) is tracked explicitly. Only changes to synapses which are active in that environment (dashed boxes in cell schematic) will generate observable RD. In the limit of sparse place cell coding the storage of the blue and orange memories readily causes drift in non-spatial inputs, but not in the spatial ones. As a result, RD in non-spatial features is proportional to the total number of memories stored, while RD in spatial tuning is due only to repetitions of the tracked memory. **b.** Simulation protocol for which the repetition rate for environment A is twice that of B. In the sparse spatial coding limit the rate correlation depends on total patterns stored (time) while the tuning correlation depends only on repetitions of the tracked pattern (session). **c.** Results of simulations when the sparseness of spatial coding $f_s = 0.1$. The drop in tuning correlation is affected both the number of repetitions as well as interference from other memories. **d.-e.** When spatially tuned inputs already exhibit drift, as has been observed in CA3 place cells, the combined effect of sparseness and drift results in the tuning correlation being dominated by the number of repetitions. See Methods for model details and parameter values.

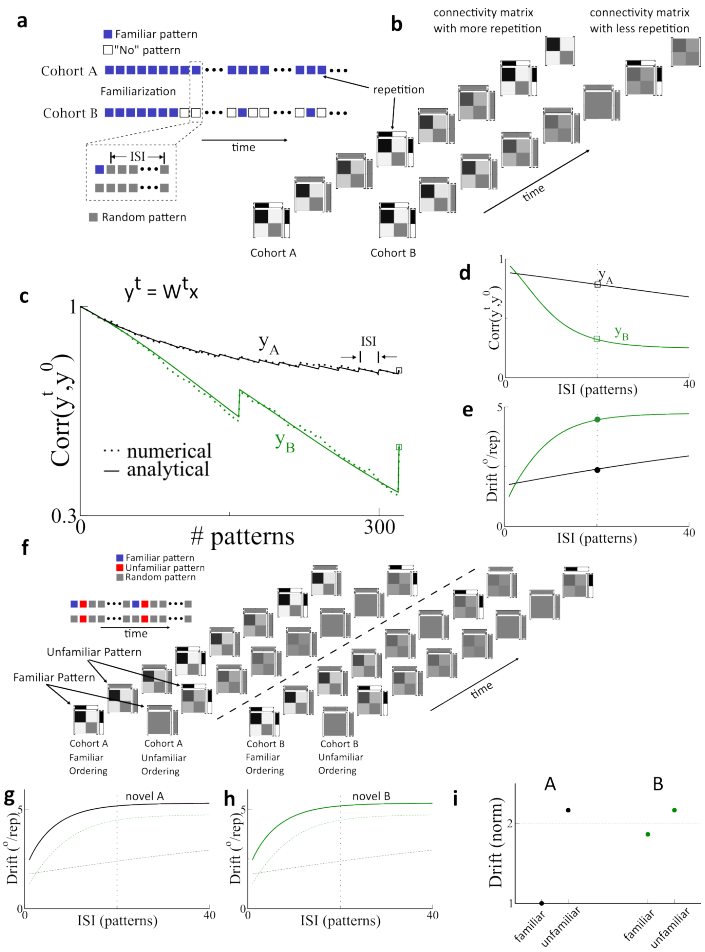


Fig. 5 |Repeated exposure to same input pattern reduces drift. a. Protocol for testing the effect of repetition rate on drift. Two cohorts are repeatedly presented a pattern (blue squares) during a familiarization period. After this time cohort A is presented the familiar pattern every session while B is presented the familiar pattern every eight sessions. Crucially, we assume that a number of additional patterns, uncorrelated with the familiar one are encoded between sessions (inter-session interval, ISI) (grey squares). **b.** Illustration of the effect of repetition rate on network connectivity. Repetition boosts network structure correlated with the familiar pattern, thereby reducing drift. **c.** Output correlation for familiarized pattern with a total of 16 repetitions for cohort A and $ISI = 20$, while the repetition rate for B was 8 times less. Dotted lines are from simulation of the network model with Hebbian plasticity while solid lines are the solution of the corresponding Markov process. **d.-e.** Output correlation and drift rate as a function of the ISI, i.e. the number of random patterns encoded between "sessions". The vertical line indicates the value of ISI used in **c. f.** Illustration of encoding of familiar and unfamiliar patterns. Because the repetition rate is the same for the unfamiliar pattern for both cohorts, the resultant drift is also the same. **g.-h.** Drift rates for the unfamiliar patterns. **i.** Drift rates normalized by the familiar case for cohort A. Parameters: $p_+ = p_- = 0.02$, $f = 0.15$, $N = 1000$. For the familiar cases the tracked pattern was encoded $k = 5$ times at time zero, whereas for the novel cases $k = 1$. See Methods for model details.

piriform cortex[25]. In the context of ongoing memory storage through Hebbian plasticity, both scenarios are possible, Figs.4 and 5, suggesting that this single mechanism provides a potential explanation.

We first note that the repeated storage of the *same* pattern tends to increase the correlation of the synaptic weight matrix with that pattern, Fig.6a-c. A period of familiarization provides a substrate for enhancing the effect of this repetition rate. The degree of this enhancement of correlation, and hence reduction in RD, depends strongly on the fraction of synapses which are updated to store each memory

(learning rate). We quantified this by varying the number of times k a pattern was repeated before time $t = 0$. For $k = 1$ the pattern was novel. When the learning rate is small, increasing k leads to a build-up of correlation which is large compared to a single presentation, Fig.6a, whereas for large learning rates this difference is smaller, Fig.6b. Large k implies a weight matrix which is entirely correlated with the tracked pattern, and hence has lost all correlation with previously stored patterns, an unreasonable limit for an actual physiological memory system. However, the difference in RD for different repetition rates is already present for small values of k and even reaches a maximum for intermediate values when learning rates are small, black lines Fig.6c. These are the parameter values used to reproduce the findings from piriform cortex in Fig.5. The effect of varying the coding sparseness f is shown in Extended Fig.6.

When the input pattern is *not* the same from one repetition to the next then the effect of repetition rate is not as straightforward. We model this by keeping a fraction s of the active units in the tracked pattern the same from repetition to repetition, while choosing the other $1 - s$ randomly, Fig.6d. When $s = 1$ repetitions reduce RD as explained above. For $s = 0$ the pattern is effectively random and uncorrelated from one repetition to the next. For intermediate s the correlation of the rth repetition with the original input pattern decreases as s^r which can lead to RD increasing for sufficiently high repetition rates, see Fig.6e. All other network parameters being the same, if RD is measured after a fixed number of repetitions, then there is a critical value of s below which repetition rate will increase RD, Fig.6f.

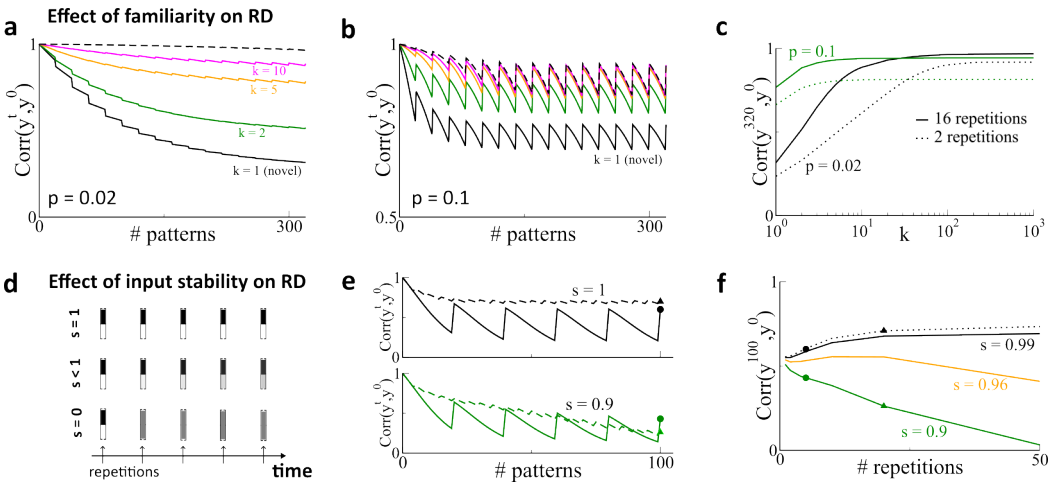


Fig. 6 |How familiarity and input stability modulate RD. a. We model familiarity by presenting a given pattern to the network k times before time $t = 0$, i.e. $k = 1$ is a novel pattern. Here $p_+ = p_- = 0.02$, $f = 0.15$ and $N = 1000$. **b.** Same as in **a.** but with $p_+ = p_- = 0.1$. **c.** Correlation after a total of 320 patterns, with either 16 or 2 repetitions, for low and high learning rates. **d.** The input patterns driving the observed activity may themselves undergo RD, which we parameterize with s. Here $s = 1$ indicates completely stable inputs over time while $s = 0$ indicates that the input vectors are completely uncorrelated from repetition to repetition. **e.** When inputs are stable, increased repetition rate decreases RD ($s = 1$), while this need not be the case when inputs themselves drift ($s = 0.9$). **f.** The repetition rate can reduce RD (black lines), increase RD (green line) or even leave it largely unchanged (orange line).

Discussion

Here we have argued that RD is, in large part, due to the storage of memories unrelated to the experiment in which RD is observed. Such a storage process implies changes in the synaptic weights of the network, which will necessarily alter its response to stimuli on subsequent sessions, even if upstream neuronal activity is unchanged. We have shown that the statistics of RD in CA1 of mice are well captured by a process of synaptic turnover in which inputs undergo rewiring from one session to the next. The fact that this synaptic turnover is consistent with the ongoing encoding of activity patterns to us strongly

points to this as the underlying mechanism. After all, episodic memory formation, a major function of the hippocampus, involves precisely this kind of ongoing storage of information related to experiences during daily life. Ongoing memory storage can furthermore account for the differential effects of time versus experience in RD in CA1, as well as the effect of repetition rate, both stabilizing as found in piriform cortex and destabilizing, as in CA1, depending on the input statistics. These nontrivial and seemingly paradoxical findings cannot be explained by a process of random synaptic turnover, but are reproducible in a network model in which patterns are encoded via a Hebbian plasticity rule.

Other computational modeling work has also proposed that RD may arise due to plasticity-driven perturbations[29–31]. In[29] the authors studied a network with prescribed synaptic weights which allow for the storage of a large number of sequential patterns. They showed that randomly perturbing the synaptic weight matrix generates RD while maintaining robust sequences. RD can also arise in a spiking network with a symmetric spike-timing dependent plasticity rule, coupled with a homeostatic mechanism[30]. Specifically, if the initial network structure (synaptic weight matrix) exhibits clustering, ongoing plasticity allows for individual cells to leave their cluster and join a new one, all the while maintaining the clustered structure at the network level. RD occurs in a similar fashion in networks which minimize the mismatch between the similarity of pairs of input patterns and the corresponding pairs of output patterns (Hebbian/antiHebbian networks)[31]. Namely, ongoing plasticity allows the network to explore the degeneracy in the solution space by undertaking a random walk along the manifold of equally optimal output patterns. An important conceptual difference between our work and previous studies is the nature of the synaptic turnover itself. While those studies showed that *already-stored* patterns of activity undergo RD in the face of ongoing plasticity, we ascribe the plasticity to the encoding of *new patterns*. Namely, we have shown that RD is consistent with the inevitable interference between patterns when learning occurs, and hence not necessarily just a consequence of noise once learning is done. The mice in the experiments we have studied are not exposed to explicit, task-dependent learning between sessions. Rather, the "learning" process may simply be the storage of episodes, unrelated to the exploration. The hippocampal circuit plays a central role in this type of memory[44,45], while there is evidence that piriform cortex participates in olfactory associative memory formation[46].

The stability paradox We have sought here to provide a plausible network mechanism for RD, and have not addressed the fundamental paradox of how to maintain stable behavior in the face of such neuronal instability. In fact, drift does not appear to adversely effect the performance of mice in a variety of memory-dependent tasks[24,25,47]. Several previous studies have addressed how this might be possible. Firstly, it has been hypothesized that there may be a low-dimensional manifold which represents the task-relevant projection of the population response, and which is invariant to RD[27,48,49]. In this scenario many distinct patterns of the high-dimensional population activity can have the same projection on the relevant, low-dimensional manifold. If RD does not affect this projection, i.e. it is constrained to the "null-space" of the manifold, then task-relevant variables can be stably read out. While neuronal representations undergoing RD do appear to be low-dimensional, the direction of drift has not been found to be orthogonal to this manifold in general[7,25,50]. Alternatively, despite ongoing changes, the patterns of neuronal activity observed in-vivo retain some significant correlation from session to session[7,16,21,24,25]. Therefore, decoders trained on a given session will perform above chance for subsequent sessions, albeit with degraded accuracy. In any case it remains unclear precisely how behavior remains unaffected by the observed drift. One potential solution to this is to allow for compensatory plasticity in downstream circuits so as to stabilize the readout performance[30,48,50–52]. Another is to seek some higher-order population structure which remains stable in the face of ongoing RD[19,21,22,27,48,49].

Yet despite recent theoretical advances and experimental findings, the stability paradox remains. It may be that behavior depends on a neuronal representation which is distributed across several cortical areas, and that RD is greatly reduced in some of the them compared to those observed up until now. In the case of the hippocampus, it is hypothesized that memories are transferred to higher-order cortical areas for long-term storage via a consolidation process which can take weeks, months or years depending on the species[53–56]. Computational models of distributed memory systems in which fast learning (and hence fast forgetting) circuits drive plasticity in slower-learning downstream circuits in a multi-layered

framework, show qualitatively enhanced memory capacity compared to single-area models[34]. Such models leverage a hierarchy of time-scales of synaptic plasticity to allow for both fast encoding as well as long lifetimes[6,34,57]. A hallmark of a model in which the timescale of plasticity is spatially distributed across cortical areas, is a concomitant array of timescales for RD. Only future experiments in which the activity of neuronal populations across several cortical areas are recorded simultaneously over long periods of time will reveal if RD acts on distinct timescales in different brain areas.

Methods

Data analysis

We analyzed previously published data[16]. In the experiment, mice repeatedly explored two familiar environments (one in the morning, the other in the afternoon) over the course of two weeks, with imaging sessions every other day (8 total sessions). We separately analyzed leftwards vs rightwards running epochs for each environment, and then pooled the data together. In total, we analyzed data from 5 mice. Of the five mice, two have only one spatial map for each environment, while three have multiple spatial maps for each environment, as previously analyzed[20]. Unless otherwise specified, all the analysis was performed separately for each map, and then results were pooled together.

Calcium events and place field maps were extracted as in the original paper[16]. Briefly, the linear tracks were divided into 24 bins, each 4cm in length. For each spatial bin, the total number of events in each session was extracted, together with the total time-occupancy of each bin. The event rate map is then calculated by dividing the total number of events per bin by the bin occupancy. The two bins at each extreme of the track were excluded from the analysis to limit reward delivery effects.

In Extended Data Fig.1, the rate maps follow two different types of normalization: in Extended Data Fig.1a, the event rate map for each cell was normalized by the maximal firing rate over all sessions. In Extended Data Fig.1c, the event rate in each session/bin as normalized according to the maximal value within each session.

Rate and tuning correlation

To compute rate correlations across sessions, we defined a rate population vector $\mathbf{r_t} \in \mathcal{R}^N$. N was the total number of neurons, and each entry of the vector was the mean firing rate of each cell in session t. We then computed the Pearson correlation coefficient to quantify the rate similarity between two sessions, as in Entended Data Fig.1d-f.

To quantify the similarity of spatial tuning across session, we defined a Tuning population vector of length $N \times n$, where n is the number of bins of the linear track. The vector then contained the rate of each cell in each spatial bin, normalized in such a way that the sum of the rates for each cell over the different spatial bins was constant in all sessions. We then computed the Pearson correlation coefficient of such vectors for pairs of session to quantify the tuning similarity. The procedure ensured that changes in mean firing rates of cells from one session to the next would not affect the tuning correlation.

Place fields

To be considered a place cell in a given session, we employed the following procedure. First, for each cell, we computed the spatial information per spike as :

$$s_i = \sum_{k=1}^{n} p_k \frac{r_i^k}{\bar{r}_i} \log\left(\frac{r_i^k}{\bar{r}_i}\right),$$

where i was the cell index, and the sum runs over all n spatial bins, r_i^k was the event rate of cell i in bin k, $\bar{r}_i$ was the mean event rate over all bins, and p_k was the occupancy of bin k (the fraction of time spent in bin k). s_i was then the spatial information (measured in bits/event) of cell i. Then, we generated

surrogate data by shuffling the position of the animal with respect to the time of the calcium events, and calculated the spatial information for each cell and each shuffle. We then compared the value of the spatial information of each cell to the null distribution generated with the surrogate data. If the value was larger than the 95th percentile of the null distribution, than the cell was defined as a place cell in that session.

The place field width of each cell was defined as the number of contiguous bins where the event rate is larger than 50% of its maximal value over all the bins.

The statistical model

In order to fit the network model to data in Fig.2, we first fit a simple statistical model to the data, and then mapped the network parameters onto the resulting fit parameters as a starting working point. For the statistical model we took cells in CA1 as binary units which received inputs from two sources, CA3 and layer III of entorhinal cortex EC. The total input to a cell i at time t (time measured in sessions) from CA3 was written x_i^t and from EC y_i^t. Both inputs were Gaussian random variables with zero mean and variances σ^2_{CA3} and σ^2_{EC} respectively. A neuron i was active at time t if its total input $h_i^t = x_i^t + y_i^t$ exceeded a threshold θ, and was otherwise silent. Specifically, the activity was written as $a_i^t = H(h_i^t - \theta)$, where $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x \leq 0$ is the Heaviside function.

To model RD we allowed for the inputs to change over time according to:

$$x_i^t = \rho_{\text{CA3}} x_i^{t-1} + \sqrt{1-\rho^2_{\text{CA3}}}\psi_i^t, \quad y_i^t = \rho_{\text{EC}} y_i^{t-1} + \sqrt{1-\rho^2_{\text{EC}}}\xi_i^t, \tag{1}$$

where the autocorrelations ρ_{CA3} and $\rho_{\text{EC}} \in [0,1]$ and ψ_i^t and ξ_i^t were Gaussian random variables with mean zero and standard deviation σ_{CA3} and σ_{EC} respectively. The update rules Eqs. (1) ensured that the input distributions were stationary. Inhibition is implicitly assumed to have the effect of subtracting the mean of both inputs so that they are centered at zero.

The statistical model had four parameters once we rescaled them by the standard deviation of the inputs from EC: the ratio $\sigma_{\text{CA3}}/\sigma_{\text{EC}}$, the rescaled threshold $\theta/\sigma_{\text{EC}}$ and the autocorrelations ρ_{CA3} and ρ_{EC}.

Note that the input dynamics Eqs. (1) could be formulated in continuous time as an Ohrnstein-Uhlenbeck processes, as shown in the Supplementary Material . The continuous formulation allowed us to calculate the time constant of the decay in correlation of the inputs analytically, yielding $\tau = \frac{1-\rho^2}{2(1-\rho)^2}$.

According to the definitions above, the state of the network was defined by a vector $\mathbf{a^t} = (a_1^t, \ldots, a_N^t)$ where N was the total number of neurons. *Fraction of active cells:* The fraction of active cells can be calculated analytically for the statistical model. The probability that a cell with a given input y is active can be written $\Pr(y > \theta - x)$. Integrating this probability over all possible values of y gives the likelihood of any cell to be active, or the fraction of active cells, f_a. This fraction is therefore

$$f_a = \int_{-\infty}^{\infty} dx \rho_{CA3}(x) \int_{\theta - x}^{\infty} dy \rho_{EC}(y), \tag{2}$$

where $\rho_\alpha(x) = \frac{e^{-x^2/(2\sigma_\alpha^2)}}{\sqrt{2\pi}\sigma_\alpha}$

Fit of the statistical model

To fit the statistical model, we only considered mice with a single map per environment (two mice in total), since with multiple maps the distribution of the number of sessions in which each cell was active and the survival fraction are not well defined (with multiple maps not all maps are visited in all sessions). We considered data from these two mice (N=1649 total recorded cells). The statistics used to fit the model were the distribution of sessions each neuron is active, the survival fraction (probability an initially active cell continued to be active on subsequent sessions), the population activity overlap between different sessions, and the fraction of active cells in each session, see Extenden Data Fig.2. To fit the data, we adjust the four free parameters of the model using least-square optimization. Specifically, to produce

Extended Fig.2g-i, we discretize the (ρ_{CA3}, ρ_{EC}) space on a 20×20 grid, and for each position on the grid, run the Scipy[58] implementation of a Basin-hopping optimization algorithm minimizing the sum of the squared residuals between numerical simulations of the statistical model, and the experimental data. For the numerical simulations of the model, we consider $N = 20000$ neurons.

Network model description

CA1 was modeled as a network of integrate-and-fire excitatory (E) and inhibitory (I) neurons with I-I, E-I and I-E connections, but no recurrent excitation, as prescribed by anatomical constraints. CA3 cells are modeled as Poisson neurons a fraction of which were spatially modulated. They projected onto both excitatory and inhibitory CA1 neurons. Additionally, CA1 neurons received excitatory inputs from a layer of non-spatial Poisson neurons (from layer III of EC). Detailed equations and parameters are given in Supplementary Information.

CA3 place fields and non-spatial inputs

CA3 neurons were modeled as Poisson neurons. In any given environment, a fraction f_{CA3} of CA3 neurons were active. Of the active neurons, a fraction f_s of the population had a spatially modulated firing rate, while the remaining fraction had a constant firing rate. For simplicity, we considered a ring topology, so that the spatial position of a virtual animal was parametrized by an angle $\phi \in [-\pi, \pi]$. The firing rate of spatially selective neurons was modulated according to a Von Mises distribution:

$$r_i(\phi) = R(\beta) e^{\beta \cos(\phi - \phi_i)} + r_0, \tag{3}$$

where ϕ_i was the center of the place field of neuron i, r_0 a baseline firing rate, and $R(\beta) = \bar{R} e^{-\beta}$ where $\bar{R}$ was a constant that set the maximal firing rate. The parameter β determined the sharpness of the place fields.

The layer of neurons providing non-spatial inputs were modeled as Poisson neurons with constant firing rate ν. Also for this layer, only a fraction f_{EC} of the population was active in any given environment.

Connectivity matrices

Connectivity matrices for all subtypes of connections within the CA1 populations were random and sparse. Each neuron had a probability of connection to other neurons in the respective subpopulations equal to $\alpha_{i,l} = K_{i,l}/N_l$, where N_l was the number of neurons of the lth population, $l \in \{E, I\}$, and $i \in \{E, I\}$. If not specified otherwise, we fixed the in-degree of CA1 pyramidal cells from interneurons to be fixed and equal to K_I (all pyramidal cell receive the same amount of inhibitory inputs). On average, each CA1 cell received projections from $K_{i,\{E,I\}}$ neurons. The connectivity between the layer of non-spatial Poisson neurons and CA1 is also random and sparse with connection probability α_{EC}. In the main text, we consider a uniform connection probability also for the CA3 $\rightarrow$ CA1 projections; in the Supplementary Information we discuss how to consider a phase bias in such connections.

Throughout, we assume all connection probabilities are the same and equal to $\alpha = 0.125$.

Input correlations in the network model

To implement the session-to-session changes in spatial and non-spatial inputs provided by Eqs. (1) in the network model, we needed to calculate the autocorrelation of inputs where changes may occur either due to changes in the input firing pattern, or in the connectivity matrices themselves. The general expression for such inputs is

$$I^t = \sum_j c_{ij}^t \nu_j^t, \tag{4}$$

where ν_j^t is the firing rate of the input neuron j at time t, c_{ij}^t is the connectivity matrix element at time t, and we omitted conductances and time constants. In the Supplementary Information we provide detailed calculations. The expression one finds for the autocorrelation is

$$\rho = \frac{\langle \nu_j^t \nu_j^{t+1} \rangle_j \left[\langle c_{ij}^t c_{ij}^{t+1} \rangle - \alpha^2 \right]}{\langle \nu_j^2 \rangle_j \alpha \left(1 - \alpha\right)}, \tag{5}$$

where α is the connection probability to the input layer. In general then, the level of correlation depends on the correlation in the input firing patterns, and on the degree of synaptic plasticity. Note that in order to obtain completely uncorrelated inputs from one session to the next, one must have $\langle c_{ij}^t c_{ij}^{t+1} \rangle = \alpha^2$, which implies $\langle c_{ij}^t c_{ij}^{t+1} \rangle = \langle c_{ij}^t \rangle \langle c_{ij}^{t+1} \rangle$, i.e. complete rewiring from one session to the other. Assuming completely uncorrelated input firing patterns from one session to the next results in:

$$\rho = \frac{\mu_\nu^2 \left[\langle c_{ij}^t c_{ij}^{t+1} \rangle - \alpha^2 \right]}{\left(\mu_\nu^2 + \sigma_\nu^2\right) \alpha \left(1 - \alpha\right)}. \tag{6}$$

On the other hand, if we assume that the input firing patterns are the same over sessions, we have:

$$\rho = \frac{\langle c_{ij}^t c_{ij}^{t+1} \rangle - \alpha^2}{\alpha \left(1 - \alpha\right)}, \tag{7}$$

which, as shown in Appendix , can be written as:

$$\rho = \frac{\Pr\left(c_{ij}^{t+1} = 1 \middle| c_{ij}^t = 1 \right) - \alpha}{1 - \alpha}. \tag{8}$$

In the following, for both spatial and non-spatial inputs, we assumed that the input firing rates were constant from one session to the next. In this case, the fraction of connections rewired from one session to the next depending on the autocorrelation ρ is shown in Fig. Extended Data Fig. 3h.

Variance of spatial and non spatial inputs

The average current from either the spatial or non-spatial inputs layer had the form:

$$I = \frac{G\tau}{\sqrt{K}} \sum_{j=1}^{N} c_{ij} \nu_j, \tag{9}$$

where we have scaled the synaptic weights as $1/\sqrt{K}$. The expected value of the input is therefore

$$\mu_I = \frac{G\tau}{\sqrt{K}} \alpha N f \nu, \tag{10}$$

where α is the connection probability, f is the fraction of active pre-synaptic cells, and ν is their mean rate. If we define the normalization factor as the mean number of active inputs $K = \alpha f N$, we have $\mu_I = G\tau\sqrt{K}\nu$. The variance of such current, using the results from the Supplemetary Information, was given by

$$\sigma^2 = \frac{G^2\tau^2}{K} \alpha \left(1 - \alpha\right) N f \nu^2, \tag{11}$$

where we neglected the intrinsic variability of the Poisson process (which goes to zero as Δt^{-1}). Again using the definition of the normalization factor K, we have

$$\sigma^2 = G^2\tau^2\nu^2 \left(1 - \alpha\right). \tag{12}$$

Given the variances ratio obtained fitting the statistical model $\tilde{\sigma} = \sigma_{\text{CA3}}/\sigma_{\text{EC}}$, we can then fix the firing rates/synaptic weights of CA3 and EC inputs via the relation:

$$\tilde{\sigma} = \frac{G_{CA3}\nu_{CA3}}{G_{EC}\nu_{EC}}, \tag{13}$$

since both the time constant and the connection probabilities are the same for the two layers. We then fix $\nu_{CA3} = \nu_{EC}$ (same average rate of the two layers), and set $G_{CA3} = \tilde{\sigma} G_{EC}$.

Fit of network simulations to data

In network simulations the virtual animal runs at a constant speed of $v = 12$ cm/s over a circular track of length $L = 84$ cm. The position of the animal on the track is parametrized with a phase $\phi \in [-\pi, \pi]$. Analogously to the experiment, we simulate 8 sessions each consisting of 20 laps on the track (i.e. one session lasts 140 seconds of simulation time). For place fields analysis, the track was divided into 20 bins each of 4.2 cm length.

Given the analytical formulas derived in the previous section we were able to match both the variances of the input distributions as well as their autocorrelations to the values from the fit of the statistical model. The means of the input distributions were not zero, as in the statistical model, but the network operated in a balanced regime in which currents from inhibitory interneurons cancelled the mean excitatory drive to cells in the mean, leaving their membrane potential near threshold to spiking. Therefore, fitting the variances and correlations alone set the network at a working point in which the statistics of RD were close to the statistical model, and hence the data. We then made slight changes to parameters by hand in order to improve the fit. In order to fit the population vector correlation we needed to compare to the calcium event rate in the data. We did this by applying an exponential kernel with time constant $\tau_c = 500$ms to the spike train from each cell in the spiking network. A calcium event was detected whenever the smoothed signal crossed a threshold θ_c, and imposed a minimum inter-event interval of 500ms. The threshold value which minimized the mean squared error of the fit was $\theta_c = 0.16837$.

Plasticity model

In the network simulations used to fit the experimental data in Fig.2 changes in inputs from one session to the next were modeled as a process of synaptic turnover, without specifying the precise mechanism responsible for this. One possibility is that this turnover is due to the storage of patterns in the network. Specifically, we assume that between sessions a fixed number of patterns are encoded, which we call the inter-session interval (ISI). Each pattern is a binary vector in which only a fraction of cells f_s are active. Of the active cells, we assume that a fraction f are strongly active, while the remaining $1 - f$ are weakly active. If we are modeling CA1, as in Figs.3 and 4, there would be three such vectors, while for Figs.5-6 there is only one input layer and hence only two vectors. To store a pattern we apply the following plasticity rule. If the pre- and post-synaptic cells are both strongly active, then the synapse is potentiated with probability p_+. If one of the cells is strongly active while the other is weakly active then we depress the synapse with probability p_-. If both are weakly active no change is made at the synapse. For the CA1 model, these probabilities can be different for CA3 versus EC inputs. Synapses are binary, as before, i.e. $c_{ij} \in \{0, 1\}$. To quantify RD we calculate the correlation in the output pattern at time t with that at time $t = 0$ given the same (tracked) input patterns. If the input patterns are the same, then RD can only be due to changes in the network connectivity. If the input patterns change, as is the case with CA3 inputs in Fig.4e, then RD is affected both by this as well as changes in network connectivity.

Figure 3: We first mapped this process onto the process of synaptic turnover in the network model by matching the decay in the correlation of the number of inputs a cell receives over time, the degree-correlation ρ_d, see Fig.3c and Supplementary Information for a detailed derivation. The parameter values used to generate Fig.3c were: $f_{EC} = f_{CA1} = f_{CA3} = 0.5$, $f_{s,EC} = f_{s,CA1} = f_{s,CA3} = 1$, $p_+^{EC} = p_-^{EC} = 0.18$, $p_+^{CA3} = p_-^{CA3} = 0.01$, $N_{EC} = N_{CA1} = N_{CA3} = 1000$. This comparison shown

in Fig.3 is purely structural. We also performed network simulations with the plasticity process and observed RD as in Fig.2, see Extended Fig.7. For those simulations we modeled a population of CA1 cells as linear threshold units, i.e. the activity of cell i was given by $r_i = [I_i - \theta]_+$, where I_i was the total input, θ was the threshold, and $[x]_+ = x$ if $x > 0$ and is zero otherwise. Active cells in EC had a constant firing rate r_{EC} while active cells in CA3 were spatially modulated according to a von Mises distribution as in Eq.3. The threshold was $\theta = 0.3$ and for the von Mises rates the parameters were $r_0 = 5$, $\bar{R} = 90$, $\beta = 19$. r_{EC} is chosen such that the ratio $\sigma_{CA3}/\sigma_{EC} = 1.16$ in order to match the fit from the statistical and network models. The initial condition for the connectivity matrix was random with connections probability $p = 0.33$. We tracked the 100th pattern stored and plotted snapshots of the activity in CA1 starting at time $t = 140$ and every 7 time steps for eight "sessions". Neurons were considered inactive if their mean firing rate was less than 0.1.

Figure 4: We used the same network model of linear-threshold units as described above with the following parameter values: $f_{EC} = f_{CA1} = f_{CA3} = 0.5$, $f_{s,EC} = f_{s,CA1} = 1$, $f_{s,CA3} = 0.1$, $p_+^{EC} = p_-^{EC} = 0.18$, $p_+^{CA3} = p_-^{CA3} = 0.02$, $N_{EC} = N_{CA1} = N_{CA3} = 1000$, $\theta = 0$. All other parameter values were the same. Drift in the CA3 inputs was modeled by allowing the position of the ith place cell (center of the von-Mises distribution) at time t to shift with respect to the position at time $t-1$ according to $\phi_i^t = \phi_i^{t-1} + \sigma\xi^t$, where ξ is uniformly distributed from $\{-\pi, \pi\}$ and $\sigma = 0.0447$. Plasticity occurred according to the rule described above. To calculate the correlation of the tracked pattern in CA1 with the activity at time $t = 0$, the neurons in CA1 at time t were driven by the input pattern $\mathbf{x}_{EC,tracked}$, $\mathbf{x}^t_{CA3,tracked}$.

Figures 5: To model the results from pririform cortex we studied a two-layer network. As before, patterns were random, binary and sparse with sparseness $f_{\text{input}} = f_{\text{output}} = f$. We stored many such patterns until the network reached a statistical steady state. We then tracked one particular pattern, for which the input vector can be written $\mathbf{x}$. For simplicity we considered linear neurons, and hence the output pattern at time t, $\mathbf{y}^t = \mathbf{C}^t\mathbf{x}$, where $\mathbf{C}^t$ was the connectivity matrix. Simulations with linear-threshold neurons revealed that the nonlinear threshold did not qualitatively affect the results (not shown). To model the effect of familiarization we presented the same (tracked) pattern to the system k times before time $t = 0$ where $k = 1$ indicated a novel pattern. We quantified the RD by calculating the output correlation

$$\text{Corr}(y^t, y^0) = \frac{\text{Cov}(y^t, y^0)}{\sqrt{\text{Var}(y^t)\text{Var}(y^0)}}. \tag{14}$$

The variance and covariance terms can be expressed in terms of first and second-order statistics of the network connectivity, which in turn can be calculated as a Markov process. Specifically, the connectivity depends on transition matrices for the presentation of random patterns or repetitions of the tracked pattern. These matrices can be applied in any order to model a given protocol, see Supplementary Information for detailed calculations.

Conversely, instead of calculating the correlation over time we can calculate the drift, which we defined as

$$\theta_{\text{drift}} = \frac{180}{\pi}\cos^{-1}\Big(\frac{(\mathbf{y^t} - \bar{\mathbf{y}}^\mathbf{t}) \cdot (\mathbf{y^0} - \bar{\mathbf{y}}^\mathbf{0})}{|\mathbf{y^t} - \bar{\mathbf{y}}^\mathbf{t}||\mathbf{y^0} - \bar{\mathbf{y}}^\mathbf{0}|}\Big). \tag{15}$$

Figure 6: For Fig.6a-c this is exactly the same model as for Fig.5. In Fig.6d-f we consider input patterns which themselves also undergo drift. To model this, we assumed that the input i for the rth repetition, x_i^r was identical to x_i^{r-1} with probability s and was otherwise set active with probability f, and otherwise inactive with probability $1 - f$. Therefore for $s = 1$ the input pattern was stable while for $s = 0$ the input pattern was random and uncorrelated from repetition to repetition. At the level of the Markov process, this meant that for each presentation of the tracked pattern we applied the transition matrices for a repetition with prefactor s^r and the transition matrices for a random vector with prefactor $1 - s^r$.

References

1. Hebb, D. *The organization of behavior* (John Wiley and Sons, 1949).

2. Hopfield, J. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554–2558 (1982).

3. Amit, D., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys Rev Lett* **55**, 1530–1533 (1985).

4. Tsodyks, M. & Feigel'man, M. The enhanced storage capacity in neural networks with low activity level. *Europhys Lett* **6**, 101–105 (1988).

5. Amit, D. & Fusi, S. Constraints on linearning in dynamics synapses. *Network: Computation in Neural Systems* **3**, 443–464 (1992).

6. Fusi, S., Drew, P. & Abbott, L. Cascade models of synaptically stored memories. *Neuron* **45**, 599–611 (2005).

7. Ziv, Y. et al. Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience* **16**, 264–266 (2013).

8. Hill, A. First occurence of hippocampal spatial firing in a new environment. *Experimental Neurology* **62**, 282–297 (1978).

9. Frank, L., Stanly, G. & Brown, E. Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience* **24**, 7681–7689 (2004).

10. Epsztein, J., Brecht, M. & Lee, A. Intracellular determinants of hippocampal ca1 place and silent cell activity in a novel environment. *Neuron* **70**, 109–120 (2011).

11. Priestly, J., Bowler, J., Rolotti, S., Fusi, S. & Losonczy, A. Signatures of rapid plasticity in hippocampal ca1 representations during novel experiences. *Neuron* **110**, 1978–1992 (2022).

12. Berners-Lee, A. et al. Hippocampal replays appear after a single experience and incorporate greater detail with more experience. *Neuron* **110**, 1829–1842 (2022).

13. Bostock, E., Muller, R. & Kubie, J. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* **1**, 193–206 (1991).

14. Leutgeb, S. et al. Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science* **309**, 619–623 (2005).

15. Wills, T., Lever, C., Cacucci, F., Burgess, N. & O'Keefe, J. Attractor dynamics in the hippocampal representation of the local environment. *Science* **308**, 873–876 (2005).

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

17. Hainmueller, T. & Bartos, M. Parallel emergence of stable and dynamics memory engrams in the hippocampus. *Nature* **558**, 292–299 (2018).

18. Mau, W. et al. The same hippocampal ca1 population simultaneously codes temporal information over multiple timescales. *Current Biology* **28**, 1499–1508 (2018).

19. Gonzalez, W., Zhang, H., Hanutyunyan, A. & Lois, C. Persistence of neuronal representations through time and damage in the hippocampus. *Science* **365**, 821–825 (2019).

20. Sheintuch, L. et al. Multiple maps of the same spatial context can stably coexist in the mouse hippocampus. *Current biology* **30**, 1467–1476 (2020).

21. Lee, J., Briguglio, J., Cohen, J., Romani, S. & Lee, A. The statistical structure of the hippocampal code for space as a function of time, context, and value. *Cell* **183**, 620–635 (2020).

22. Levy, S., Kinsky, N., Mau, W., Sullivan, D. & Hasselmo, M. Hippocampal spatial memory representations in mice are heterogeneously stable. *Hippocampus* **31**, 244–260 (2020).

23. Keinath, A., Mosser, C. & Brandon, M. The representation of context in mouse hippocampus is preserved despite neural drift. *Nature Communications* **13**, 2415 (2021).

24. Driscoll, L., Petit, N., Minderer, M., Chettih, S. & Harvey, C. Dynamical reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986–999 (2017).

25. Schoonover, C., Ohashi, S., Axel, R. & Fink, A. Representational drift in primary olfactory cortex. *Nature* **594**, 541–546 (2021).

26. Marks, T. & Goard, M. Stimulus-dependent representational drift in primary visual cortex. *Nat Commun* **12**, 5169 (2021).

27. Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. *Curr Biol* **31**, 1–13 (2021).

28. Aschauer, D. et al. Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization. *Cell Reports* **38**, 110340 (2022).

29. Gillett, M., Pereira, U. & Brunel, N. Characteristics of sequential activity in networks with temporally asymmetric hebbian learning. *PNAS* **117**, 29948–29958 (2020).

30. Kossio, Y., Goedeke, S., Klos, C. & Memmesheimer, R. Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation. *PNAS* **118**, e2023832118 (2021).

31. Quin, S. et al. Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning. *Nat Neurosci* **26**, 339–349 (2023).

32. Ratzon, A., Derdikman, D. & Barak, O. Representational drift as a result of implicit regularization. *eLife* **12**, RP90069 (2024).

33. Amit, D. & Fusi, S. Learning in neural networks with material synapses. *Neural Computation* **6**, 957–982 (1994).

34. Roxin, A. & Fusi, S. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Biol* **9**, e1003146 (2013).

35. Attardo, A., Fitzgerald, J. & Schnitzer, M. Impermanence of dendritic spines in live adult ca1 hippocampus. *Nature* **523**, 592–596 (2015).

36. Marr, D. Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci* **262**, 23–81 (1971).

37. Gluck, M. & Myers, C. Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* **3**, 491–516 (1993).

38. Treves, A. & Rolls, E. Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374–391 (1994).

39. McClelland, J. & Goddard, N. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* **6**, 654–655 (1996).

40. Benna, M. & Fusi, S. Place cells may simply be memory cells: memory compression leads to spatial tuning and history dependence. *PNAS* **118**, e2018422118 (2021).

41. Geva, N., Deitch, D., Rubin, A. & Ziv, Y. Time and experience differentially affect distinct aspects of hippocampal representational drift. *Neuron* **111**, 1–10 (2023).

42. Hansel, D. & van Vreeswijk, C. The mechanism of orientation selectivity in primary visual cortex without a functional map. *Journal of Neuroscience* **32**, 4049–4064 (2012).

43. Sheintuch, L., Geva, N., Deitch, D., Rubin, A. & Ziv, Y. Organization of hippocampal ca3 into correlated cell assemblies supports a stable spatial code. *Cell reports* **42**, 112119 (2023).

44. Burgess, N., Maguire, E. & O'Keefe, J. The human hippocampus and spatial and episodic memory. *Neuron* **35**, 625–641 (2002).

45. Moscovitch, M., Cabeza, R., Winocur, G. & Nadel, L. Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annu Rev Psychol* **67**, 105–134 (2016).

46. Gottfried, J., Smith, A., Rugg, M. & Dolan, R. Remembrance of odors past. *Neuron* **42**, 687–695 (2004).

47. Sadeh, N., Zemer, M., Rubin, A. & Ziv, Y. Long-term spatial memory is maintained in the presence of continuous drift in hippocampal representations (2021). Cosyne abstract.

48. Rule, M., O'Leary, T. & Harvey, C. Causes and consequences of representational drift. *Current Opinion in Neurobiology* **58**, 141–147 (2019).

49. Rubin, A. et al. Revealing neural correlates of behavior without behavioral measurements. *Nat Commun* **10**, 4745 (2019).

50. Rule, M. et al. Stable task information from an unstable neural population. *eLife* **9**, e51121 (2020).

51. Raman, D. & O'Leary, T. Optimal plasticity for memory maintenance during ongoing synaptic change. *eLife* **10**, e62912 (2021).

52. Rule, M. & O'Leary, T. Seld-healing codes: How stable neural populations can track continually reconfiguring neural representations. *PNAS* **199**, e2106692119 (2022).

53. Scoville, W. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatr* **20**, 11–21 (1957).

54. Squire, L. & Wixted, J. The cognitive neuroscience of human memory since H.M. *Annu Rev Neurosci* **34**, 259–288 (2011).

55. Squire, L. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychol Rev* **99**, 195–231 (1992).

56. Buzsáki, G. The hippocampal-neocortical dialogue. *Cereb Cortex* **6**, 81–92 (1996).

57. Benna, M. & Fusi, S. Computational principles of synaptic memory consolidation. *Nat Neurosci* **19**, 1697–1706 (2016).

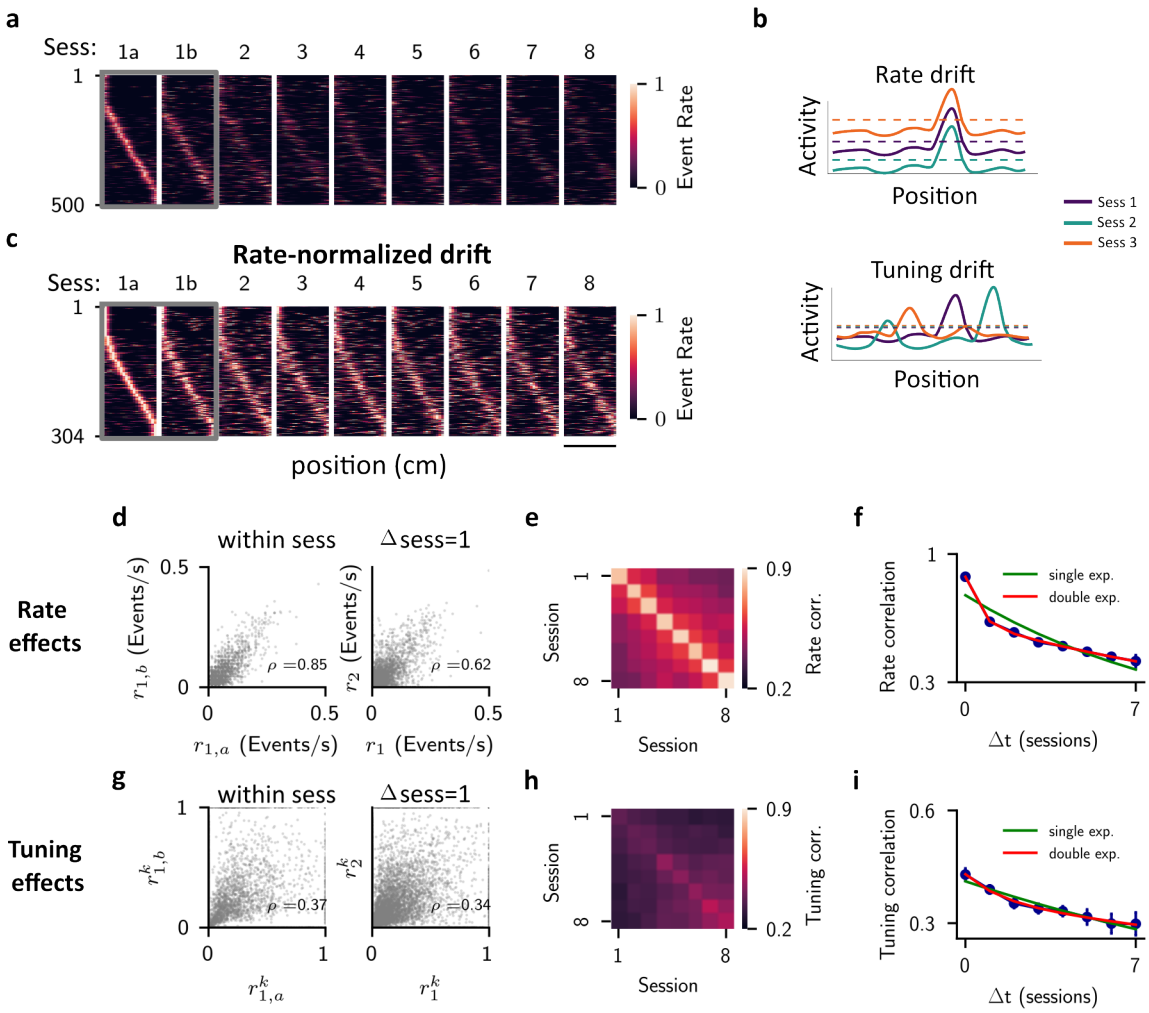
58. Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

Acknowledgements

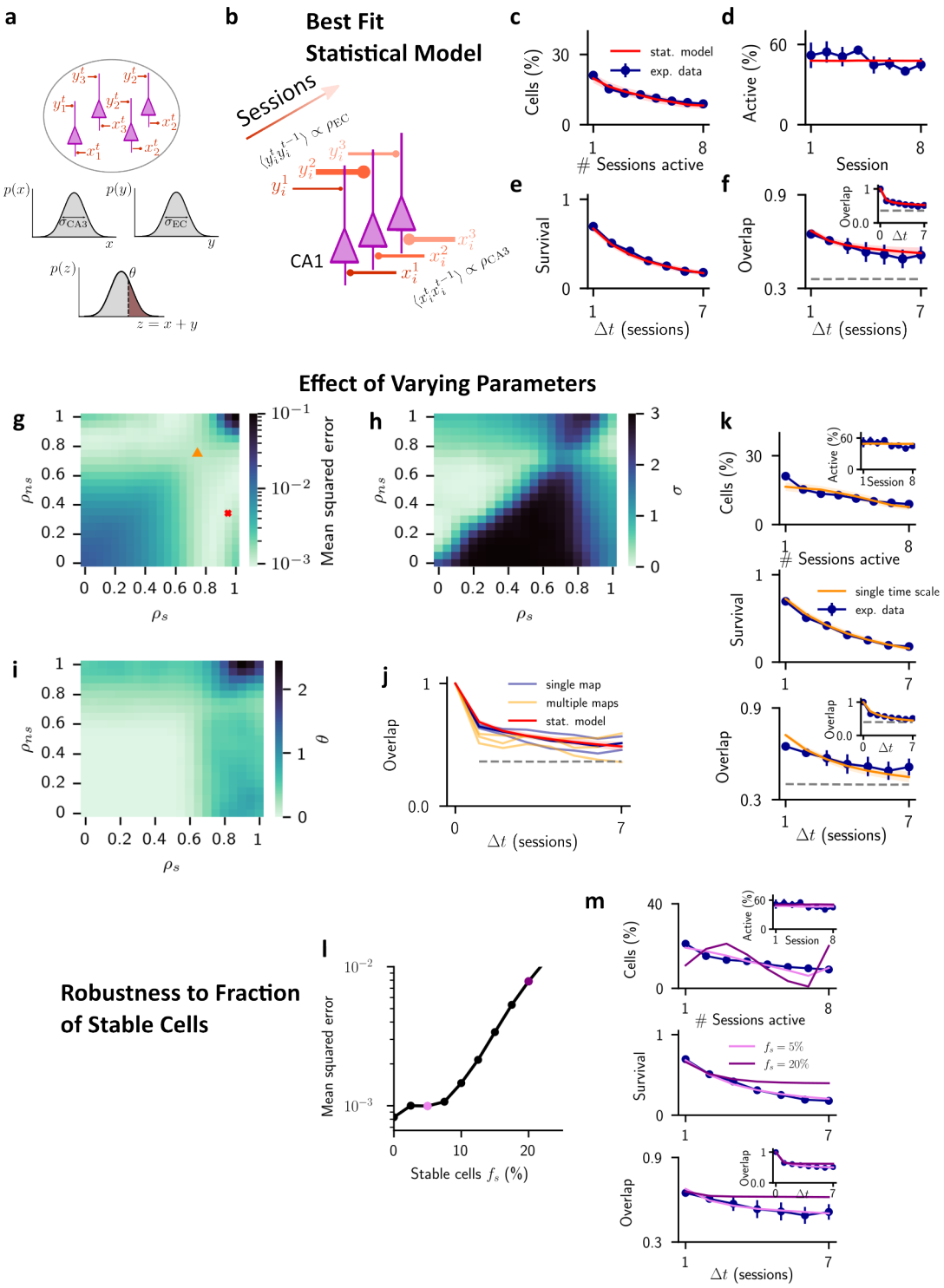
AR and FD thank Yaniv Ziv and Alon Rubin for the experimental data as well as for many helpful and enlightening discussions. AR and FD also thank Klaus Wimmer, Alex Hyafil and Jose M. Esnaola for helpful discussions and feedback. AR acknowledges "Retos" project RTI2018-097570-B-100 from the Ministry of Science and Innovation of the Spanish Government, Flag-Era project from the EU for the Human Brain Project HIPPOPLAST (Era-ICT code PCI2018-093095), "Red de Investigación"

RED2018-102323-T from the Ministry of Science and Innovation of the Spanish Government. This work is supported by the Spanish State Research Agency, through the Severo Ochoa and Maria de Maeztu program for Centers and Units of Excellence in R&D (CEX2020- 001084-M). We thank CERCA Program/Generalitat de Catalunya for institutional support.

Supplementary Figures

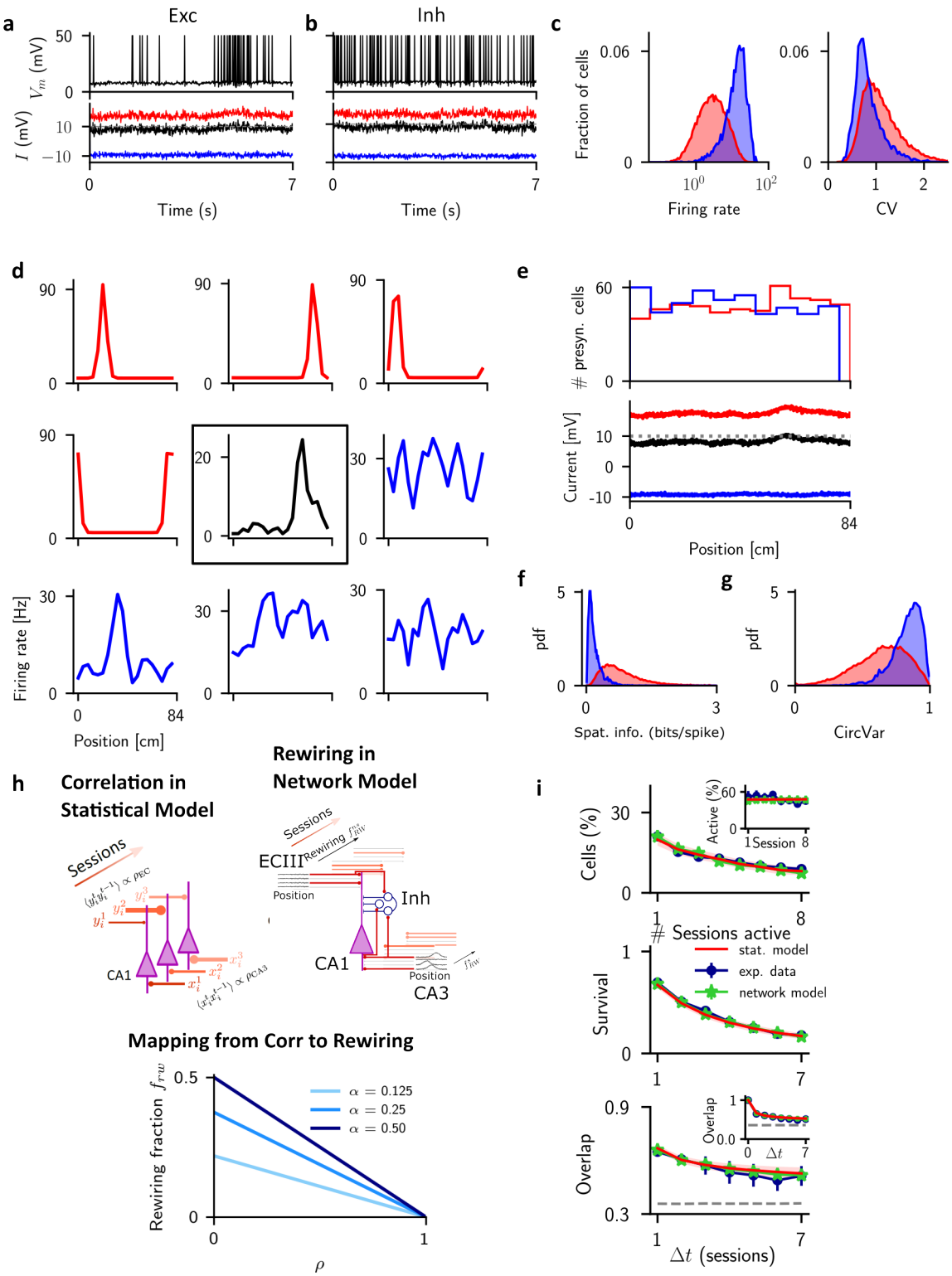


Extended Data Fig. 1 |Analysis of data reveals rate and tuning drift occur on distinct timescales **a**, Place field maps of cells active on even trials on session 1. 1a and 1b stand for even and odd trials of session 1. Cells are ordered according to their peak rate position in session 1a. Each cell event rate is normalized to the maximum reached over the 8 sessions. The scale bar is 80cm. **b**, Illustration of different changes cells may undergo over time. In the top panel, only the mean rate changes from session to session, while tuning is preserved. In the bottom panel, mean rate is preserved while tuning changes over time. **c**, Same as **a**, but plotting cells active on all sessions, and normalizing the event rate within each session. **d**, Scatter plots of mean event rates for all cells. Left: mean event rate on odd trials (ordinate axis) versus mean event rate on even trials. Right: mean event rate on session 2 (ordinate axis) versus mean event rate on session 1. ρ is the Pearson correlation coefficient. **e**, Color map showing the rate correlation for all possible session pairs. The bright diagonal indicates strong within-session correlation compared to across-session correlation. **f**, Mean rate correlation as a function of elapsed sessions (blue points), averaging over all possible session pairs. A double exponential fit (red) reveals the presence of two time scales in the data. **g** Scatter plots of position-specific event rates. Left: event rate in bin k on odd trials (ordinate axis) versus the event rate in the same bin on even trials. Right: event rate in bin k on session 2 (ordinate axis) versus the event rate in the same bin on session 1. **h**, Same as **e**, but for tuning correlation. **i**, Same as **f**, but for tuning correlation. Note that a single time scale captures the decay over time of the correlation.



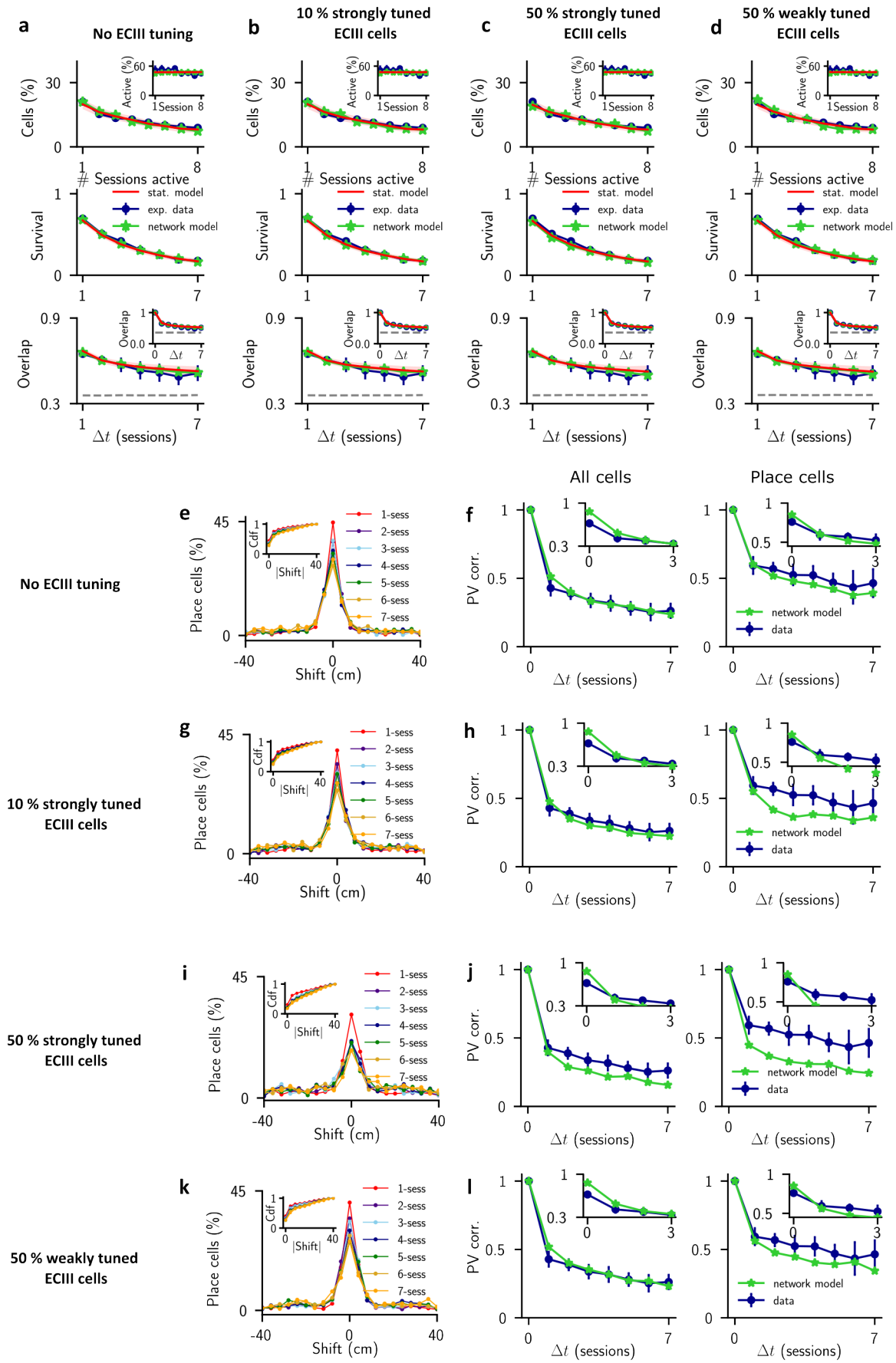
Extended Data Fig. 2 |Modeling session-to-session changes in two sources of input with disparate temporal correlations accounts for data on active cells. **a**, A pool of CA1 cells (in pink), where each cell i receives two distinct inputs x_i^t and y_i^t in each session t. Inputs are Gaussian distributed across cells with variances σ^2_{CA3} and σ^2_{EC} respectively. Cells are active whenever their total input $z_i^t = x_i^t + y_i^t$ exceeds a threshold θ. **b**,The inputs change from session to session with autocorrelations ρ_{CA3} and ρ_{EC} respectively, while preserving overall input statistics (see Methods). **c-f**, Best fit of the model (red line) to the experimental data (blue points): **c**, Distribution of the number of sessions each cell is active; **d**, Fraction of active cells over time; **e**, Survival probability; **f**, Overlap of the population activity vector. Parameters: $\theta = 0.55$, $\sigma_{CA3}/\sigma_{EC} = 1.16$, $\rho_{\text{CA3}} = 0.95$, $\rho_{\text{EC}} = 0.35$. **Fit of the data. a,** Colormap

showing the mean squared error of the fit in the $(\rho_{\mathrm{CA3}}, \rho_{\mathrm{EC}})$ plane. Red cross indicates parameters chosen for Fig.2, orange triangle parameters of panel **e**. **b,c** Colormaps showing parameters σ and θ for the solutions found in panel **a**. **d** Overlap over time for mice with a single map per environment (blue, also shown in Fig.2) and multiple maps per environment (yellow). **e**, Simulations of the statistical model where we impose both inputs to evolve on the same time scale (dark triangle in panel **a**). **Fit of the model with stable fraction of cells. a,** Error of the fit as a function of the percentage of stable cells in the model. Pink and purple dots correspond to simulations of panel **b**. **b**, Best fit of the model with 5 % (pink) and 20 % (purple) of stable cells.

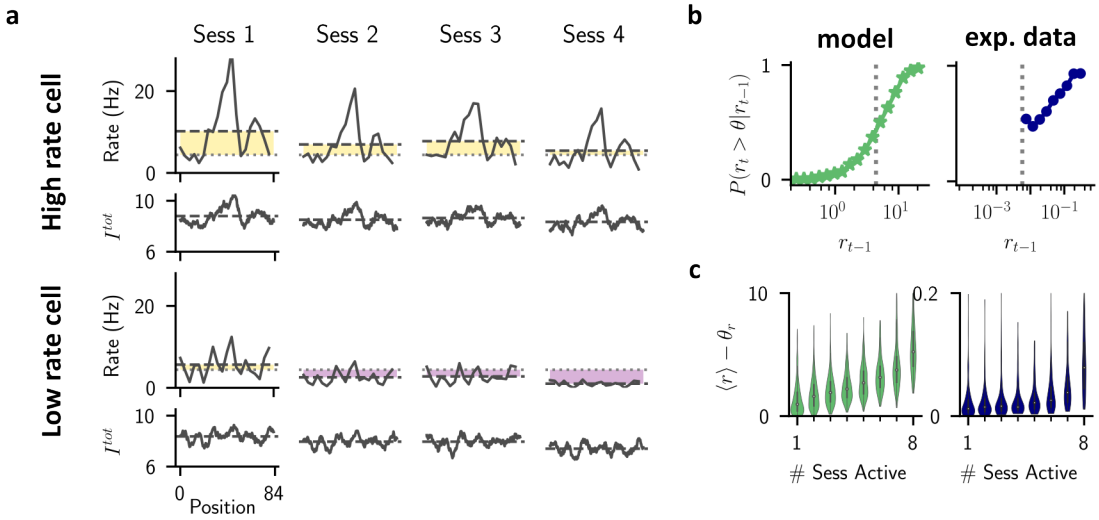


Extended Data Fig. 3 | The network operates in the balanced regime. a, Membrane voltage (top) and excitatory (red), inhibitory (blue) and total (black) currents (bottom) into an example excitatory neuron over one lap on the track (lap time: 7s). Excitatory and inhibitory currents balance, so that the total input is right below threshold. Shading indicated the place field of the cell. **b**, Same as **a**, but for an example inhibitory cell. **c,** Distribution of mean firing rates (left) and ISI coefficient of variations (right) for excitatory (red) and inhibitory (blue) cells. Data pooled across all sessions. Only cells with >10 spikes per session are considered. **Emergence of selectivity in the network. a,** Tuning curve of the same cell as in panel a of Fig.**??** (black), and of 8 of its excitatory (red) and inhibitory (blue) neurons.

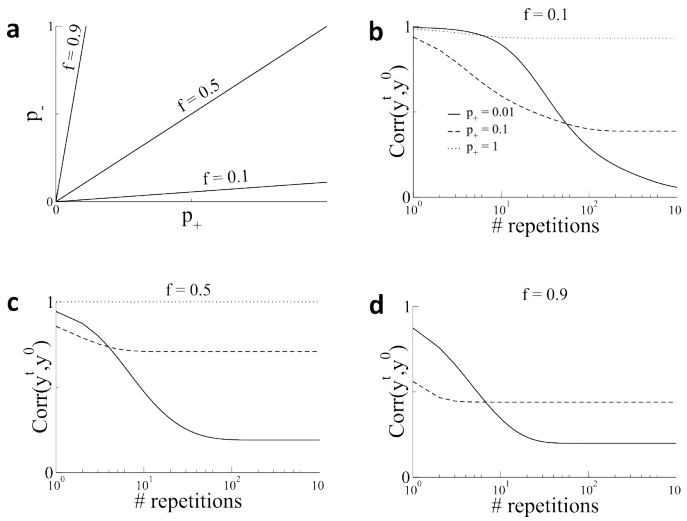
Only tuned excitatory neurons are considered. **b,** Distribution of place field positions of all inputs to the cell highlighted in panel **a** (top), and the corresponding mean excitatory (red), inhibitory (blue) and total (black) synaptic input. **Fraction of connections changing from one session to the next f_{rw} as a function of the correlation** ρ. The value is given by the formula $f_{rw} = 2\alpha\,(1-\alpha)\,(1-\rho)$.


a
No ECIII tuning
b
10 % strongly tuned ECIII cells
c
50 % strongly tuned ECIII cells
d
50 % weakly tuned ECIII cells
Cells (%)
Active (%)
Session
Sessions active
stat. model
exp. data
network model
Survival
Overlap
Δt (sessions)
All cells
Place cells
e
f
g
h
i
j
k
l
No ECIII tuning
10 % strongly tuned ECIII cells
50 % strongly tuned ECIII cells
50 % weakly tuned ECIII cells
Place cells (%)
Cdf
|Shift|
Shift (cm)
1-sess
2-sess
3-sess
4-sess
5-sess
6-sess
7-sess
PV corr.
network model
data

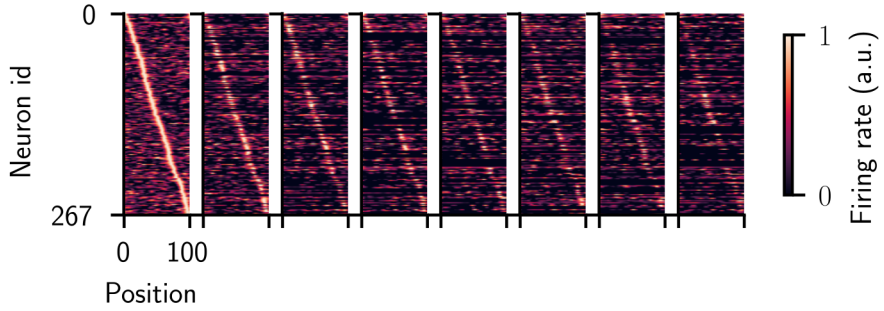
Extended Data Fig. 4 | The effect of tuning in EC on RD. a, Adding tuning to EC does not alter the statistics of active cells. **b,**: Increasing the tuning in EC cells leads to increased diffusion in the position of place cell centroids (decreased peak in histogram of centroids shifts), and a sharper drop in PV correlation of place cells. This affect is most prominent when cells in EC are tuned according to a von Mises distribution, as in CA3, i.e. "strongly tuned". Weaker tuning, of a cosine type, does not lead to significant effects, even if the fraction of tuned cells is large, see bottom row. All parameters are as in Fig.3.



Extended Data Fig. 5 | The correlation between firing rate and stability predicted in network model is observed in the data a. The firing rate and mean total input for two sample cells with high (top) and low (bottom) mean firing rate. Dashed lines indicate averages over all spatial positions, while the light gray dotted line in the threshold. Yellow and purple shadings highlight the distance from threshold, from above and below respectively. **b.** The probability of a cell being above threshold (for detection as active) conditioned on its mean firing rate from the previous session, in the network model (left) and experimental data (right). **c.** Distributions of mean firing rates across sessions, for cells active on different numbers of sessions, in the network model (left) and in the data (right).



Extended Data Fig. 6 | The effect of coding sparseness and learning rates on RD a. The parameters are varied so as to keep the mean connectivity $\bar{c}$ fixed. **b.** Low coding sparseness, $f = 0.1$. **c.** Cofing sparseness $f = 0.5$. For this value the learning rates for potentiation and depression are the same, i.e. $p_- = p_+$. **d.** Coding sparseness $f = 0.9$.



Extended Data Fig. 7 | RD in a network model with Hebbian Plasticity Snapshots of the activity of place cells in a network of threshold linear neurons with Hebbian plasticity. Parameters are chosen to match the decrease in the degree correlation ρ_d from the spiking network, see Methods and Fig.3.

Supplementary Information

We characterized how spatially tuned versus untuned features of neuronal activity differentially contributed to the observed drift in the data. In the case of mice running on a linear track, using the ordering of place cells from the first session shows that there is significant drift on subsequent sessions[7], Fig.2a.

We asked to what extent this drift was due to shifts in the place-field location of cells or simply to changes in rate, Fig.2b. To do so we normalized the time-averaged spatial profile of each place cell by its maximal firing rate on each session. Doing so revealed much less drift, and specifically there was no longer a large difference between the drift from session 1 to session 2 and the degree of variability within session 1, Fig.2c. To quantify this effect we compared scatter plots of the mean event rate of all cells within session 1, and then between sessions 1 and 2, Fig.2d. The correlation of the mean rates evaluated on only even trials versus only odd trials within the session 1 was much higher than the correlation of mean rates between sessions 1 and 2. The pronounced drop in correlation occurred reliably from one session to the next regardless of the session identity, as can be seen in the matrix of rate correlations in Fig.2e. The correlation as a function of session separation averaged over all sessions showed a steep initial drop, followed by a slower decay, Fig.2f. In fact, the curve was much better fit by a double exponential than a single exponential suggesting the presence of two processes operating at distinct time-scales.

On the other hand, when we compared mean rates for each spatial bin of the rate-normalized profiles, we found a much smaller decrease in correlation, Fig.2g. Again this drop was slight from one session to the next irrespective of the session identity, and the curve of the average correlation as a function of session separation showed a shallow decay which was much better fit by a single exponential function than the non-normalized curve, Fig.2h-i. The time-scale of this decay was the same as that of the slower of the two time-scales of the mean rates in Fig.2f. This strongly suggested that the slow time-scale decay in the overall RD was due to changes in the spatial tuning of cells, while the fast drop-off in correlation from one session to the next was related to changes in mean activity.

Statistical Model

We hypothesized that these two time-scales might reflect plasticity along the two main afferent excitatory pathways to CA1: Schaffer collaterals from CA3, and inputs originating from layer III of entorhinal cortex (EC). To test this we investigated if changes in synaptic input alone could account for key experimental findings in a simple, statistical model. Specifically, we considered a population of binary neurons, each one of which received two inputs on any given session t, which we call x^t and y^t, Fig.2a. On the first session we distributed these inputs randomly across neurons according to Gaussian distributions with zero mean and variances σ^2_{CA3} and σ^2_{EC} respectively. This led naturally to a large degree of heterogeneity in the inputs to single cells, with some receiving strong excitatory drive and others being strongly inhibited. The total input to a cell $z^t = x^t + y^t$ was then compared to a threshold θ to determine if the cell was active or not, i.e. cell activity was binary. Therefore, in this simplified model we did not take continuous changes in the event rate of the cells into account, nor did we explicitly include spatial tuning. To model RD we drew these inputs anew from session to session from the same Gaussian distributions, but allowed for some temporal correlation. Namely, the session-to-session autocorrelation for inputs from CA3 was

$$\rho_{CA3} = \langle x^t x^{t-1} \rangle / \sigma^2_{CA3}, \tag{16}$$

where the brackets indicate an average over all neurons, Fig.2b, and there is an equivalent relation for inputs from EC. There were four model parameters in total after normalization: the ratio of the variances, the threshold, and the two autocorrelations. We optimized these parameters to fit four characteristic measures of neuronal activity from the experimental data: 1 - the distribution of the number of sessions in which a cell is active, 2 - the fraction of active neurons per session, 3 - Survival fraction. This measure includes only those cells which are initially active in the first session, and quantifies the fraction of these which continue to be active on subsequent sessions, and 4 - the overlap in the vector of active neurons from session to session. The model was able to reproduce these measures quantitatively Fig.2c-f, but

only when the temporal correlation along one input pathway was close to one ($\rho = 0.95$) and the other sufficiently small ($\rho = 0.35$), Extended Fig.2g-i. In this regime, one of the two inputs changed only very slowly from session to session while the other was almost entirely redrawn randomly for each session, consistent with the two disparate time-scales we observed from direct analysis of the data. Because the model only considered mean inputs, and cell activity was just the sum of the two, it was not possible to say which of the two inputs should vary slowly over time, and which quickly. Once spatial tuning was taken into account (see next section), fast plasticity in EC inputs and slow plasticity in CA3 inputs was the only choice compatible with the data. Finally, we note that a small fraction of cells in the data were active on all 8 sessions, Fig.2c, raising the possibility of a sub-population of stable cells which reliably participated in the hippocampal code. We modeled this by assuming that a fraction f_s of cells were always active, and then allowing the remaining fraction $1 - f_s$ to evolve according to the model described above. We found that the mean squared error of the model fit simultaneously to the four above-mentioned measures increased for increasing f_s, but that it remained low when the stable fraction was below 10 percent, Extended Fig.2l. Therefore, while the model fit was always best without dedicated stable cells, the data would nonetheless also be consistent with a small fraction of stable cells.

Input dynamics as an Ornstein-Uhlenbeck process

The update rules Eq.1 can be rewritten, in continous time, in this form:

$$\epsilon \frac{dz}{dt} = \sqrt{1-\rho^2}\sigma\eta(t) - (1-\rho)\, z(t), \tag{17}$$

where $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(s) \rangle = \delta(t-s)$.
Eq. (17) is a Ornstein-Uhlenbeck with drift coefficient $\mu = (1-\rho)/\epsilon$ and diffusion coefficient $D = \sigma^2 \left(1-\rho^2\right)/(2\epsilon^2)$. The associated Fokker-Planck equation for the probability distribution of the input $P(z,t)$ is then:

$$\frac{\partial P(z,t)}{\partial t} = \mu \frac{\partial (zP)}{\partial z} + D \frac{\partial^2 P}{\partial z^2}. \tag{18}$$

The stationary solution of equation (18) is a gaussian distribution centered at zero with variance

$$\sigma^2 = \frac{D}{\mu} = \frac{\sigma^2 \left(1-\rho^2\right)}{2\epsilon\,(1-\rho)}. \tag{19}$$

We therefore set $\epsilon = \left(1-\rho^2\right)[2\,(1-\rho)]^{-1}$ so that, if the initial distribution of z is $\mathcal{N}(0,\sigma)$ the process is stationary. Eq. (17) is then

$$\frac{dz}{dt} = \sqrt{2D}\eta(t) - \mu z(t), \tag{20}$$

where

$$\mu = \frac{2}{(1-\rho^2)}\,[1-\rho]^2 \quad D = \sigma^2\mu. \tag{21}$$

In the limit $p \to 0$, we have $dz/dt = 0$. When p goes to 1, $dz/dt = 2(\sigma\eta - z)$.
The covariance of a process of the type Eq.(17) is given by:

$$\langle z(t)z(0) \rangle = \sigma^2 e^{-t/\tau}, \tag{22}$$

where we defined $\tau = \mu^{-1}$. Note that $\lim_{\rho\to 0}\tau = 1/2$ and $\lim_{\rho\to 1}\tau = \infty$: when there is no plasticity, the autocorrelation decay time is infinite.

Network equations

Each model neuron follows the LIF dynamics:

$$\tau \dot{V}_i^l = -V_i^l + \tau I_i^l \tag{23}$$

with $i = 1, \ldots, N_l$ and $l \in \{E, I\}$. I_i^l is the synaptic current. The neuron emits a spike when it crosses $V_{\text{th}} = 10$ mV, where it is then resetted to $V_r = 0$ mV.
The synaptic current to excitatory neurons is:

$$I_i^E = \frac{g_s}{\tau_s} \sum_{j=1}^{N_{\text{CA3}}} c_{ij}^{E,s} \sum_k e^{-\left(t - t_{j,k}^s\right)/\tau_s} - \frac{g_{EI}}{\tau_s} \sum_{j=1}^{N_I} c_{ij}^{E,I} \sum_k e^{-\left(t - t_{j,k}^I\right)/\tau_s} + \frac{g_{ns}}{\tau_s} \sum_{j=1}^{N_{\text{EC}}} c_{ij}^{E,ns} \sum_k e^{-\left(t - t_{j,k}^{ns}\right)/\tau_s} \quad (24)$$

where $g_{s,EI}$ are the feedforward and E $\leftarrow$ I synaptic conductances, $t_{j,k}^l$ are the times of the kth action potential of neuron j of the l population, and $C^{l,r}$ the connectivity matrices.
The synaptic current to inhibitory neurons is:

$$I_i^I = \frac{g_{IE}}{\tau_s} \sum_{j=1}^{N_E} c_{ij}^{I,E} \sum_k e^{-\left(t - t_{j,k}^E\right)/\tau_s} - \frac{g_{II}}{\tau_s} \sum_{j=1}^{N_I} c_{ij}^{I,I} \sum_k e^{-\left(t - t_{j,k}^I\right)/\tau_s} + \frac{g_{Is}}{\tau_s} \sum_{j=1}^{N_{\text{CA3}}} c_{ij}^{E,s} \sum_k e^{-\left(t - t_{j,k}^s\right)/\tau_s} + \frac{g_{Ins}}{\tau_s} \sum_{j=1}^{N_{\text{EC}}} c_{ij}^{E,ns} \sum_k e^{-\left(t - t_{j,k}^{ns}\right)/\tau_s} \quad (25)$$

Unless otherwise stated, we set $g_{Is} = g_s$ and $g_{Ins} = g_{ns}$. To set the network in the so called balance state, the synaptic weights are scaled according to:

$$g_l = \frac{G_l}{\sqrt{K}}, \quad (26)$$

where K is the average number of active inputs to each subpopulation, and G_l is indepedent of K

Network parameters

The membrane time constant is set to $\tau = 10$ ms, and the synaptic time constant $\tau_s = 5$ ms. The synaptic weights of CA1 recurrent connections are $G_{EI} = 2.8$ mV, $G_{IE} = 2.0$ mV, $G_{II} = 2.0$ mV[1]. The other synaptic weights (that determine the variances of spatial and non-spatial inputs) are set in the following considering the fit of the statistical model. Parameters of the CA3 place fields are $\beta = 19.85$, $r_b = 5$ Hz, $\bar{R} = 90$ Hz.
Unless specified otherwise, we consider $N_E = 4000$ and $N_I = 1000$ CA1 cells. The connections probability is the same for all connectivity matrices and equal to $\alpha = 0.125$. The sparsity of the CA3 and EC layer are set respectively to $f_{CA3} = 0.5$ and $f_{EC} = 0.5$. We wish to have about the same number of excitatory CA1, CA3 and EC active cells, so that we set $N_{CA3} = N_E/f_{CA3} = 8000$ and $N_{EC} = N_E/f_{EC} = 8000$. The fraction of tuned CA3 inputs f_s is set to $f_s = 0.5$, unless specified otherwise.

Input correlations in the network model

In this appendix, we calculate the amount of correlations of the inputs expected in the network model, for the same or different environments. In the network model, average synaptic inputs to a cell i in environment A can be written in the form:

$$I_i^A = \sum_j c_{ij}^A \nu_j^A, \quad (27)$$

[1]Note that these weights are then rescaled according to Eq.26.

where we omitted conductances and time constants. c_{ij}^A is the matrix element during visits of environment A (if visits do not occur at the same time plasticity may have occurred), and ν_j^A is the firing rate of input neuron j in environment A.

In general, the Pearson correlation of inputs from different environments can be written as:

$$\rho = \frac{\langle I^A I^B \rangle - \mu_A \mu_B}{\sigma_A \sigma_B}, \tag{28}$$

where we have defined $\mu_{A,B} = \langle I^{A,B} \rangle$ and dropped the index i for simplicity. Note that the bracket averages are all over the index i (i.e. we consider one specific realization of the input neurons rate ν). The expected values take the form:

$$\mu_{A,B} = \sum_j \langle c_j^{A,B} \nu_j^{A,B} \rangle = \alpha \sum_j \nu_j^{A,B}, \tag{29}$$

where for the second equality we have assumend that the connection probability α in the two environments is the same. On the other hand, we have for the expected value of the product of the inputs:

$$\langle I^A I^B \rangle = \sum_j \sum_l \langle c_j^A c_l^B \nu_j^A \nu_l^B \rangle = \sum_j \langle c_j^A c_j^B \rangle \nu_j^A \nu_j^B + \sum_{j \neq l} \langle c_j^A c_l^B \rangle \nu_j^A \nu_l^B. \tag{30}$$

We then have

$$\langle I^A I^B \rangle = \sum_j \nu_j^A \nu_j^B \langle c_j^A c_j^B \rangle + \alpha^2 \sum_{j \neq l} \nu_j^A \nu_j^B. \tag{31}$$

For the within-environment variances, we have:

$$\left(\sigma^{A,B}\right)^2 = \sum_j \sum_l \nu_j \nu_l \langle c_j c_l \rangle - \mu^2 = \alpha \sum_j \nu_j^2 + \alpha^2 \sum_{j \neq l} \nu_j \nu_l - \alpha^2 \sum_{j,l} \nu_j \nu_l, \tag{32}$$

where we dropped the indeces A,B for simplicity and substituted the previously found expression for μ. Adding and removing the element where $j = l$ in the second term gives:

$$\sigma^2 = \alpha \left(1 - \alpha\right) \sum_j \nu_j^2. \tag{33}$$

Using the same trick in the numerator of adding and removing one element from the sum, we get for the correlation:

$$\rho = \frac{\sum_j \nu_j^A \nu_j^B \left(\langle c_j^A c_j^B \rangle - \alpha^2 \right)}{\alpha \left(1 - \alpha\right) \left[\sum_j \left(\nu_j^A \right)^2 \right]^{1/2} \left[\sum_j \left(\nu_j^B \right)^2 \right]^{1/2}}. \tag{34}$$

If we now consider that the input layer is very large, then all sums over j can be rewritten in terms of mean values. If we further assume that the input patterns are drawn from the same distribution for the different environments, then we have:

$$\rho = \frac{\langle \nu_j^A \nu_j^B \rangle_j \left[\langle c_j^A c_j^B \rangle - \alpha^2 \right]}{\langle \nu_j^2 \rangle_j \alpha \left(1 - \alpha\right)}, \tag{35}$$

where $\langle \cdot \rangle_j$ stands for the average over the input layer.

Eq.(35) is the general form of the inputs considering different environments/directions or the same environment at different times. We now consider three separate cases:

1. Visits of the same environment at different times (same inputs firing rates but different connectivity matrices)

2. Partially correlated environments at approximately the same time (different input firing rates but constant connectivity)

3. Orthogonal environment at different times.

Case 1. If the environment is the same, at least for the spatial input we can consider the input firing rates to be constant over visits (in first approximation). This implies $\langle \nu_j^A \nu_j^B \rangle_j = \langle \nu_j^2 \rangle_j$, which then gives:

$$\rho_t = \frac{\langle c_j^t c_j^{t+1} \rangle - \alpha^2}{\alpha (1-\alpha)}, \tag{36}$$

where t stands for the time of the visit. Eq. (36) can be made more explicit by expressing $\langle c_j^t c_j^{t+1} \rangle$ in terms of probabilities. In fact, since c_j's are binary matrix elements, we can write:

$$\langle c_j^t c_j^{t+1} \rangle = \text{Pr}\left(c_j^t = 1, c_j^{t+1} = 1\right) = \text{Pr}\left(c_j^{t+1} = 1 \middle| c_j^t = 1\right) \text{Pr}\left(c_j^t = 1\right). \tag{37}$$

Since we have $\text{Pr}\left(c_j^t = 1\right) = \alpha$, we then have

$$\rho_t = \frac{\text{Pr}\left(c_j^{t+1} = 1 \middle| c_j^t = 1\right) - \alpha}{1-\alpha}. \tag{38}$$

If there is no plasticity, then $\text{Pr}\left(c_j^{t+1} = 1 \middle| c_j^t = 1\right) = 1$ and $\rho_t = 1$, while if there is a complete rewiring $\text{Pr}\left(c_j^{t+1} = 1 \middle| c_j^t = 1\right) = \alpha$ and $\rho_t = 0$.

Case 2. In this case, we assume that the connectivity matrix is the same for the different visits, and the input patterns are partially correlated. In the data we have, this would for example be the case of considering the two running directions as different environments. We then have here that $\langle c_j^A c_j^B \rangle = \alpha$, so that the correlation is

$$\rho_{LR} = \frac{\langle \nu_j^L \nu_j^R \rangle_j}{\langle \nu_j^2 \rangle_j}, \tag{39}$$

where L and R stands for left and right. Input neurons j are either silent or active with firing rate ν, so that we have $\langle \nu_j^2 \rangle_j = f\nu^2$, where f is the sparsity of the input layer (fraction of active cells). As for the previous case, we can write more explicit the correlation using probabilities:

$$\rho_{LR} = \frac{\nu^2 \text{Pr}\left(\nu^L = \nu, \nu^R = \nu\right)}{f\nu^2} = \text{Pr}\left(\nu^L = \nu \middle| \nu^R = \nu\right); \tag{40}$$

the correlation between running directions is just the probability that given an input cell is active in one running idrection, it is active also in the other.

Case 3. Since environments are assumed to be orthogonal or independent, then we have $\langle \nu^A \nu^B \rangle = f\nu^2$, so that we get for the correlation

$$\rho_{AB} = \frac{f\left(\langle c^A c^B \rangle - \alpha^2\right)}{\alpha(1-\alpha)}. \tag{41}$$

We summarize here the three correlations found:

- $\rho_t = \dfrac{\text{Pr}\left(c_j^{t+1} = 1 \middle| c_j^t = 1\right) - \alpha}{1-\alpha}$,
- $\rho_{LR} = \text{Pr}\left(\nu^L = \nu \middle| \nu^R = \nu\right)$,
- $\rho_{AB} = \dfrac{f\left(\langle c^A c^B \rangle - \alpha^2\right)}{\alpha(1-\alpha)}$.

Steady-state Connectivity Statistics in Hebbian plasticity model

We assume that the learning process has gone on long enough so that the connectivity has reached a stationary state and hence is independent of the initial configuration. Then we consider the encoding of one pattern and calculate the mean and the variance in the number of inputs to a cell in the postsynaptic layer (CA1). We assume N neurons in the presynaptic layer, e.g. EC, and patterns with sparseness $f_{\text{pre}} = f_{\text{post}} = f$ for simplicity. The synapse between two co-active cells is potentiated with probability p_+, and a synapse between one active and one inactive cell is depressed with probability p_-. Synapses between inactive cells are not updated. We consider binary synapses such that the connection from cell j in the presynaptic layer to cell i in the postsynaptic layer is $c_{ij} \in \{0, 1\}$. The degree of a cell i can be written $d_i = \sum_{j=1}^{N} c_{ij}$. The mean degree in the network is then $\mu_d = \langle d_i \rangle = N\langle c_{ij} \rangle = N\bar{c}$, and the average over synapses can be written as

$$\bar{c} = f_{\text{post}} \langle c_{ij} \rangle_{i \in \Omega_f} + (1 - f_{\text{post}}) \langle c_{ij} \rangle_{i \notin \Omega_f}, \tag{42}$$

where ω_f is the set of neurons which are part of the current, activated pattern. By considering the probability that a synapse is in the potentiated state depending on whether or not the neuron is part of the pattern, one arrives at

$$\langle c_{ij} \rangle = \frac{f^2 p_+}{f^2 p_+ + 2f(1-f)p_-} \tag{43}$$

The variance in the degree $\sigma_d^2 = \langle d_i^2 \rangle - \mu_d^2$. The first term can be written

$$\langle d_i^2 \rangle = \langle \sum_j c_{ij}^2 + \sum_j \sum_{k \neq j} c_{ij} c_{ik} \rangle = N \langle c_{ij}^2 \rangle + N(N-1) \langle c_{ij} c_{ik} \rangle$$

The first term $\langle c_{ij}^2 \rangle = \langle c_{ij} \rangle = \bar{c}$, while the second term can be calculated by considering the probability of both connections being potentiated as a function of whether or not the pre- and postsynaptic cells are active in the current pattern, or not. For simplicity we write $\bar{c}_{11} = \langle c_{ij} c_{ik} \rangle = Pr(c_{ij} = 1, c_{ik} = 1)$. The result is

$$\bar{c}_{11} = \frac{f^2 p_+ \Big(f(p_+ + 2(1-p_+)\bar{c}) + 2(1-f)(1-p_-)\bar{c} \Big)}{1 - f\big(f(1-p_+) + (1-f)(1-p_-)\big)^2 - (1-f)(1-fp_-)^2} \tag{44}$$

Finally, the autocorrelation of the degree can be written

$$\rho_d = \frac{\langle d_i^t d_i^{t-1} \rangle - \mu_d^2}{\sigma_d^2}, \tag{45}$$

where we can write $\langle d_i^t d_i^{t-1} \rangle = N \langle c_{ij}^t c_{ij}^{t-1} \rangle + N(N-1) \langle c_{ij}^t c_{ik}^{t-1} \rangle$. The two covariances can, once again, be calculated by considering the probability that the product is equal to one depending on the participation of the cells in the patterns at time $t-1$ and t. The result is

$$\langle c_{ij}^t c_{ij}^{t-1} \rangle = \bar{c} \Big(1 - 2p_- f(1-f) \Big), \tag{46}$$

$$\langle c_{ij}^t c_{ik}^{t-1} \rangle = \bar{c} f^2 p_+ + \bar{c}_{11} \Big(1 - f^2 p_+ - 2p_- f(1-f) \Big) \tag{47}$$

Analytical solution for RD in Plasticity Model

The relative simplicity of the plasticity model allows us to calculate how an output pattern changes in time, given a particular input. The drop in correlation of the output pattern, or the drift in the representation, can therefore be calculated analytically. We recall that the parameters of the model are the learning rates p_+ and p_-, the coding fraction f, and the population size N. The theoretical solution will be exact in the limit $N \to \infty$. We will consider the scenario for which the network has reached a steady state after many statistically equivalent patterns have been encoded. We then encode the pattern to be tracked, and

allow for an arbitrary protocol of repetitions and interleaved random patterns. In this way the effects of familiarization, repetition rate as well as the number of interleaved unrelated episodes can all be studied.

We calculate the correlation between the activity of the network at time t and the activity at time $t = 0$ which can be written

$$\mathrm{Corr}(y^t, y^0) = \frac{\mathrm{Cov}(y^t, y^0)}{\sqrt{\mathrm{Var}(y^t)\mathrm{Var}(y^0)}}, \tag{48}$$

where

$$\mathrm{Cov}(y^t, y^0) = \frac{1}{N}\sum_{i=1}^{N} y_i^t y_i^0 - \bar{y}^t \bar{y}^0, \tag{49}$$

and $\mathrm{Var}(y^t) = \mathrm{Cov}(y^t, y^t)$, while the mean output is

$$\bar{y}^t = \frac{1}{N}\sum_{i=1}^{N} y_i^t. \tag{50}$$

The vector y^t is the output vector of the network at time t in response to an input vector which we call $x \in \{0, 1\}$. For simplicity we assume that the input vector does not change, and hence the activity of a cell i is

$$y_i^t = \frac{1}{N}\sum_{j=1}^{N} c_{ij}^t x_j, \tag{51}$$

where $c_{ij}^t = 1$ if there is a synaptic connection, or else $c_{ij}^t = 0$. We can choose the ordering of the presynaptic and postsynaptic cells such that pattern we are tracking is the one where the first fN cells are the active ones. In that case

$$y_i^t = \frac{1}{N}\sum_{j=1}^{fN} c_{ij}^t, \tag{52}$$

where the mean and variance of the input will depend on whether $i \leq fN$ or not. For $i \leq fN$ the cell will have been active in the tracked memory. Specifically, when the tracked memory is encoded, we have ordered the cells so that in the first quadrant of the connectivity matrix there is an fN by fN size block in which the synapses have been subjected to potentiations. Below this there is an $(1 - f)N$ by fN block where the synapses have been subjected to depressions. The remaining two blocks in the matrix are not relevant for determining the output correlation since the input x is a vector of ones for the first fN entries and otherwise zero.

The mean input $\bar{y}^t$ is

$$\begin{aligned}
\bar{y}^t &= \frac{1}{N}\sum_{i=1}^{N} y_i^t, \\
&= \frac{1}{N}\sum_{i=1}^{fN} y_i^t + \frac{1}{N}\sum_{i=fN+1}^{N} y_i^t, \\
&= f\langle y^t\rangle_P + (1-f)\langle y^t\rangle_D, \\
&= f^2\langle c^t\rangle_P + f(1-f)\langle c^t\rangle_D,
\end{aligned} \tag{53}$$

where for the last line we have used Eq.52, and the subscripts P and D refer to the statistics of inputs which have been potentiated or depressed during the storage of the tracked memory, respectively.

The variance of the input is

$$\begin{aligned}
\mathrm{Var}(y^t) &= \frac{1}{N}\sum_i (y_i^t)^2 - (\bar{y}^t)^2, \\
&= \frac{1}{N}\sum_{i=1}^{fN}(y_i^t)^2 + \frac{1}{N}\sum_{i=fN+1}^{N}(y_i^t)^2 - (\bar{y}^t)^2, \\
&= \frac{f}{N^2}\Big(\sum_{j=1}^{fN}\langle c_{ij}^t\rangle_P + \sum_{j=1}^{fN}\sum_{k\neq j}\langle c_{ij}^t c_{ik}^t\rangle_P\Big) + \frac{1-f}{N^2}\Big(\sum_{j=1}^{fN}\langle c_{ij}^t\rangle_D + \sum_{j=1}^{fN}\sum_{k\neq j}\langle c_{ij}^t c_{ik}^t\rangle_D\Big) - (\bar{y}^t)^2, \\
&= \frac{f^2}{N}\Big(\langle c^t\rangle_P - \langle cc^t\rangle_P\Big) + f^3\langle cc^t\rangle_P + \frac{f(1-f)}{N}\Big(\langle c^t\rangle_D - \langle cc^t\rangle_D\Big) + f^2(1-f)\langle cc^t\rangle_D \\
&\quad -(\bar{y}^t)^2, \qquad (54)
\end{aligned}$$

where for simplicity we have written $\langle c_{ij}^t c_{ik}^t\rangle = \langle cc^t\rangle$.

The covariance of the input at times 0 and t is

$$\begin{aligned}
\mathrm{Cov}(y_t, y_0) &= \frac{f^2}{N}\Big(\langle c_{ij}^t c_{ij}^0\rangle_P - \langle c_{ij}^t c_{ik}^0\rangle_P\Big) + f^3\langle c_{ij}^t c_{ik}^0\rangle_P + \frac{f(1-f)}{N}\Big(\langle c_{ij}^t c_{ij}^0\rangle_D - \langle c_{ij}^t c_{ik}^0\rangle_D\Big) \\
&\quad + f^2(1-f)\langle c_{ij}^t c_{ik}^0\rangle_D - \bar{y}^t\bar{y}^0, \qquad (55)
\end{aligned}$$

Markov Processes for 1st and 2nd-order Statistics of Synapses

i. Mean weight in potentiated quadrant, $\langle c^t\rangle_P$:

A single presentation of the tracked pattern will shift the steady-state probabilities in the potentiated quadrant according to

$$\mathbf{c_P} = \mathbf{R_P}\mathbf{c_{ss}}, \qquad (56)$$

where $\mathbf{c_{ss}} = (\bar{c}, 1-\bar{c})$ and

$$\mathbf{R_P}\begin{pmatrix} 1 & p_+ \\ 0 & 1-p_+ \end{pmatrix}. \qquad (57)$$

The mean weight at time t is then given by

$$\langle c^t\rangle = \mathbf{e_1}^{\mathrm{T}}\mathbf{A}^t\mathbf{c_P}, \qquad (58)$$

where

$$\mathbf{A} = \begin{pmatrix} 1-2f(1-f)(1-p_-) & f^2p_+ \\ 2f(1-f)p_- & 1-f^2p_+ \end{pmatrix}, \qquad (59)$$

and $\mathbf{e_1} = (1, 0)$.

ii. Mean weight in depressed quadrant, $\langle c^t\rangle_D$:

The Markov process is identical, with the sole difference being the initial condition, where here $\mathbf{c_D} = \mathbf{R_D}\mathbf{c_{ss}}$ with

$$\mathbf{R_D}\begin{pmatrix} 1-p_- & 0 \\ p_- & 1 \end{pmatrix}. \qquad (60)$$

iii. 2nd-order statistic for weights in potentiated quadrant, $\langle cc^t\rangle_P$

A single presentation of the tracked pattern shift the steady-state 2nd-order statistics in the potentiated quadrant according to

$$\xi_{\mathbf{P}} = \mathbf{Q_P}\xi_{\mathbf{ss}}, \qquad (61)$$

with $\xi_{\mathbf{ss}} = (\bar{c}_{11}, \bar{c}_{01}, \bar{c}_{00})$, where $\bar{c}_{11}$ is given by Eq.44,

$$\bar{c}_{00} = \frac{f(1-f)\Big(p_-^2 + 2p_-(1-\bar{c})(2-p_- - fp_+)\Big)}{1-f(1-fp_+-(1-f)p_-)^2-(1-f)(1-fp_-)^2} \qquad (62)$$

and $\bar{c}_{01} = (1 - \bar{c}_{11} - \bar{c}_{00})/2$. The transition matrix for a repetition is

$$\mathbf{Q_P} = \begin{pmatrix} 1 & 2p_+ & p_+^2 \\ 0 & 1-p_+ & p_+(1-p_+) \\ 0 & 0 & (1-p_+)^2 \end{pmatrix}. \tag{63}$$

The 2nd-order statistic at time t is then given by

$$\langle cc^t \rangle_P = \mathbf{e_1}^{\mathrm{T}} \mathbf{B}^t \xi_{\mathbf{P}}, \tag{64}$$

where $\mathbf{e_1} = (1, 0, 0)$ and the coefficients of the matrix $\mathbf{B}$ are

$$\begin{aligned}
B_{11} &= f^3 + (1-f)^3 + f(1-f)(1-p_-)(3-p_-), \\
B_{12} &= 2f^3 p_+ + 2f^2(1-f)p_+(1-p_-), \\
B_{13} &= f^3 p_+^2, \\
B_{21} &= f(1-f)p_-(2-p_-), \\
B_{22} &= f(1-f) + (1-f)^3 + 2f(1-f)(1-p_-) + f^3(1-p_+) - f^2(1-f)p_+(1-p_-) + f^2(1-f)p_+p_-, \\
B_{23} &= f^2 p_+(1-fp_+), \\
B_{31} &= f(1-f)p_-^2, \\
B_{32} &= 2f(1-f)(2-f)p_- + 2f^2(1-f)p_-(1-p_+), \\
B_{33} &= 1 - f + f(1-fp_+)^2.
\end{aligned}$$

iv. 2nd-order statistic for weights in depressed quadrant, $\langle cc^t \rangle_D$:

The Markov process is identical, with the sole difference being the initial condition, where here $\xi_{\mathbf{D}} = \mathbf{Q_D} \xi_{\mathbf{ss}}$ with

$$\mathbf{Q_D} = \begin{pmatrix} (1-p_-)^2 & 0 & 0 \\ p_-(1-p_-) & 1-p_+ & 0 \\ p_-^2 & 2p_- & 1 \end{pmatrix}. \tag{65}$$

Then

$$\langle cc^t \rangle_D = \mathbf{e_1}^{\mathrm{T}} \mathbf{B}^t \xi_{\mathbf{D}}, \tag{66}$$

v. Two-point statistic for synaptic weight in potentiated quadrant, $\langle c_{ij}^t c_{ij}^0 \rangle_P$:

$$\begin{aligned}
\langle c_{ij}^t c_{ij}^0 \rangle_P &= \Pr(c_{ij}^t = 1, c_{ij}^0 = 1)_P, \\
&= \Pr(c_{ij}^t | c_{ij}^0 = 1) \Pr(c_{ij}^0 = 1)_P, \\
&= \left[\mathbf{e_1}^{\mathrm{T}} \mathbf{A^t} \mathbf{e_1}\right] \mathbf{e_1}^{\mathrm{T}} \mathbf{c_P}.
\end{aligned} \tag{67}$$

vi. Two-point statistic for synaptic weight in depressed quadrant, $\langle c_{ij}^t c_{ij}^0 \rangle_D$:

$$\langle c_{ij}^t c_{ij}^0 \rangle_D = \left[\mathbf{e_1}^{\mathrm{T}} \mathbf{A^t} \mathbf{e_1}\right] \mathbf{e_1}^{\mathrm{T}} \mathbf{c_D}. \tag{68}$$

Modeling different protocols with the Markov Process

Using this Markov process to model a simulation protocol involves using the transition matrices in the proper order. For example, to determine the state of the mean weights in the potentiated quadrant after the presentation of the tracked pattern, a number ISI of random pattern, and another repetition, is given by

$$\bar{\mathbf{c}}^{\mathrm{final}} = \mathbf{R_p} \mathbf{A}^{\mathrm{ISI}} \mathbf{R_p} \bar{\mathbf{c}}, \tag{69}$$

where $\bar{\mathbf{c}} = (\bar{c}, 1 - \bar{c})$. A similar operation must be performed for the second-order statistics in the potentiated quadrant, using the corresponding transition matrices, and the same for the depressed quadrant.

Drift in input pattern

If the input pattern itself undergoes drift from repetition to repetition, then this must be accounted for in the calculation of the correlation. We assume that if the tracked input pattern at time $t = 0$ is $\mathbf{x}$, then the ith input for the first repetition x_i^1 will be identical to that at time 0 with probability s and otherwise its state will be set to active with probability f or inactive with probability $1 - f$. In a slight abuse of notation we can write $\mathbf{x^1} = s\mathbf{x^0} + (1 - s)\xi^1$ where ξ is a random, binary vector with sparseness f. If this is the case, then it is straightforward to show that the rth input vector has correlation s^r with the input vector at time $t = 0$. To model the Markov process corresponding to this choice of input vectors, for the rth repetition we applied the repetition transition matrix with prefactor s^r and the random transition matrix with prefactor $1 - s^r$. Hence, for updating the mean weights in the potentiation vector we used $s^r\mathbf{R}_p + (1 - s^r)\mathbf{A}$.